MIT CRITICAL DATA

Análisis Secundario de Historias Clínicas Electrónicas

Leo Anthony Gutierrez Celi
Peter Harcourt Charlton
Mohammad Mahdi Ghassemi
Alistair Johnson
Matthieu Jean Claude Komorowski
Dominic Charles Marshall
Tristan Josef Naumann
Kenneth Eugene Paik
Tom Joseph Pollard
Jesse Daniel Raffa
Justin Daniel Salciccioli

EDITORIAL DUNKEN

MIT Critical Data está integrado por un grupo de científicos de datos y médicos provenientes de distintos países del mundo y reunidos por una visión en común: generar un sistema de salud basado en datos apoyado por la informática en salud sin fronteras. En este ecosistema, la creación de evidencia y de herramientas para apoyar las decisiones clínicas se inicia, actualiza, mejora y perfecciona al ampliar el acceso y el uso significativo de los datos clínicos.

Este libro ha sido escrito para que pueda iniciarse su lectura por diferentes secciones, de acuerdo al nivel de conocimiento previo de los lectores. En el caso de los investigadores principiantes, el libro debería leerse desde el inicio al fin. Aquellos que ya están familiarizados con los desafíos de la informática en salud, pero desearían una orientación para realizar el análisis de forma más efectiva, deberían leer el libro a partir de la segunda parte. Finalmente, la parte del estudio de casos que provee consideraciones prácticas específicas sobre el diseño y metodología de estudio, es recomendada para todos los lectores.

Ha llegado el momento de aprovechar los datos que generamos durante la atención habitual de los pacientes para formular un compendio más completo de recomendaciones basadas en evidencias y para apoyar a la toma de decisiones compartida con los pacientes. Este libro tiene como objetivo entrenar a la próxima generación de científicos, representando diferentes disciplinas pero colaborando para expandir la base de conocimiento que guiará la práctica médica en el futuro.

MIT CRITICAL DATA

ANÁLISIS SECUNDARIO DE HISTORIAS CLÍNICAS ELECTRÓNICAS

Leo Anthony Gutierrez Celi - Peter Harcourt Charlton - Mohammad Mahdi Ghassemi - Alistair Johnson- Matthieu Jean Claude Komorowski - Dominic Charles Marshall - Tristan Josef Naumann - Kenneth Eugene Paik - Tom Joseph Pollard - Jesse Daniel Raffa- Justin Daniel Salciccioli

EDITORIAL DUNKEN

Buenos Aires

2021

Análisis Secundario de Historias Clínicas Electrónicas / Leo Anthony Gutierrez Celi... [et al.].
1a ed.- Ciudad Autónoma de Buenos Aires : Dunken, 2021.
Libro digital, EPUB

Archivo Digital: descarga y online
ISBN 978-987-85-1727-8

1. Medicina. I. Gutierrez Celi, Leo Anthony.
CDD 616.09

Contenido y corrección a cargo de los autores.
Ayacucho 357 (C1025AAG) - Capital Federal
Tel/fax: 4954-7700 / 4954-7300
E-mail: *info@dunken.com.ar*
Página web: *www.dunken.com.ar*

Hecho el depósito que prevé la ley 11.723

ISBN 978-987-85-1727-8

EQUIPO DE TRADUCCIÓN

Directores

María del Pilar Arias López (@pilar262004)
[Argentina]
Médica especialista en Terapia Intensiva Pediátrica. Hospital de Niños Ricardo Gutierrez.
Directora del Comité de Gestión, Calidad y Escores. Sociedad Argentina de Terapia Intensiva.
Ex presidente Sociedad Latinoamericana de Cuidados Intensivos Pediátricos (SLACIP).

Ariel Leonardo Fernández (@hardineros)
[Argentina]
Magister en Efectividad Clínica – Universidad de Buenos Aires.
Investigador y desarrollador. Especialista en Interoperabilidad, Data Management, Analítica y Sistemas de Información en Salud.
Responsable del Programa de Calidad SATI-Q: Quality Benchmarking en Unidades de Cuidados Intensivos, Sociedad Argentina de Terapia Intensiva (SATI).

Juan Sebastian Osorio (@JS-Osorio)
[Colombia]
Project Manager, MIT Critical Data – Sana MIT, Cambridge, MA.
Co-fundador Clubes de Ciencia Colombia y ScienteLab, Bogotá, Colombia.
Estudiante MPH en Salud Global, Universidad de Washington, Seattle, WA.

Colaboradores

Xavier Borrat Frigola (@xborrat)
[España]
Médico Especialista en Anestesiología y Reanimación. Unidad de Cuidados Intensivos Quirúrgica. (Hospital Clinic de Barcelona).

Doctor en modelización de efectos farmacológicos - UB (Universidad de Barcelona)
Profesor Asociado de Bioingenieria-UB.

Javier E. Camacho Cogollo (@biocamacho)
[Colombia]
Ingeniero Biomédico
MsC. Gestión de Innovación Tecnológica. PhD (c) Ingeniería
Docente investigador en Ingeniería Clínica e Informática Médica. Universidad EIA. Colombia.

Paula Caporal (@paucaporal)
[Argentina]
Médica especialista en Terapia Intensiva Pediátrica
Hospital Sor María Ludovica, La Plata, Buenos Aires, Argentina.
Jefa de Trabajos Prácticos del Curso Superior Terapia Intensiva Pediátrica de la Sociedad Argentina de Terapia Intensiva (SATI)

Pedro Ignacio Faedo (@pedrofaedo1)
[Argentina]
Médico - Facultad de Ciencias Médicas - Universidad de Buenos Aires (Argentina)
Ayudante de Farmacología - Facultad de Ciencias Médicas - Universidad de Buenos Aires (Argentina)

Paloma Gutierrez
[Argentina]
Médica, Facultad de Ciencias Médicas - Universidad de Buenos Aires (Argentina)
Ayudante de Salud Pública, Facultad de Ciencias Médicas - Universidad de Buenos Aires (Argentina)

Joaquín Lanuza
[Argentina]
Estudiante de Medicina en Facultad de Ciencias Médicas - Universidad de Buenos Aires (Argentina)

Diego M. López
[Colombia]
Profesor Titular, Departamento de Telemática, Universidad del Cauca, Popayán, Colombia.

Felipe Mejía Medina (@FelipeMejiaMV)
[Colombia]
Consultor en Salud Global y Desarrollo Social.

Ezequiel Monteverde
[Argentina]
Médico especialista en Cuidados Intensivos Pediátricos y en Estadística para Ciencias de la Salud.
Médico de staff Unidad de Cuidados Intensivos del Hospital de Niños Ricardo Gutiérrez.
Director del Registro de Trauma, Fundación Trauma

Malena Pilar Rosso (@Malena Rosso)
[Argentina]
Estudiante de Derecho - Universidad Torcuato Di Tella (Argentina)

David Bigio Roitman (@dbigio)
[Colombia]
Profesional Distinguido y Permanente. Departamento de Ingeniería Biomédica, Universidad de Los Andes. Bogotá, Colombia.

Sebastián Torres Montoya (@sebtomon)
[Colombia]
Estudiante doctorado en Universidad de California, Santa Cruz, CA.

AGRADECIMIENTOS

Deseamos agradecer al **Dr. Leo Anthony Celi**, por confiarnos este proyecto y por su incansable labor dirigida a promover una cultura de trabajo colaborativo entre médicos, informáticos, ingenieros y científicos de datos, que permita generar evidencia a partir de datos provenientes de la práctica médica rutinaria.

A **MIT CRITICAL DATA,** por ceder los derechos de publicación de este libro, con el objetivo de promover la capacitación de los científicos de Latinoamérica y España, en el análisis de los datos almacenados en los registros médicos electrónicos.

A **MIT International Science and Technology Initiatives (MISTI) y MIT Sloan LatinAmerica Office** por brindar el financiamiento que ha permitido la impresión de la obra.

A **Leandro N. Notarfrancesco** y **HARDINEROS SAS**, por su compromiso con el proyecto y por gestionar los aspectos administrativos que han permitido la publicación de este libro en Latinoamérica.

PREFACIO

Las tecnologías terapéuticas y diagnósticas continúan evolucionando rápidamente, y tanto los profesionales individuales como los equipos médicos se enfrentan a decisiones cada vez más complejas. Desafortunadamente, el estado de conocimiento científico actual no brinda orientación alguna para tomar la mayoría de las decisiones clínicas basándose en evidencia. De acuerdo con el Informe realizado por el Instituto de Medicina en el año 2012, solamente entre el 10% y el 20% de las decisiones clínicas están basadas en evidencia. El problema inclusive se extiende a la creación de guías de práctica clínica (GPCs). Cerca del 50% de las recomendaciones realizadas a partir de guías de sociedades especializadas confían más en opiniones de expertos que en datos obtenidos mediante investigaciones científicas. Además, el proceso de creación de GPCs está "afectado por métodos débiles y conflictos de interés financieros", haciendo a las GPCs potencialmente menos confiables.

La infraestructura actual de la investigación es ineficiente y con frecuencia produce resultados poco fiables y que, por eso, no pueden ser replicables. Inclusive los ensayos controlados aleatorizados (ECAs), los principales "*gold standard*" de la investigación tienen sus limitaciones. Los mismos pueden ser costosos, laboriosos, lentos y pueden devolver resultados que rara vez son generalizables para todos los pacientes. Es imposible, para un ECA estrictamente controlado, capturar los detalles completos, interactivos y contextuales de los problemas clínicos que surgen en el ámbito de atención ambulatoria y de internación. A su vez, muchos problemas clínicos importantes no resueltos no parecen haber atraído el interés necesario que incentive al ámbito de la investigación, que en cambio se ha centrado, en las investigaciones celulares y moleculares y efectos individuales (como por ejemplo, drogas o dispositivos). Para los médicos, el resultado final es un desierto de datos a la hora de tomar decisiones.

Las Historias Clínicas Electrónicas (HCEs) con frecuencia son almacenadas en forma digital y posteriormente pueden ser extraídas y analizadas. Entre el año 2011 y el año 2019, se espera que la prevalencia de uso de HCE

crezca de un 34% a un 90%, y que la mayoría de los hospitales ya hayan reemplazado, o bien se encuentren en el proceso de reemplazo, de los sistemas de registro en papel por HCEs. El poder de escalar, intrínseco de esta transformación digital abre la puerta a una cantidad masiva de información que se encuentra sin explotar. La información, si es analizada e interpretada de manera adecuada, podría mejorar enormemente nuestra concepción y el desarrollo de mejores prácticas. Las posibilidades de una mejora de calidad, mayor seguridad, optimización de procesos y personalización de las decisiones clínicas son desde impresionantes hasta revolucionarias. El National Health Institute (NHI) y otras importantes organizaciones han empezado a reconocer el poder del "big data" para la creación de conocimiento y han comenzado a ofrecer subvenciones para apoyar a los investigadores en esta área.

Este libro, escrito con el apoyo del National Institute for Biomedical Imaging and Bioengineering de Estados Unidos de América (EUA) por medio de la beca R01 EB017205-01A1, tiene el objetivo de funcionar como una guía ilustrativa para los científicos, ingenieros, y médicos que estén interesados en realizar investigaciones retrospectivas utilizando información extraída de HCEs. La obra está dividida en tres grandes partes.

La primer parte muestra la situación actual y describe el cuerpo de conocimientos que dictan las guías de práctica clínica, incluyendo sus limitaciones y los desafíos. Esto sienta las bases para presentar la motivación detrás del análisis secundario de datos de las HCEs. También describe el panorama de los datos, quiénes son los actores principales y qué tipo de bases de datos son las más útiles para cada tipo de preguntas. Finalmente, se delinean los desafíos políticos, regulatorios y técnicos que enfrentan los informáticos clínicos, además de proponer sugerencias para sortearlos.

En la segunda parte, el proceso de análisis de una pregunta clínica y su conversión en un diseño y metodología de estudio se divide en cinco pasos. El primer paso explica cómo formular la pregunta de investigación correcta y cómo seleccionar el equipo de trabajo adecuado. El segundo paso esboza las estrategias para la identificación, extracción y pre-procesamiento de la información de las HCEs para comprender y abordar las cuestiones de interés de la investigación.

El tercer paso presenta técnicas en análisis exploratorio y visualización de datos. En el cuarto paso, se provee una guía detallada sobre cómo elegir el tipo de análisis que mejor responde y se condice con la pregunta de investigación. Finalmente, el quinto paso describe cómo validar los resultados, usando validaciones cruzadas, análisis de sensibilidad, pruebas de hipótesis nula y otras técnicas comunes en el campo.

La tercer y última parte del libro ofrece una colección integral de análisis de casos. Estos estudios de casos destacan líneas de investigación presentadas en la segunda parte del libro, y ayuda a situar al lector en el mundo real del análisis de datos.

Hemos escrito este libro para que lectores que se encuentran en diferentes niveles, puedan fácilmente comenzar su lectura en partes diferentes. En el caso de los investigadores novatos, el libro debería leerse desde el inicio al fin. Los individuos que ya están familiarizados con los desafíos de la informática en salud, pero desearían una orientación para realizar análisis en forma más efectiva, deberían leer el libro a partir de la segunda parte. Finalmente, la parte del estudio de casos provee consideraciones prácticas específicas sobre el diseño y metodología de estudio y es recomendada para todos los lectores.

Ha llegado el momento de aprovechar los datos que generamos durante la atención habitual de los pacientes para formular un compendio más completo de recomendaciones basadas en evidencias y para apoyar a la toma de decisiones compartida con los pacientes. Este libro va a entrenar a la próxima generación de científicos, representando diferentes disciplinas pero colaborando para expandir la base de conocimiento que guiará la práctica médica en el futuro.

Nos gustaría aprovechar esta oportunidad para agradecerle al profesor Roger Mark, cuya visión de crear una base de datos clínicos de alta calidad abierta a investigadores de todo el mundo nos ha inspirado para escribir este libro.

Cambridge, EUA.

MIT CRITICAL DATA

MIT Critical Data es en un grupo de científicos de datos y médicos provenientes de todo el mundo y reunidos por la visión en común de generar un sistema de salud basado en datos, apoyado por la *informática en salud sin fronteras.* En este ecosistema, la creación de evidencia y de herramientas para apoyar las decisiones clínicas se inicia, actualiza, mejora y perfecciona al ampliar el acceso y el uso significativo de los datos clínicos.

Leo Anthony Celi ha ejercido la medicina en tres continentes, lo cual le ha dado una amplia perspectiva sobre la asistencia médica. Su investigación se basa en el análisis secundario de las historias clínicas electrónicas y la informática en salud global. Fundó y codirige "Sana" en el Instituto de Ingeniería y Ciencias Médicas, del Instituto de Tecnología de Massachusetts (MIT). También ocupa un puesto de profesor en la Escuela de Medicina de Harvard como médico especialista en Cuidados Intensivos en el Centro Médico Beth Israel Deaconess, además de ser el director de investigación clínica para el Laboratorio de Fisiología Computacional del MIT. Finalmente, es uno de los directores del curso HST.936: análisis secundario de las historias clínicas electrónicas del MIT (Innovación en informática en salud global y HST.953).

Peter Charlton obtuvo un Máster en Ingeniería Científica en el año 2010, de la Universidad de Oxford. Desde entonces, ha ocupado cargos de investigador, trabajando conjuntamente con la Fundación Guy's y St. Thomas NHS y con el King's College London. La investigación de Peter se centra en el monitoreo fisiológico de pacientes hospitalizados, dividido en tres áreas. La primera se ocupa del desarrollo de técnicas de procesamiento de señales para estimar parámetros clínicos a partir de señales fisiológicas. Se ha enfocado en la estimación no invasiva de la frecuencia respiratoria para su uso en entornos ambulatorios, en la estimación invasiva del gasto cardiaco para su uso en cuidados intensivos y en el desarrollo de nuevas técnicas de análisis de señales de oximetría de pulso (fotopletismografía). En segundo lugar, investiga la efectividad de tecnologías para la adquisición de mediciones fisiológicas continuas e intermitentes en el ámbito de

cuidados intensivos y ambulatorio. En tercer lugar, está desarrollando técnicas para transformar el monitoreo continuo de datos en medidas apropiadas para crear alertas en tiempo real del deterioro de pacientes.

Mohammad Mahdi Ghassemi es candidato a un doctorado en el Instituto de Tecnología de Massachusetts. Como estudiante de pregrado, estudió ingeniería electrónica y se graduó como erudito en Goldwater y como "Ingeniero sobresaliente en la Universidad". En el año 2011, Mohammad recibió un MPhil en Ingeniería Informática de la Universidad de Cambridge, donde también recibió la beca Gates-Cambridge. Desde que llegó al MIT, ha realizado investigaciones en la interfaz del aprendizaje automático y la informática médica. El enfoque doctoral de Mohammed se centra en el procesamiento de señales y técnicas de aprendizaje automático en el contexto de set de datos multimodales y multiescala. Ha colaborado en organizar la colección más grande de EEGs de coma post-anóxico en el mundo. Además, Mohammad ha trabajado con Samsung Corporation y varias entidades en todo el campus construyendo "dispositivos inteligentes "entre los que se encuentra: un dispositivo portátil con múltiples sensores que monitorea de forma pasiva la actividad fisiológica, de audio y video de un usuario para estimar un estado emocional latente.

Alistair Johnson recibió su Licenciatura en Ingeniería Biomédica y Electrónica en McMaster University, Canadá y posteriormente realizó un Posgrado en Filosofía en Innovación en Salud en la Universidad de Oxford. Su tesis fue titulada "Mortalidad y evaluación de la agudeza en cuidados críticos" y su enfoque incluyó el uso de técnicas de aprendizaje automático para predecir la mortalidad y desarrollar nuevos puntajes de gravedad de enfermedad para pacientes ingresados en unidades de cuidados intensivos. Alister también pasó un año como asistente de investigaciones en el Hospital John Radcliffe en Oxford, donde trabajó en la construcción de modelos de alerta temprana para pacientes después del alta de la Unidad de Cuidados Intensivos. Los intereses de investigación de Alistair giran en torno al uso de datos registrados durante la práctica clínica habitual para mejorar la atención del paciente.

Matthieu Komorowski posee Certificaciones en Anestesiología y Cuidados Intensivos tanto en Francia como en Reino Unido. Es un ex investigador médico de la Agencia Espacial Europea, completó un máster en ingeniería biomédica centrado en el aprendizaje automático. El Dr.

Komorowski actualmente realiza un Ph.D. en el Imperial College y una beca de investigación de cuidados intensivos en el Hospital Charing Cross en Londres. En su investigación, combina su experiencia en aprendizaje automático y en cuidados intensivos para generar nueva evidencia clínica y así construir la próxima generación de herramientas clínicas como sistemas de soporte de decisiones, con un interés particular en el shock séptico, el cuadro con mayor mortalidad en cuidados intensivos y la condición más costosa tratada en hospitales.

Dominic Marshall es médico en la Fundación Académica en Oxford, Reino Unido. Dominic estudió Biología Molecular y Celular en la Universidad de Bath y trabajó en Eli Lily en su programa de investigación para búsqueda de drogas para la enfermedad del Alzheimer. Él completó su entrenamiento médico en el Imperial College London, donde ganó la beca Santander Undergraduate Schollarship por su desempeño académico y ocupó el primer lugar en su clase al graduarse. Sus intereses de investigación incluyen desde la biología molecular hasta el análisis de grandes set de datos clínicos y ha recibido fondos no provenientes de la industria para poder desarrollar nuevos antibióticos y agentes quimioterapéuticos. Al mismo tiempo que en el entrenamiento médico, se encuentra involucrado en un gran número de proyectos de investigación enfocados en el análisis de las historias clínicas electrónicas.

Tristan Naumann es candidato a un Doctorado en Ingeniería Electrónica y Ciencias de la Computación en el MIT, trabajando con el Dr. Peter Szolovits en el grupo de Toma de Decisiones Médicas CSAIL. Su investigación incluye la exploración de relaciones en datos complejos no estructurados utilizando técnicas de aprendizaje no supervisadas y la aplicación de técnicas de procesamiento del lenguaje natural en datos en salud. Ha sido organizador de talleres y “datathones”, los cuales reúnen a participantes con diversa formación laboral y académica para abordar preguntas biomédicas y clínicas de una manera confiable y reproducible.

Kenneth Paik es un informático clínico que democratiza el”acceso a la asistencia de la salud “mediante la innovación tecnológica, con su formación multidisciplinaria en medicina, inteligencia artificial, gestión de negocios y estrategia tecnológica. Es un científico investigador en el Laboratorio de Fisiología Computacional del MIT, donde investiga el análisis secundario de datos de salud y construye un sistema inteligente de soporte

de decisiones. Como codirector de Sana, dirige programas y proyectos que impulsan la mejora de la calidad y el desarrollo de capacidades en torno a la salud global. Recibió su Doctorado en Medicina y su Máster en Administración de Empresas de la Universidad de Georgetown y completó su entrenamiento en informática biomédica en la Escuela de Medicina de Harvard y en el Laboratorio de Ciencias de la Computación del Hospital General de Massachusetts.

Tom Joseph Pollard es un asociado postdoctoral en el Laboratorio de Fisiología Computacional del MIT. Más recientemente ha trabajado con sus colegas para desarrollar MIMIC-III, una base de datos de cuidados intensivos de código abierto. Antes de unirse al MIT en el año 2015, Tom completó su Doctorado (Ph.D.) en el University College, Londres, Reino Unido, donde exploró modelos de salud en pacientes de cuidados intensivos en un proyecto interdisciplinario entre Mullard Space Science Laboratory y University College Hospital. Tom tiene un gran interés en mejorar la forma en que los datos clínicos son manejados, compartidos y analizados, para el beneficio de los pacientes. Es un becario del Instituto de Sustentabilidad de Software.

Jesse Raffa es investigador científico del Laboratorio de Fisiología Computacional del MIT en Cambridge, Estados Unidos. Ha recibido su Doctorado (Ph.D.) en Bioestadística en la Universidad de Waterloo (Canadá) en el año 2013. Sus intereses metodológicos principales están relacionados con el modelado de datos longitudinales complejos, modelos de variables latentes e investigaciones reproducibles. Además de sus contribuciones metodológicas, él ha colaborado y publicado más de 20 artículos académicos con colegas en una gran diversidad de áreas, las que incluyen enfermedades infecciosas, adicciones y cuidados intensivos, entre otros. Jesse recibió el Premio al artículo científico estudiantil distinguido en la conferencia de la Eastern North American Region International Biometric Society en el año 2013 y el premio al Nuevo Investigador del Año por The Canadian Association of HIV/AIDS Research en el año 2004.

Justin Salciccioli es médico de la Academic Foundation en Londres, Reino Unido. Oriundo de Toronto, Canadá, Justin completó sus estudios de grado y posgrado en los Estados Unidos antes de realizar sus estudios en Medicina en el Imperial College de Londres. Sus investigaciones comenzaron cuando era un estudiante de pregrado mientras completaba su formación en

bioquímica. Posteriormente, trabajó en ensayos clínicos en medicina de emergencia y medicina de cuidados intensivos en el Beth Israel Deaconess Medical Center en Boston y completó un Master con una tesis cuyo foco era la falta de vitamina D en pacientes enfermos de sepsis que se encontraban en estado crítico. En el último tiempo ha desarrollado un gran interés en métodos estadísticos y programación, particularmente en SAS y R. Es coautor de más de 30 manuscritos revisados por pares y, además de su formación clínica actual, continúa con su interés en la investigación sobre métodos analíticos para datos de ensayos clínicos y observacionales así como educación en análisis de datos para estudiantes de medicina y médicos.

ÍNDICE

PARTE I

CREANDO EL ESCENARIO:
FUNDAMENTO Y DESAFÍOS SUBYACENTES
EN EL ANÁLISIS DE DATOS EN SALUD

INTRODUCCIÓN

Mientras que todos los días aparecen en las noticias maravillosos descubrimientos e innovaciones médicas, aquellos que proveen la atención sanitaria luchan continuamente con el uso de la información. La incertidumbre y las preguntas clínicas sin respuesta son una realidad diaria para los decisores responsables de brindar asistencia sanitaria. Es posible que la mayor limitación para tomar las mejores decisiones para los pacientes sea que la información disponible no se encuentra enfocada en cada individuo o situación.

Existen, por ejemplo, guías de práctica clínica que establecen el valor ideal de presión arterial para un paciente con una infección severa. Sin embargo, el valor de presión arterial óptimo difiere sustancialmente de paciente a paciente e incluso puede variar para un mismo paciente en distintos momentos del tratamiento. La creciente informatización de los registros clínicos representa una oportunidad para superar esta limitación. Al analizar los datos electrónicos de las experiencias de muchos profesionales de la salud con muchos pacientes podemos acercarnos a responder la eterna pregunta: ¿qué es verdaderamente mejor para cada paciente?

El análisis secundario de los datos recolectados rutinariamente –en contraste con el análisis primario conducido durante el proceso de cuidado de cada paciente individual– ofrece una oportunidad para extraer mayor conocimiento que nos guiará hacia el objetivo del cuidado óptimo. Un informe de la Academia Nacional de Medicina nos dice hoy que la mayoría de los médicos basan sus decisiones diarias en guías basadas en opiniones de expertos (a veces sesgadas) o en ensayos clínicos de pequeño tamaño muestral. Sería mejor si estas guías se basaran en estudios grandes, multicéntricos y aleatorizados, con condiciones controladas de ejecución que garantizaran que los resultados tuvieran la mayor confiabilidad posible Sin embargo, este tipo de estudios son costosos y difíciles de realizar, y aún así excluyen una cantidad importante de grupos de pacientes en base a su edad, enfermedad y condiciones sociales.

Parte del problema tiene que ver con que los registros clínicos suelen almacenarse en papel, haciendo difícil su análisis en grandes cantidades. Como resultado, la mayoría de lo que los profesionales de la salud pudieran haber aprendido de su experiencia se pierde o es, como mínimo, inaccesible. El sistema digital ideal debería recolectar y almacenar la mayor cantidad de información clínica de la mayor cantidad de pacientes posible para luego utilizar esa información del pasado –como la presión arterial, glucemia, frecuencia cardíaca y otros parámetros de funciones vitales– para guiar a futuros profesionales de la salud en la toma de decisiones diagnósticas y terapéuticas en pacientes similares.

Sin embargo, el "big data" en salud ha estado marcado por la "Superfluidad de Sillicon Valley", el lenguaje mediante el que Sillicon Valley endulza y llena de grandes promesas el campo para atraer inversores y futuros usuarios. La moda de la "medicina de precisión" llega a los oídos del público con escasas menciones sobre los fracasos de la "medicina personalizada", su antecesora.

Esta parte del libro configura el escenario del análisis secundario de las historias clínicas electrónicas (HCE). El capítulo 1 comienza con la racionalidad subyacente a este tipo de investigación. El capítulo 2 provee una lista de bases de datos clínicos ya existentes y ya en uso para investigación. El capítulo 3 se sumerge en las oportunidades y, más importante aún, en los desafíos que existen en el análisis retrospectivo de las HCE. El capítulo 4 presenta ideas acerca de cómo la información podría ser aplicada de manera sistemática y más efectiva en un sistema sanitario diseñado con tal fin. El profesor Roger Mark, el visionario que creó la base de datos MIMIC (del inglés, Medical Information Mart for Intensive Care) usada en este libro, cuenta la historia detrás del proyecto en el capítulo 5. El capítulo 6 se sumerge en el futuro y describe la integración de las HCE con información no relacionada con la clínica para una representación más rica del proceso salud-enfermedad. El capítulo 7 se enfoca en el rol de la HCE en dos áreas importantes de investigación –resultados y servicios de salud–.

Finalmente, el capítulo 8 enfrenta la ruina de los estudios observacionales que utilizan las HCE: la confusión residual.

Enfatizamos sobre la importancia de integrar profesionales clínicos de cabecera, como enfermeros, farmacéuticos y médicos con científicos de datos para identificar las preguntas adecuadas y conducir el análisis

apropiado en forma colaborativa. Aún más, creemos que esta asociación para la investigación del personal asistencial y los investigadores provee a los cuidadores y a los pacientes las mejores opciones individualizadas de diagnóstico y tratamiento ante la ausencia de un ensayo clínico randomizado. Al trabajar y familiarizarnos con la información disponible hoy en los hospitales, podemos reducir las incertidumbres que han obstaculizado el cuidado de la salud durante demasiado tiempo.

Capítulo 1

OBJETIVOS DEL ANÁLISIS SECUNDARIO DE LOS DATOS DE LAS HISTORIAS CLÍNICAS ELECTRÓNICAS

Sharukh Lokhandwala y Barret Rush

Puntos clave

- La medicina clínica se sustenta en una robusta investigación que permita construir la evidencia necesaria, base para informar las mejores prácticas y optimizar los cuidados médicos. Sin embargo, los ensayos clínicos aleatorizados (ECA) de gran tamaño muestral son costosos y a veces inviables. Afortunadamente, existe una gran cantidad de información disponible en la forma de historias clínicas electrónicas (HCE).
- La información puede ser abrumadoramente compleja o incompleta para cualquier individuo. Por lo tanto, requerimos en forma urgente de equipos de investigación multidisciplinarios conformados por clínicos y científicos de datos para desentrañar el lenguaje médico y permitir un análisis apropiado de los datos.

1.1 Introducción

La industria de la salud rápidamente se ha convertido en informatizada y digital. La mayoría de los servicios de salud provistos hoy en América dependen de o utilizan tecnología. La informática en salud moderna genera y almacena enormes cantidades de información detallada sobre los pacientes y los procesos clínicos. Es escasa la cantidad de información del mundo real de los pacientes que se ha utilizado para avanzar en el campo de la salud. Una importante barrera para la utilización de estos datos es la inaccesibilidad para los investigadores. Facilitar el acceso a estas bases de datos, así como integrar la información permitiría que más investigadores respondieran preguntas fundamentales del cuidado de la salud.

1.2 Estado actual de la investigación

Muchos tratamientos carecen de evidencia sobre su eficacia e incluso pueden causar daño [1]. Diversas sociedades médicas difunden guías para asistir en la toma de decisiones clínicas y para estandarizar las prácticas; sin

embargo la evidencia utilizada para formular estas guías es inadecuada. Dichas guías también derivan comúnmente de ECA realizados en cohortes de pacientes limitadas, utilizando estrictos criterios de inclusión y exclusión, y cuyos resultados pueden no ser generalizables. Los ECA, "gold standard" en la investigación clínica, respaldan únicamente del 10 al 20% de las decisiones médicas [2] y la mayoría de las decisiones nunca han sido respaldadas por los mismos [3]. Aún más, sería imposible realizar un ECA para cada una de las innumerables decisiones que los médicos enfrentan día a día en el cuidado de los pacientes, por numerosas razones, incluyendo la limitada disponibilidad de recursos financieros y humanos. Por este motivo, los médicos e investigadores deben aprender a encontrar evidencia clínica de las fuentes de datos que ya existen: las HCEs.

1.3 El poder de las Historias Clínicas Electrónicas

Muchos de los trabajos que han utilizado grandes bases de datos en los últimos 25 años se han basado en bases de datos administrativas (registros de alta hospitalaria) y en bases de datos de Registros Clínicos. Las bases de datos de altas hospitalarias fueron creadas inicialmente con fines de facturación y carecen de la granularidad de información clínica a nivel paciente, clínicamente útil, precisa y completa que permita responder preguntas de investigación complejas. Las bases de datos de registros clínicos, por su parte, suelen contener información limitada, misión dependiente y requieren de una exhaustiva recolección de datos extracurricular. El futuro de la investigación clínica está en la utilización de "big data" para mejorar el cuidado de los pacientes.

Aunque se han creado varias bases de datos comerciales y no comerciales utilizando información clínica y contenida en HCE, su función primaria ha sido analizar diferencias en el grado de enfermedad, resultados y costos de tratamientos entre los centros participantes. Se han creado registros específicos para enfermedades como lesión renal aguda [4], síndrome de distress respiratorio agudo [5] y shock séptico [6]. Adicionalmente, bases de datos como la Dartmouth Atlas utiliza información de los reclamos de Medicare para rastrear discrepancias entre costos y resultados a lo largo de los Estados Unidos [7]. Mientras que estas bases de datos coordinadas contienen una gran cantidad de pacientes,

suelen tener un alcance acotado (es decir, para severidad de enfermedad, costos o resultados específicos por enfermedad) y carecen de otra información clínica relevante necesaria para responder un amplio espectro de preguntas de investigación, ocultando muchas probables variables confusoras.

Por ejemplo, la base de datos “APACHE Outcomes” fue creada uniendo “APACHE (Acute Physiology and Chronic Health Evaluation)” [8] con el Proyecto IMPACT [9] e incluye información de aproximadamente 150.000 admisiones en unidades de cuidados intensivos (UCI) desde el año 2010 [1]. Si bien la base de datos “APACHE Outcomes” es grande y ha contribuido significativamente a la literatura médica, las mediciones fisiológicas y de laboratorio son incompletas y no incluye notas clínicas o datos de sensores. La base de datos del Programa Phillips eICU [10], proveedor de soporte en cuidados intensivos mediante telemedicina, contiene información sobre 2 millones de admisiones en UCI. Aunque incluye la documentación ingresada por los profesionales de la salud en el software, no cuenta con notas clínicas o datos de sensores. Además, considerando que las bases de datos con distintos objetivos primarios (por ej.: costos, mejoras de calidad o investigación) se focalizan en distintas variables y resultados, hay que tener cuidado al interpretar las conclusiones de su análisis.

Desde el año 2003, el Laboratorio de Fisiología Computacional del Instituto de Tecnología de Massachusetts se asoció con el Centro Médico Beth Israel Deaconess y con Philips Healthcare, apoyados por el Instituto Nacional de Imágenes Biomédicas y Bioinformáticas (NIBIB, del inglés National Institute of Biomedical Imaging and Bioinformatics) para desarrollar y mantener la base de datos MIMIC (del inglés, Medical Information Mart for Intensive Care) [11]. MIMIC es una base de datos de acceso público que contiene información clínica exhaustiva de más de 60.000 admisiones de pacientes a UCIs en el Centro Médico Beth Israel Deaconess. La información desidentificada se comparte libremente y a la fecha la han utilizado cerca de 2.000 investigadores de 32 países. MIMIC contiene tanto datos fisiológicos como de laboratorio, datos de sensores, información numérica verificada por enfermería y documentación clínica. Esta base de datos de alta resolución y accesibilidad ha servido para conducir investigaciones en cuidados críticos y para asistir en el desarrollo

de nuevos algoritmos de soporte para la toma de decisiones, y servirá de modelo para la mayor parte de este libro.

1.4 Obstáculos y desafíos

Los médicos y científicos de datos deben aplicar el mismo nivel de rigor académico al analizar investigaciones realizadas a partir de bases de datos que el que aplican cuando realizan los métodos tradicionales de investigación clínica. Para asegurar la validez interna y externa, los investigadores deben determinar si los datos son precisos y si están ajustados adecuadamente, correctamente analizados y presentados convincentemente [12]. En relación a los proyectos de mejora de calidad, que frecuentemente utilizan bases de datos de hospitales, es necesario tener la seguridad de que los investigadores están aplicando rigurosos estándares en el desarrollo y presentación de sus de investigación [13].

A pesar del gran valor que contienen las HCE, muchos investigadores dudan en utilizar todas sus capacidades debido a su gran complejidad y a la imposibilidad de utilizar métodos tradicionales de procesamiento de datos con grandes volúmenes de datos. Como solución para la creciente complejidad asociada a este tipo de investigaciones, sugerimos que los investigadores trabajen en colaboración con equipos multidisciplinarios que incluyan científicos de datos, médicos y bioestadísticos. Esto puede requerir un cambio en los incentivos financieros y académicos para que los grupos de investigación individuales no compitan por financiamiento o publicaciones; los incentivos deberían así promover la autoría y financiación conjunta. Esto permitiría a los investigadores focalizarse en la fidelidad de su trabajo y estar más dispuestos a compartir sus datos para otras investigaciones en lugar de retenerlos para ser los “primeros” en arribar a una solución.

Algunos han argumentado que el uso de grandes volúmenes de datos puede aumentar la frecuencia del llamado “p-hacking” en el que investigadores buscan resultados significantes en lugar de perseguir respuestas a preguntas relevantes clínicamente. Si bien parece que el “p-hacking” es una práctica extendida, el alcance del efecto que se le puede atribuir no suele socavar los resultados científicos de grandes estudios y metaanálisis. El uso de estos grandes volúmenes de datos puede, en efecto,

reducir la probabilidad del “p-hacking” al asegurar a los investigadores la capacidad de responder preguntas incluso con efectos de pequeño tamaño, haciendo innecesaria la interpretación y análisis selectivo de los datos para obtener resultados significativos. Si se realizan importantes descubrimientos utilizando grandes bases de datos, dicho trabajo puede ser utilizado como base de ensayos clínicos rigurosos que confirmen tales descubrimientos. En el futuro, una vez que las grandes bases de datos sean más accesibles para los investigadores, se espera que estos recursos puedan ser utilizados como generadores de hipótesis y escenarios de prueba para preguntas que en última instancia serán sometidas a ECAs. Si no hay fuertes indicios observados en grandes estudios retrospectivos preliminares, proseguir con un ECA que consuma tiempo y recursos puede no ser aconsejable.

1.5 Conclusiones

Con los avances en la recolección de datos y en la tecnología, los investigadores tienen más acceso a información sobre pacientes que en cualquier otro momento de la historia. Actualmente, muchos de esos datos se encuentran inaccesibles y son subutilizados. La habilidad de aprovechar las HCEs permitiría desarrollar sistemas de aprendizaje continuo, en los que datos específicos de un paciente podrían ser ingresados a bases de datos poblacionales obteniendo apoyo para tomar decisiones en tiempo real para ese paciente individual basadas en casos y escenarios similares. Tanto los médicos como los pacientes podrían tomar mejores decisiones con dichos recursos y los resultados volverían a ser ingresados a la misma base de datos poblacional [14].

La gran cantidad de información disponible para médicos y científicos plantea enormes desafíos así como también grandes oportunidades. La Academia Nacional de Medicina realizó una convocatoria entre médicos e investigadores para crear sistemas que “fomenten el aprendizaje continuo, de modo que las lecciones de cada investigación y de cada experiencia de cuidado sean sistemáticamente capturadas, evaluadas y traducidas en cuidados seguros” [2]. Para capturar, evaluar y traducir estos datos debemos utilizar el poder de las HCEs para crear repositorios de datos, al mismo tiempo que brindar a los médicos y pacientes herramientas de

soporte para la toma de decisiones basadas en datos que permitan mejor atención de los pacientes en la práctica diaria.

Referencias

1. Celi LA, Mark RG, Stone DJ, Montgomery RA (2013) “Big data” in the intensive care unit. Closing the data loop. Am J Respir Crit Care Med 187:1157-1160.
2. Smith M, Saunders R, Stuckhardt L, McGinnis JM (2013) Best care at lower cost: the path to continuously learning health care in America. National Academies Press.
3. Mills EJ, Thorlund K, Ioannidis JP (2013) Demystifying trial networks and network meta-analysis. BMJ 346: f2914.
4. Mehta RL, Kellum JA, Shah SV, Molitoris BA, Ronco C, Warnock DG, Levin A, Acute Kidney Injury N (2007) Acute Kidney Injury Network: report of an initiative to improve outcomes in acute kidney injury. Crit Care 11: R31.
5. The Acute Respiratory Distress Syndrome Network (2000) Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. N Engl J Med 342:1301-1308.
6. Dellinger RP, Levy MM, Rhodes A, Annane D, Gerlach H, Opal SM, Sevransky JE, Sprung CL, Douglas IS, Jaeschke R, Osborn TM, Nunnally ME, Townsend SR, Reinhart K, Kleinpell RM, Angus DC, Deutschman CS, Machado FR, Rubenfeld GD, Webb SA, Beale RJ, Vincent JL, Moreno R, Surviving Sepsis Campaign Guidelines Committe including the Pediatric S (2013) Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2012. Crit Care Med 41:580-637.
7. The Dartmouth Atlas of Health Care. Lebanon, NH. The Trustees of Dartmouth College 2015. Consultado 10 Julio 2015. Disponible en http://www.dartmouthatlas.org/.

8. Zimmerman JE, Kramer AA, McNair DS, Malila FM, Shaffer VL (2006) Intensive care unit length of stay: Benchmarking based on Acute Physiology and Chronic Health Evaluation (APACHE) IV. Crit Care Med 34:2517-2529.
9. Cook SF, Visscher WA, Hobbs CL, Williams RL, Project ICIC (2002) Project IMPACT: results from a pilot validity study of a new observational database. Crit Care Med 30:2765-2770.
10. eICU Program Solution. Koninklijke Philips Electronics N. V, Baltimore, MD (2012).
11. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L-W, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG (2011) Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database. Crit Care Med 39:952.
12. Meurer S (2008) Data quality in healthcare comparative databases. MIT Information Quality Industry symposium.
13. Davidoff F, Batalden P, Stevens D, Ogrinc G, Mooney SE, group Sd (2009) Publication guidelines for quality improvement studies in health care: evolution of the SQUIRE project. BMJ 338: a3152.
14. Celi LA, Zimolzak AJ, Stone DJ (2014) Dynamic clinical data mining: search engine-based decisión support. JMIR Med Informatics 2: e13.

Capítulo 2

UNA REVISIÓN DE BASES DE DATOS MÉDICAS

Jeff Marshall, Abdullah Chahin
y Barret Rush

Puntos clave

- Existen diversas bases de datos de acceso libre que promueven efectivas investigaciones retrospectivas de efectividad comparativa.
- Estos set de datos contienen un número variable de datos con variables representativas, propicias para tipos específicos de investigaciones y poblaciones. Comprender las características de cada base de datos particular será crucial para una elaboración apropiada de las conclusiones de la investigación.

2.1 Introducción

Desde la aparición de la primera historia clínica electrónica (HCE) en los años 60, la información sobre pacientes se acumuló durante décadas sin una estructura clara que la hiciera significativa y utilizable. Con el tiempo, las instituciones comenzaron a implementar bases de datos que archivaban y organizaban la información en repositorios centrales. Los hospitales fueron entonces capaces de combinar información de servicios auxiliares, incluyendo farmacias, laboratorios y estudios por imágenes, con varios componentes de la asistencia clínica (como planes de enfermería, registros de administración de medicación y prescripciones médicas). En este capítulo presentamos al lector diversas bases de datos grandes que son de acceso público o rápidamente accesibles con escasa dificultad. A medida que avanza la investigación en salud utilizando grandes volúmenes de datos, es probable que se pueda acceder a otras fuentes de datos de acceso libre, en un entorno de código abierto.

2.2 Antecedentes

Inicialmente, las HCE fueron diseñadas para archivar y organizar los registros médicos de los pacientes. Luego fueron codificadas para facturación y con el propósito de mejora de calidad. Con el tiempo, las bases

de datos generadas a partir de las HCEs se fueron volviendo más detalladas, dinámicas e interconectadas. Sin embargo, la industria médica ha quedado atrás con respecto a otras industrias en el uso de "big data". Las investigaciones que utilizan estas grandes bases de datos se han visto obstaculizadas por la pobre calidad y organización de la información recopilada. En la actualidad, los datos médicos evolucionaron más allá de los meros registros clínicos, brindando la oportunidad de un análisis más detallado. Tradicionalmente, la investigación médica se ha basado en los registros de enfermedad o sistemas de gestión de enfermedades crónicas. Estos repositorios son conjuntos de datos recolectados *a priori* que suelen ser específicos para una enfermedad. Son incapaces de trasladar datos o conclusiones a otras enfermedades y suelen contener datos de cohortes de pacientes de un área geográfica en particular, limitando su generalización.

A diferencia de los registros de enfermedades, las HCEs contienen mayor cantidad de variables y mejor calidad de datos, lo que resulta ideal para estudiar interacciones y decisiones clínicas complejas. Esta nueva fuente de conocimientos integra varios set de datos que actualmente se encuentran completamente informatizados y accesibles. Desafortunadamente, la gran mayoría de las grandes bases de datos en el ámbito sanitario recolectadas alrededor del mundo, restringen el acceso a los datos. Algunas explicaciones posibles para estas restricciones involucran cuestiones de privacidad, intenciones de obtener rédito económico de la información y cierta reticencia a que investigadores externos tengan acceso directo a información respecto a la calidad de atención de un determinado establecimiento. Existe un movimiento creciente que busca hacer estos repositorios de acceso libre y abierto a investigadores.

2.3 La base de datos MIMIC (del inglés, Medical Information Mart for Intensive Care)

La base de datos MIMIC (http://mimic.physionet.org) se estableció en octubre del año 2003 como una colaboración de investigación en bioingeniería entre MIT, Philips Medical Systems y el Centro Médico Beth Israel Deaconess. El proyecto está financiado por el Instituto Nacional de Imágenes Biomédicas y Bioingeniería [1].

Esta base de datos se originó a partir de todos los ingresos médicos y quirúrgicos de todas las Unidades de Cuidados Intensivos (UCI) del Centro

Médico Beth Israel Deaconess, un hospital universitario, urbano y de tercer nivel. La tercera versión de la base de datos, MIMIC-III, contiene más de 40 mil pacientes con miles de variables. La base de datos está desidentificada, comentada y disponible en forma abierta para la comunidad de investigadores. Sumada a la información de los pacientes procedente del hospital, la base de datos MIMIC-III contiene información fisiológica y clínica detallada [2]. Además de investigación en "big data" en cuidados críticos, el proyecto tiene como objetivo desarrollar y evaluar sistemas avanzados de monitoreo de pacientes y de soporte a la toma de decisiones que mejoren la eficacia, precisión y oportunidad de las decisiones clínicas en cuidados críticos.

A través de la minería de datos, esta base de datos permite realizar numerosos estudios epidemiológicos que vinculan información de los pacientes con diferentes prácticas clínicas y resultados. El elevado grado de detalle de los datos brinda la posibilidad de realizar análisis minuciosos de problemas clínicos complejos.

2.3.1 Variables incluidas

Esencialmente, hay dos grandes tipos de datos en la base MIMIC-III; datos clínicos extraídos de la HCE como los datos demográficos, diagnósticos, valores de laboratorio, informes de imágenes, signos vitales, etc. (Fig. 2.1). Estos datos son almacenados en una base de datos relacional de aproximadamente 50 tablas. El segundo tipo de datos primario comprende las curvas procedentes de los monitores con sus parámetros asociados y eventos almacenados en archivos binarios planos (con descriptores de encabezado en ASCII). Esta librería única incluye datos de alta resolución obtenidos a partir de electroencefalogramas (EEGs), electrocardiogramas (ECGs) y registro en tiempo real de los signos vitales de los pacientes en UCI. El Comité de Ética en Investigación determinó que era necesario un consentimiento individual de cada paciente ya que toda la información pública se encontraba "desidentificada".

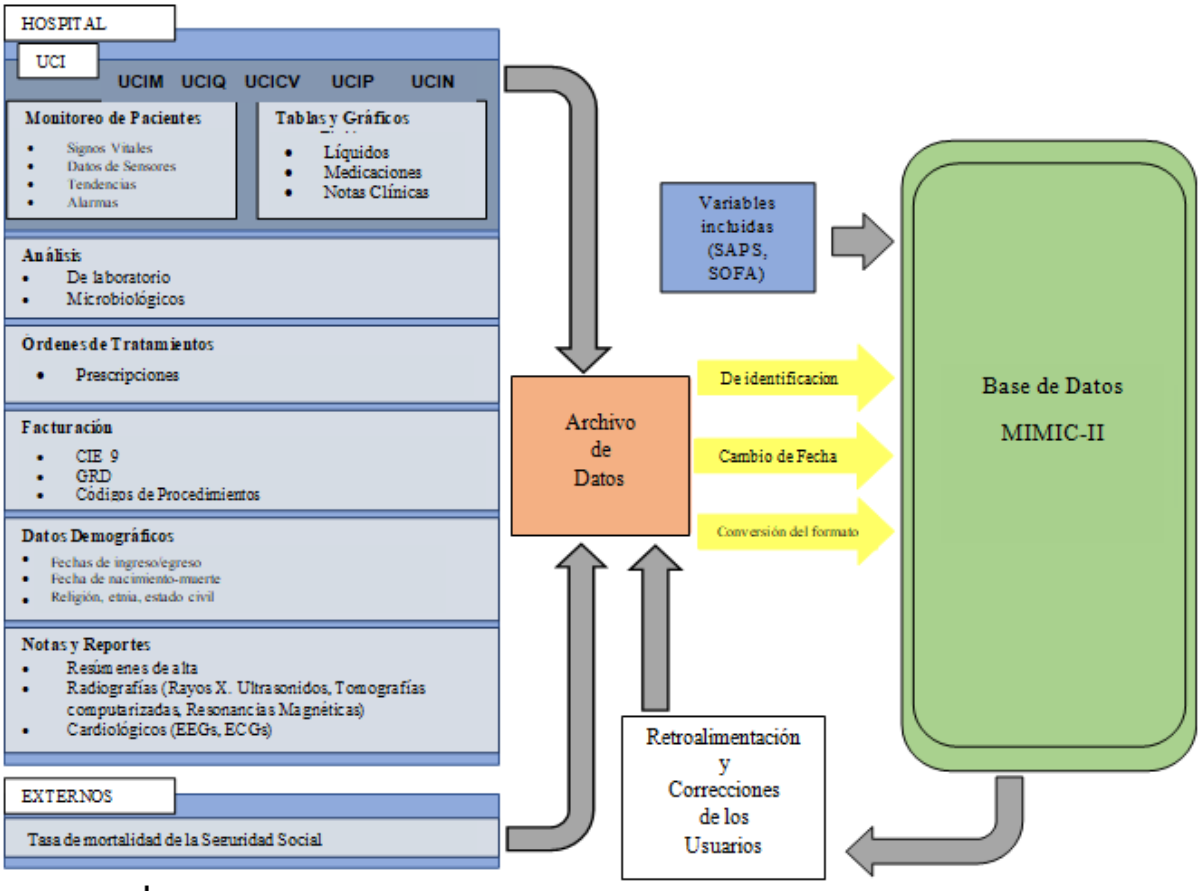

Fig. 2.1 Estructura general de la base de datos de MIMIC

UCIM: Unidad de cuidados Intensivos Medica; UCIQ: Unidad de Cuidados Intensivos Quirúrgica; UCICV: Unidad de Cuidados Intensivos Cardiovascular; UCIP: Unidad de Cuidados Intensivos Pediátrica; UCIN: Unidad de Cuidados Intensivos Neonatal

2.3.2 Acceso e interfaz

MIMIC-III es una base de datos de acceso libre para cualquier investigador del mundo, entrenado en el manejo de información sensible proveniente de pacientes. La base de datos es administrada por PhysioNet (http://physionet.org), un grupo heterogéneo de informáticos, físicos, matemáticos, investigadores biomédicos, médicos y educadores de todo el mundo. La tercera versión fue publicada en el año 2015 y se espera que sea actualizada continuamente con nuevos pacientes a lo largo del tiempo.

2.4 PCORnet

La Red Nacional de Investigación Clínica Centrada en el Paciente, PCORnet, es una iniciativa del Instituto de Investigación de Resultados Centrados en el Paciente (PECORI, del inglés Patient Centered Outcomes Research Institute). El PECORI involucra en la gobernanza de la red y en la selección de las preguntas de investigación tanto a los pacientes como a aquellos que cuidan de ellos.

La iniciativa de la PCORnet comenzó en el año 2013 buscando integrar información clínica proveniente de múltiples Redes de Investigación de Datos Clínica (CDRNs, del inglés Clinical Data Research Networks) y Redes de Investigación Impulsadas por Pacientes (PPRNs, del inglés Patient Powered Research Networks) [3]. Su centro coordinador está conformado por 9 socios: Harvard Pilgrim Health Care Institute, Duke Clinical Research Institute, Academy Health, Brookings Institution, Center for Medical Technology Policy, Center for Democracy&Technology, Group Health Research Institute, Johns Hopkins Berman Institute of Bioethics y America's Health Insurance Plans. PCORnet incluye a su vez 29 redes individuales que en conjunto permitirán el acceso a grandes cantidades de datos clínicos y sanitarios. El objetivo de la PCORnet es mejorar la eficiencia y capacidad de conducir investigaciones de efectividad comparativa.

2.4.1 Variables incluidas

Las variables de la base de datos PCORnet se extraen de varias HCE utilizadas en los nueve centros que conforman la red. Contiene datos clínicos e información sanitaria obtenida diariamente durante las consultas rutinarias de los pacientes. Adicionalmente, PICORnet incorpora datos compartidos por los individuos a través de sus registros personales de salud o compartido en redes sociales con otros pacientes a medida que manejan sus patologías durante su vida cotidiana. Esta iniciativa facilitará la investigación sobre varias condiciones médicas, involucrará una gran cantidad de pacientes de distintos sistemas sanitarios y representará una excelente oportunidad para la conducir estudios multicéntricos.

2.4.2 Acceso e interfaz

La PCORnet está concebida como un recurso nacional para investigación que posibilitará que los equipos de investigadores sanitarios y los pacientes trabajen en conjunto sobre preguntas de interés común. Estos equipos

podrán presentar preguntas de investigación y recibir los datos para conducir sus estudios. Los participantes actuales de la red (CDRNs, PPRNs y PCORI) están desarrollando las estructuras de gobernanza durante la fase de construcción y expansión que dura 18 meses [4].

2.5 Open NHS

El Servicio Nacional de Salud de Inglaterra (NHS, del inglés National Health System) es un organismo ejecutivo público no departamental dependiente del Departamento de Salud, una entidad gubernamental. El NHS contiene uno de los repositorios de datos sobre la salud de las personas más grande en el mundo. A su vez, es uno de los pocos sistemas de salud capaz de ofrecer información sobre salud de todos los sectores y a lo largo de la vida de una población entera.

Open NHS es un programa iniciado en octubre del año 2011. El NHS en Inglaterra ha realizado un esfuerzo activo por abrir los repositorios de datos utilizados en sus agencias y departamentos. El objetivo principal del giro hacia una base de datos de acceso abierto fue aumentar la transparencia y rastrear los resultados y la eficiencia del sector de salud británico [5]. Se espera que la información de alta calidad empodere al sector de la salud y de los servicios sociales para identificar prioridades y cumplir con las necesidades de las poblaciones locales. El NHS espera que al permitir a pacientes, médicos y comisionados comparar los servicios prestados y su calidad en distintas regiones del país se pueda identificar de forma más efectiva y rápida dónde el cuidado de la salud es inferior al ideal.

2.5.1 Variables incluidas

Open NHS es una base de datos abierta que contiene información pública del gobierno y otros organismos públicos.

2.5.2 Acceso e interfaz

Antes de la creación de la plataforma Open NHS, el SUS (del inglés, Secondary Uses Service) fue incorporado como parte del Programa Nacional para TICs en el NHS para proveer información sobre la planificación, puesta en marcha, administración, investigación y auditoría. Open NHS pasó a reemplazar a SUS como plataforma de acceso a la base de datos nacional del Reino Unido.

La Red de Investigación Clínica (CRN, del inglés Clinical Research Network) del Instituto Nacional de Investigaciones en Salud (NIHR, del inglés National Institute of Health Research), ha desarrollado e implementado una herramienta online conocida como Plataforma Open Data.

Además de los estudios retrospectivos que utilizan dichas bases de datos, se encuentra en marcha otra forma de investigación para comparar la calidad de los datos recolectados mediante HCE en relación con aquellos recolectados en forma manual por enfermeros investigadores. El personal de la CRN puede acceder a la Plataforma Open Data y conocer tanto la cantidad de pacientes reclutados en estudios de investigación en un determinado hospital como las investigaciones que se están haciendo en dicho hospital. Así pueden establecer qué hospitales son más exitosos reclutando pacientes, a qué velocidad lo hacen y en qué especialidades.

2.6 Otras investigaciones en curso

Las siguientes son otras bases de datos que se encuentran en etapa de desarrollo o tienen limitaciones de acceso más restrictivas:

2.6.1 eICU-Philips

Como parte de su colaboración con MIT, Philips brindará acceso a datos de cientos de miles de pacientes que fueron recolectados y anonimizados a través del programa de telemedicina eICU Philips del Hospital al Hogar. La información estará disponible para investigadores a través de PhysioNet, de manera similar que con la base de datos MIMIC.

2.6.2 VistA

El Sistema de Información en Salud VistA, (del inglés, Veterans Health Information System and Technology Architecture) es un sistema de información empresarial construido alrededor de las HCEs utilizadas en el sistema médico del Departamento de Asuntos de Veteranos de Estados Unidos (VA, del inglés Veterans Affairs). El sistema médico de VA opera en más de 125 hospitales, 800 clínicas ambulatorias y 135 centros de atención de pacientes crónicos. Todos estos establecimientos utilizan la interfaz VistA desde el año 1997. El sistema VistA aglutina hospitales, clínicas, farmacias y servicios complementarios para más de 8 millones de veteranos estadounidenses. Mientras que la red de salud tiene limitaciones y sesgos

para la investigación inherentes al gran porcentaje de pacientes hombres, el gran volumen de registros fehacientes disponibles sobrepasa dicha limitación. La base de datos de VA ha sido utilizada por numerosos investigadores en los últimos 25 años para conducir estudios de referencia en múltiples áreas [6, 7].

Esta base de datos tiene una larga historia de compromiso con la investigación médica y de colaboración con investigadores que son parte del sistema de VA. Históricamente, el acceso a la información se ha limitado a aquellos que tienen vinculación con VA. Sin embargo, con la tendencia reciente hacia el libre acceso a las grandes bases de datos, es cada vez mayor la discusión para habilitar su acceso a más investigadores. El amplio repositorio de información que contiene la base de datos permitiría a un gran grupo de investigadores mejorar la práctica clínica en múltiples campos. La fortaleza de los datos radica en la capacidad de rastrear pacientes a lo largo de Estados Unidos, tanto en el ámbito de la internación como en el ámbito ambulatorio. Como todas las prescripciones de drogas se encuentran cubiertas por el sistema de VA, la conexión con esta base de datos permite conducir grandes estudios fármaco-epidemiológicos con relativa facilidad.

2.6.3 NSQUIP

El Proyecto Nacional de Mejora de Calidad Quirúrgica (NSQUIP, del inglés National Surgical Quality Improvement Project) constituye un esfuerzo internacional realizado por el Colegio Americano de Cirujanos (ACS, del inglés American College of Surgeons) con el objetivo de mejorar los servicios quirúrgicos en todo el mundo [8]. El ACS trabaja con instituciones para implementar variadas intervenciones que mejoren la calidad de las cirugías en los hospitales. Una consecuencia del sistema es la recopilación de grandes cantidades de datos relativos a procedimientos quirúrgicos, resultados y eventos adversos. Toda la información proviene de las HCEs de las instituciones participantes.

La base de datos del NSQUIP está disponible gratuitamente para los miembros de las instituciones afiliadas, que son más de 653 centros en todo el mundo. Esta base de datos contiene grandes cantidades de información relativa a los procedimientos quirúrgicos, complicaciones, demografía e información del hospital. Aunque no cuenta con el grado de detalle de la MIMIC, tiene datos de numerosos hospitales del mundo y es, por lo tanto,

más generalizable a la práctica quirúrgica real. Es una base de datos particularmente importante en lo que concierne a cuidados quirúrgicos y calidad del cuidado, en especial en relación a los detalles de las complicaciones y eventos adversos de los procedimientos.

Nota: Las imágenes de este capítulo se encuentran disponibles a color en el ebook; disponible en www.hardineros.com

Referencias

1. Lee J, Scott DJ, Villarroel M, Clifford GD, Saeed M, Mark RG (2011) Open-access MIMIC-II database for intensive care research. In: Annual international conference of the IEEE engineering in medicine and biology society, pp 8315-8318.
2. Scott DJ, Lee J, Silva I et al (2013) Accessing the public MIMIC-II intensive care relational database for clinical research. BMC Med Inform DecisMak 13:9.
3. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS (2014) Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc JAMIA 21 (4): 578-582.
4. Califf RM (2014) The patient-centered outcomes research network: a national infrastructure for comparative effectiveness research. N C Med J 75 (3): 204-210.
5. Open data at the NHS [Internet]. Disponible en: https://www.nhsbsa. nhs. uk/open-data-portal-odp.
6. Maynard C, Chapko MK (2004) Data resources in the department of veterans affairs. Diab Care 27 (Suppl 2): B22-B26.
7. Smith BM, Evans CT, Ullrich P et al (2010) Using VA data for research in persons with spinal cord injuries and disorders: lessons from SCI QUERI. J Rehabil Res Dev 47 (8): 679-688.
8. NSQUIP at the American College of Surgeons [Internet]. Disponible en: https://www.facs.org/quality-programs/acs-nsqip.

Capítulo 3

DESAFÍOS Y OPORTUNIDADES EN EL ANÁLISIS SECUNDARIO DE DATOS DE LA HISTORIA CLÍNICA ELECTRÓNICA

Sunil Nair, Douglas Hsu y Leo Anthony Celi

Puntos clave

- Las historias clínicas electrónicas (HCE) son cada vez más útiles para realizar estudios observacionales, con un poder que compite con el de los ensayos controlados aleatorizados.
- El análisis secundario de datos de la HCE puede contribuir con la toma de decisiones a gran escala del sistema de salud (ej: farmacovigilancia) o para decisiones clínicas en la práctica clínica diaria (ej: elección de medicación).
- Los médicos, investigadores y científicos de datos necesitarán sortear numerosos desafíos frente al análisis de "big data" –incluyendo la interoperabilidad de los sistemas, el intercambio de información y la seguridad de los datos– para poder utilizar todo el potencial de las HCEs y los estudios basados en "big data".

3.1 Introducción

La creciente incorporación de la HCE ha creado nuevas oportunidades para que los investigadores, incluyendo a médicos y científicos de datos, tengan acceso a grandes bases de datos. Con estos datos, los investigadores están en posición de abordar la investigación con un poder estadístico nunca antes visto. En este capítulo presentamos y discutimos los desafíos en el uso secundario de los datos de la HCE, así como exploramos las oportunidades únicas que ofrecen estos datos.

3.2 Desafíos en el Análisis Secundario de datos de la HCE

Si bien se han dado grandes pasos al poner a disposición de los científicos de datos y los médicos clínicos, los registros de salud agrupados para actividades de investigación en salud, aún queda mucho por hacer para aprovechar toda la capacidad del "big data" en la atención médica. En todos los campos relacionados con la salud, los propietarios de los datos –es decir, las empresas farmacéuticas, las empresas de dispositivos médicos, los

sistemas de salud, y los ahora crecientes proveedores de HCE– están enfrentando simultáneamente presiones para proteger su capital intelectual y sus plataformas, garantizar la seguridad de los datos, y cumplir con las normativas de privacidad, sin obstaculizar la investigación que depende del acceso a las mismas bases de datos. Las historias de éxito del "big data" son cada vez más frecuentes, como resaltaremos a continuación, pero los desafíos no son menos abrumadores de lo que fueron en el pasado, y quizás se han vuelto incluso más exigentes a medida que despega el campo del análisis de datos en la asistencia sanitaria.

Los científicos de datos y sus socios clínicos tienen que lidiar con una cultura de la investigación que es altamente competitiva –tanto dentro de los círculos académicos, como entre los socios clínicos y de la industria. Si bien se ha escrito poco sobre la naturaleza del secreto de los datos en los círculos académicos, es una realidad que los presupuestos cada vez más ajustados y la mayor preocupación por la seguridad de los datos han empujado a los investigadores a utilizar los datos que tienen a mano, en lugar de buscar la integración de bases de datos separadas. Compartir datos de una manera segura y escalable es extremadamente difícil y costoso, o imposible incluso en la misma institución. Con el acceso restringido o bloqueado a los datos más relevantes, el poder estadístico y la capacidad de análisis longitudinales se reducen o pierden. No significa que los investigadores tengan malas intenciones– de hecho, muchos apreciarían la oportunidad de mayor colaboración en sus proyectos. Sin embargo, el tiempo, la financiación y la infraestructura para estos esfuerzos son sencillamente deficientes. Los datos también suelen estar segmentados en varios lugares y no se encuentran almacenados sistemáticamente en formatos similares en las distintas bases de datos clínicas o de investigación. Por ejemplo, la mayor parte de los datos clínicos se guarda en una variedad de formatos no estructurados, haciendo difícil su consulta directamente a través de algoritmos digitales [1]. En muchos hospitales, los datos clínicos de los departamentos de emergencias o de los pacientes ambulatorios pueden existir por separado de las HCE de internación y de la unidad de cuidados intensivos (UCI), por lo cual el acceso a uno de ellos no garantiza el acceso al otro. Las imágenes de Radiología y Patología suelen almacenarse por separado en sistemas diferentes y por ende, no son datos fáciles de relacionar con los resultados. La base de datos Medical Information Mart of

Intensive Care (MIMIC), descripta más adelante en este capítulo, que contiene los datos de las HCE de la UCI del Beth Israel Deaconess Medical Center (BIDMC), aborda y resuelve estas divisiones artificiales, pero requiere una gran cantidad de personal de ingeniería y apoyo que no está disponible en todas las instituciones.

Luego de años de preocupación por la confidencialidad de los datos, la industria farmacéutica recientemente ha dado un giro, poniendo a disposición de los investigadores datos detallados de sus ensayos clínicos ajenos a sus organizaciones. GlaxoSmithKline fue uno de los primeros en el año 2012 [2], seguido por una mayor iniciativa –the Clinical Trial Data Request– a la cual se han adherido otras grandes empresas farmacéuticas [3]. Los investigadores pueden solicitar el acceso a información a gran escala, e integrar bases de datos para meta análisis y otras revisiones sistemáticas. La próxima frontera será la liberación de los registros médicos que se conservan a nivel del sistema de salud. La ley de Tecnologías de la Información Sanitaria para la Salud Clínica y Económica (HITECH) del año 2009 fue un boom para el sector de las tecnologias de información (TIC) [4], pero los estándares de interoperabilidad entre los sistemas de registros siguen rezagados [5]. La brecha ha comenzado a resolverse mediante intercambios de información sanitaria patrocinados por los gobiernos, así como mediante la creación de nuevas redes de investigación [6,7], pero la mayoría de los expertos, los científicos de datos y los médicos siguen lidiando con datos incompletos.

Muchos de los obstáculos comerciales y técnicos mencionados anteriormente tienen sus raíces en las preocupaciones por la privacidad sostenida por los vendedores, los proveedores y sus pacientes. Dichas preocupaciones no son sin fundamento –las violaciones de los datos de grandes sistemas de salud se están volviendo alarmantemente frecuentes [8]. Recientemente, empleados de Partners Healthcare en Boston fueron blanco de un plan de suplantación de identidad o "*phishing*", proporcionando de forma involuntaria información personal que permitió a los hackers el acceso no autorizado a información de los pacientes [9]; los pacientes del Seton Healthcare en Texas sufrieron una violación de datos similar solo unos meses antes [10]. Las violaciones de datos no se limitan a los proveedores de atención médica– 80 millones de clientes de Anthem pudieron haber sufrido pérdida de su información personal a causa de un

ciberataque, el mayor de este tipo hasta la fecha [11]. No sorprende que en el contexto de estas violaciones, las compañías proveedoras de salud tengan uno de los puntajes más bajos de todos los sectores en seguridad de correo electrónico y prácticas de privacidad [12]. Dichos reportes ponen de manifiesto la necesidad de prudencia en medio de la exuberancia cuando se utilizan las HCEs agrupadas para el análisis de "big data"; dicho uso implica una responsabilidad ética de proteger los datos de la población y a nivel personal de actividades delictivas y otros fines nefastos. Para este propósito, las agencias federales han convocado a grupos de trabajo y audiencias públicas para resolver los vacíos en la seguridad de la información sanitaria, como la "*desidentificación"* de los datos fuera de las entidades no protegidas por la HIPAA y guías de consenso sobre lo que constituye "daño" de una violación de datos [13].

Incluso cuando los problemas de acceso de datos, integridad, interoperabilidad, seguridad y privacidad hayan sido resueltos, permanecerán los importantes costos en infraestructura y capital humano. Aunque el costo marginal de cada consulta adicional de "big data" es pequeño, el costo inicial para instalar un centro de datos y emplear científicos de datos dedicados puede ser significativo. No existen cifras para la creación de un centro de "big data" sanitario, y de todas formas, esas cifras podrían ser variables según la dimensión y el tipo de datos. Sin embargo, no debería sorprender que los ejemplos comúnmente citados de HCEs agrupadas con capacidades analíticas – MIMIC (BIDMC), STRIDE (Standford), el "data mart" Memorial Care (Memorial Health System, California, renta anual de US$2,2 billones), y el High Value Healthcare Collaborative (gestionado por Dartmouth, con otros 16 miembros y financiación del Center for Medicare and Medicaid Services) [14] – provienen de grandes sistemas de salud, con altos ingresos y experiencia regional en "big data".

Además de los aspectos anteriores, la confiabilidad de los estudios publicados usando métodos de "big data" es una preocupación importante para los expertos y los médicos. La cuestión específica es si estos estudios son simplemente amplificaciones de señales de bajo nivel que carecen de importancia clínica, o si son generalizables más allá de la base de datos de la cual derivan. Estas son preocupaciones genuinas en un ambiente médico y académico ya saturado de innumerables estudios de calidad variable. Los

escépticos están preocupados de que los análisis de big data sólo "aumentarán el ruido", desviando la atención y los recursos de otros escenarios de investigación científica, como los tradicionales ensayos clínicos controlados aleatorizados (ECA). Mientras que las limitaciones de los ECAs y la comparación favorable de los resultados de los grandes estudios observacionales con los hallazgos de los ECAs se discuten a continuación; de todas formas estos sentimientos tienen fundamento y deben ser considerados seriamente a medida que el análisis secundario de los datos de las HCE continua creciendo. Los líderes de opinión han sugerido que se expongan los principios del "big-data" descritos anteriormente para crear espacios de aprendizaje abiertos y colaborativos, a través de los cuales los datos *desidentificados* puedan ser compartidos entre investigadores. De esta manera, las distintas bases de datos pueden ser agrupadas para obtener un mayor poder o se pueden realizar investigaciones similares con diferentes conjuntos de datos, para ver si llegan a conclusiones similares. Los costos para dicha transparencia podrían ser asumidos por una única institución– por ejemplo, gran parte del costo de la creación de MIMIC ya ha sido invertido, por ende el costo incremental de hacer que los datos estén abiertos a otros investigadores es mínimo –o almacenados en una colaboración comprometida– como el High Value Healthcare Collaborative, financiada por sus miembros [16] o PCORnet, financiada por el gobierno federal [7]. Estas iniciativas de colaboración tendrían estructuras de gobierno y estándares transparentes de acceso a los datos, permitiendo la validación de los estudios y la revisión continua de los trabajos publicados y no publicados [15], y mitigaría los efectos de los sesgos de selección y confundidores en cualquier estudio individual [17].

A medida que las historias clínicas electrónicas agrupadas alcancen una escala cada vez mayor, los científicos de datos, investigadores y otras partes interesadas esperan que el costo del *hosting*, clasificación, formateo y análisis de estos registros se distribuyan entre un mayor número de partes interesadas, reduciendo el costo del análisis de HCEs agrupadas para todos los involucrados. Es posible que tengan que entrar en vigencia nuevas normas para el intercambio de datos de forma que las instituciones se sientan verdaderamente cómodas con esto. También dentro de las instituciones y las investigaciones colaborativas existentes, pueden implementarse prácticas para la seguridad de los datos y fomentarse una

mayor colaboración a través de la estandarización del registro y el almacenamiento de los datos. Deben trazarse líneas claras de responsabilidad para el acceso a los datos, y los repositorios de datos deben hacerse accesibles para aclarar el alcance de la información disponible para cualquier institución o grupo de investigación. La era del "big data" ha llegado al ámbito sanitario, y sólo a través de una continua adaptación y perfeccionamiento puede alcanzarse todo su potencial.

3.3 Oportunidades en el Análisis Secundario de los datos de la historia clínica electrónica

La creciente implementación de las historias clínicas electrónicas en el sistema de salud de los Estados Unidos ha creado vastas oportunidades para que los científicos clínicos, informáticos y otros investigadores de salud realicen consultas en grandes bases de datos de información clínica integrada para responder grandes y pequeños interrogantes. Con un tesoro de datos para explorar, médicos y científicos están en condiciones de evaluar preguntas de eficacia clínica y costo-efectividad –asuntos de máxima preocupación en la atención sanitaria estadounidense del siglo XXI– con una calidad y un poder estadístico rara vez alcanzado anteriormente en la investigación médica. La base de datos comercial de APACHE Outcomes, por ejemplo, contiene mediciones fisiológicas y de laboratorio de más de un millón de registros pacientes en 105 UCIs desde el año 2010 [18]. El Beth Israel Deaconess Medical Center –un hospital de atención terciaria con 649 camas, incluyendo 77 camas de cuidados críticos– proporciona una base de datos única del centro, de acceso abierto, que contiene datos de más de 60000 estadías en UCI [19].

Las bases de datos de unidades individuales y multicéntricas como las anteriores permiten realizar consultas a gran escala sin el gasto y dificultad, muchas veces insostenible, de un ensayo clínico aleatorizado (ECA), permitiendo así responder interrogantes previamente no verificables en los ECAs o estudios prospectivos de cohorte. Esto también puede hacerse con mayor precisión en la evaluación de diagnósticos o terapéuticas para sub-poblaciones seleccionadas, y para la detección de efectos adversos de medicamentos u otras intervenciones con mayor eficacia, entre otras ventajas [20]. En este capítulo, ofrecemos mayor información sobre la

utilidad del análisis secundario de los datos de la HCE para investigar interrogantes clínicos relevantes y proveer un apoyo útil para la toma de decisiones a los médicos, al equipo de salud y a los pacientes.

3.4 Análisis secundarios de HCE como alternativa a los Ensayos Clínicos Controlados Aleatorizados

Entre las limitaciones relativas de los ECA para fundamentar la toma de decisiones clínicas en el mundo real, se incluyen las siguientes: muchas de las comparaciones de tratamientos de interés para los médicos no han sido abordadas por los ECA; cuando se han realizado y evaluado los ECA, la mitad de las revisiones sistemáticas reportaron evidencia insuficiente para apoyar una intervención médica; y existen limitaciones reales de costos y proyectos que impiden a los ECA explorar escenarios clínicos específicos. Esto último incluye enfermedades raras, eventos clínicamente poco comunes o dispares, y una lista creciente de combinaciones reconocidas de subgrupos de pacientes, enfermedades concomitantes (genéticas, crónicas, agudas y adquiridas por la atención médica), y opciones de diagnóstico y tratamiento [20, 21].

Las consultas en las bases de datos de HCE que abordan interrogantes clínicos son esencialmente grandes estudios observacionales no aleatorizados. En comparación con los ECA, son relativamente más eficientes y menos caros de realizar [22], la mayor parte del costo ha sido absorbido por la instalación inicial del sistema y su mantenimiento, y el resto consiste principalmente en los salarios del personal de investigación, los costos del servidor o del espacio en la nube. Existe literatura que sugiere un alto grado de correlación entre los reportes de efectos del tratamiento en estudios no aleatorizados y ensayos clínicos aleatorizados. Ioannaidis y col. [23] hallaron una significativa correlación (Coeficiente de Spearman 0.75, $p<0.001$) entre los efectos de tratamientos reportados en ensayos aleatorizados versus estudios no aleatorizados en 45 temas diferentes de medicina interna general, que van desde la anticoagulación en el infarto de miocardio hasta la terapia con bajo nivel de láser en la osteoartritis. Es de particular interés que la variabilidad significativa en el resultado informado del tratamiento “se observó tan frecuentemente entre ensayos aleatorizados, como entre estudios aleatorizados y no aleatorizados”, y

observaron que la variabilidad era común tanto entre ensayos aleatorizados como entre los estudios no aleatorizados [23]. Vale la pena señalar que los mayores efectos del tratamiento fueron reportados más frecuentemente en estudios no aleatorizados que en ensayos aleatorizados (p=0,009) [23]. Sin embargo, esto no tiene por qué ser evidencia de sesgo de publicación, ya que el tamaño relativo del estudio y el protocolo conservador del ensayo también podrían causar este hallazgo. Los resultados de Ionnadis y col. [24], se repiten en un meta análisis de Cochrane más reciente, que no encontró diferencias significativas en las estimaciones de efectos entre los ECA y los estudios observacionales, independientemente de la heterogeneidad en el diseño del estudio observacional.

Para reducir aún más los confundidores en los estudios observacionales, los investigadores han empleado el puntaje de propensión [25], que permite equilibrar numerosas covariables entre los grupos de tratamiento, así como la estratificar las muestras por el puntaje de propensión para un análisis más detallado [26]. Kitsios y colegas compararon 18 estudios de puntaje de propensión en el ámbito de la UCI con al menos un ECA evaluando la misma pregunta clínica y encontraron un alto grado de acuerdo entre sus estimaciones de riesgo relativo y tamaño del efecto. Hubo una diferencia sustancial en la magnitud del tamaño de los efectos en un tercio de las comparaciones, alcanzando una significación estadística en un caso [27]. Aunque el ECA permanece en la cima de la medicina basada en la evidencia, es difícil de ignorar el poder de los grandes estudios observacionales que incluyan un ajuste adecuado de las covariables, así como de los estudios realizados cuidadosamente a partir de la revisión de HCEs. El alcance de los datos agrupados de la HCE –sean sesenta mil o un millón de registros– permite conocer los pequeños efectos del tratamiento que pueden ser sub-registrados o incluso pasar desapercibidos en ECAs de poco poder estadístico. Debido a que los costos son bajos en comparación con los ECA, también es posible investigar preguntas que carecen de patrocinadores. Finalmente, en el caso de los estudios observacionales realizados a partir de bases de datos, es más factible mejorar y repetir, o simplemente repetir, los estudios según sea necesario para investigar la precisión, heterogeneidad de los efectos y nuevos conocimientos clínicos.

3.5 Demostrando el Poder del Análisis Secundario de la HCE: Ejemplos en Farmacovigilancia y Atención Clínica

La seguridad de los fármacos es de gran preocupación tanto para los pacientes como para los médicos. Sin embargo, los métodos para garantizar la detección de efectos adversos luego de su comercialización son menos robustos de lo que sería deseable. Los fármacos a menudo son prescriptos a grandes y diversas poblaciones de pacientes que pueden no haber estado representadas adecuadamente en ensayos clínicos. De hecho, los ECA de cohortes pueden deliberadamente ser relativamente homogéneos con el fin de captar los efectos previstos de una medicación sin el "ruido" de las comorbilidades que podrían modificar los efectos del tratamiento. Humphreys y colegas (2013) informaron que en ensayos clínicos muy citados, el 40% de los pacientes identificados con la afección en cuestión no eran seleccionados, principalmente debido a criterios de selección restrictivos [29]. La variación en el diseño de los ensayos (comparadores, objetivos, duración del seguimiento) así como el tamaño de los mismos limita su capacidad de detectar efectos secundarios y efectos adversos de baja frecuencia o de largo plazo [28]. Los informes de vigilancia luego de la comercialización son recogidos deficientemente, no se integran regularmente, y pueden no ser de acceso público para respaldar la toma de decisiones clínicas por parte de los médicos o para informar la toma de decisiones por parte de los pacientes.

Las preguntas en las HCEs agrupadas –que esencialmente realizan estudios secundarios observacionales en grandes poblaciones– podrían compensar estas brechas en la farmacovigilancia. Los enfoques unicéntricos para esta y otras preguntas similares, relacionadas con la seguridad de los medicamentos en el ambiente clínico son prometedores. Por ejemplo, los hallazgos ampliamente difundidos del Kaiser Study en Vioxx® corroboraron las sospechas previas de una asociación entre el colecoxib y un aumento del riesgo de enfermedad coronaria grave [30]. Estos resultados se hicieron públicos en abril del 2004 luego de la presentación en una conferencia internacional; posteriormente Vioxx® fue retirado voluntariamente del mercado en septiembre del mismo año. Graham y colegas fueron capaces de aprovechar 2.302.029 años-persona de seguimiento de la base de datos permanente de Kaiser, para encontrar 8143 casos de enfermedad coronaria

en todos los AINEs que estaban en consideración, y posteriormente examinar los odds ratio apropiados [31].

Utilizando la base de datos MIMIC mencionada anteriormente, los investigadores del Beth Israel Deaconess Medical Center pudieron describir por primera vez un mayor riesgo de mortalidad en pacientes de UCI que habían tomado inhibidores selectivos de la recaptación de la serotonina (ISRS) antes de la admisión [32]. Un análisis más detallado demostró que la mortalidad variaba según los ISRS específicos, con una mayor mortalidad entre los pacientes que tomaban ISRS de mayor afinidad (por ejemplo, aquellos con mayor inhibición de serotonina); por otro lado, la mortalidad no podía ser explicada por los efectos adversos comunes de los ISRS, como el impacto sobre las variables hemodinámicas [32].

La utilidad del análisis secundario de datos de la HCE no se limita al descubrimiento de los efectos del tratamiento. A falta de estudios publicados para guiar la decisión potencial de anticoagular a un paciente con Lupus pediátrico con múltiples factores de riesgo de trombosis, médicos de Standford recurrieron a su propia plataforma de consulta de HCE (Standford Transnational Research Integrated Database Environment-STRIDE) para crear una cohorte electrónica de pacientes con lupus pediátrico con el fin de estudiar las complicaciones de esta enfermedad [33]. En cuatro horas, un sólo médico determinó que los pacientes con complicaciones del Lupus similares tenían un alto riesgo relativo de trombosis, y se tomó la decisión de administrar anticoagulación [33].

3.6 Un Nuevo Paradigma para Apoyar la Clínica Basada en la Evidencia y Consideraciones Éticas

Experiencias innovadoras como las previamente mencionadas, combinadas con la evidencia que apoya la eficacia de los ensayos observacionales para guiar adecuadamente la práctica clínica, validan el concepto de HCEs agrupadas como grandes poblaciones de estudio que poseen gran cantidad de información esperando ser aprovechada para el apoyo en la toma de decisiones clínicas y para la seguridad del paciente. Uno puede imaginar a un futuro médico solicitando una consulta grande o pequeña como las que se describieron anteriormente. Dichas preguntas podrían estar relacionadas con la eficacia de una intervención en una

subpoblación, o con un único paciente complicado cuyas condiciones no son plasmadas satisfactoriamente en ningún ensayo publicado. Quizás esto es suficiente para que un médico recomiende una nueva práctica clínica; o quizás diseñarán un estudio observacional pragmático para obtener más matices –evaluando la respuesta a la dosis o el perfil de efectos adversos en todas las subpoblaciones. A medida que se toman las decisiones clínicas y se determina el rumbo de la atención del paciente, esta intervención y los resultados se ingresan en la historia clínica electrónica, creando efectivamente un circuito de retroalimentación para consultas futuras [34].

Por supuesto, las ventajas del análisis secundario de datos de la historia clínica electrónica siempre deben equilibrarse con consideraciones éticas. Al contrario que en los ECA, no existe un proceso de consentimiento explícito para el uso de datos demográficos, clínicos y otros datos potencialmente confidenciales capturados en la HCE. Consultas suficientemente específicas podrían arrojar resultados muy estrictos, en teoría lo suficientemente específicos como para identificar a un paciente en particular. Por ejemplo, en una investigación sobre pacientes con una enfermedad rara, con ciertos grupos de edad, y admitidos en un plazo de tiempo limitado, podría incluir a alguien que pueda ser conocido por su comunidad. Un ejemplo tan extremo destaca la necesidad de cumplir con las leyes federales de privacidad, así como de garantizar altos estándares institucionales de seguridad de datos, tales como servidores seguros, acceso limitado, *firewalls* y otros métodos de seguridad de los datos.

Yendo más allá, los científicos de datos deberían considerar medidas adicionales diseñadas para proteger el anonimato de los pacientes, por ejemplo, el cambio de fecha, tal como se implementa en la base de datos MIMIC (ver Sección 5.1, Capítulo 5). En situaciones en las que las consultas pudieran potencialmente re-identificar pacientes, como en la investigación de enfermedades raras, o en el curso de un brote contagioso, los investigadores y las comités de ética en investigación deberían buscar acuerdos con este subconjunto relativamente pequeño de pacientes potencialmente afectados y sus grupos de apoyo, para asegurar su comodidad con los análisis secundarios. Podría ser necesaria la divulgación de los propósitos y métodos de investigación por parte de quienes buscan acceso a los datos, y debería ofrecerse a los pacientes la posibilidad de prohibir el uso de sus propios datos.

Es responsabilidad de los investigadores y científicos de datos explicar los beneficios de la participación en un análisis secundario a los pacientes y a los grupos de pacientes. Este intercambio permite al sistema médico crear una base de datos clínica de suficiente magnitud y calidad para beneficiar individuos y a grupos de pacientes, en tiempo real o en el futuro. Además, la recolección pasiva de datos clínicos permite que el paciente contribuya, con un riesgo relativo muy bajo y sin costo personal, a la atención clínica continua y futura de otros. Creemos que las personas son lo suficientemente altruistas para considerar las contribuciones de sus datos a la investigación, siempre y cuando los riesgos potenciales del uso de datos sean pequeños y se encuentren bien descritos.

En última instancia, el análisis secundario de la HCE solo tendrá éxito si los pacientes, los reguladores y otras partes interesadas tienen la seguridad y la tranquilidad de que sus datos de salud se mantendrán seguros, y si los procesos para su uso se hacen transparentes para asegurar el beneficio para todos.

Nota: Las imágenes de este capítulo se encuentran disponibles a color en el ebook; disponible en www.hardineros.com

Referencias

1. Riskin D (2012) Big data: opportunity and challenge. Healthcare IT News, 12 June 2012. Disponible en http://www.healthcareitnews.com/news/big-data-opportunity-and-challenge.
2. Harrison C (2012) GlaxoSmithKline opens the door on clinical data sharing. Nat Rev Drug Discov 11 (12): 891-892. Doi: 10.1038/nrd3907 [Medline: 23197021].

3. Clinical Trial Data Request. Disponible en https://clinicalstudydatarequest.com/. [Consultado 11 de Agosto 2015]. [Web Cite Cache ID 6TFyjeT7t].
4. Adler-Milstein J, Jha AK (2012) Sharing clinical data electronically: a critical challenge for fixing the health care system. JAMA 307 (16): 1695-1696.
5. Verdon DR (2014) ONC's plan to solve the EHR interoperability puzzle: an exclusive interview with National Coordinator for Health IT Karen B. DeSalvo. Med Econ. Disponible en https://www.medicaleconomics.com/view/oncs-plan-solve-ehr-interoperability-puzzle.
6. Green M (2015) 10 things to know about health information exchanges. Becker's Health IT CIO Rev. Disponible en: https://www.beckershospitalreview.com/healthcare-information-technology/10-things-to-know-about-health-information-exchanges.html.
7. PCORnet. Disponible en http://www.pcornet.org/. [Consultado 11 de Agosto 2015].
8. Dvorak K (2015) Big data's biggest healthcare challenge: making sense of it all. Fierce Health IT, [Consultado 4 de Mayo 2015] Disponible en: http://www.fiercehealthit.com/story/big-datas-biggest-healthcare-challenge-making-sense-it-all/2015-05-04.
9. Bartlett J (2015) Partners healthcare reports data breach. Boston Bus J. Disponible en http://www.bizjournals.com/boston/blog/healthcare/2015/04/partners-healthcare-reports-potential-data-breach.html.
10. Dvorak K (2015) Phishing attack compromises info of 39 K at Seton healthcare family. Fierce Health IT, [Consultado 28 de abril 2015]. Disponible en: http://www.fiercehealthit.com/story/phishing-attack-compromises-info-39k-seton-healthcare-family/2015-04-28.
11. Bowman D (2015) Anthem hack compromises info for 80 million customers. Fierce Health Payer, [Consultado 5 de febrero 2015]. Disponible en: http://www.fiercehealthpayer.com/story/anthem-hack-compromises-info-80-million-customers/2015-02-05.
12. Dvorak K (2015) Healthcare industry 'behind by a country mile' in email security. Fierce Health IT, [Consultado 20 de febrero 2015]. Disponible en http://www.fiercehealthit.com/story/healthcare-industry-behind-country-mile-email-security/2015-02-20.
13. White house seeks to leverage health big data, safeguard privacy. Health Data Manage. Disponible en: http://www.healthdatamanagement.com/news/White-House-Seeks-to-Leverage-Health-Big–Data-Safeguard-Privacy-50829-1.html.
14. How big data impacts healthcare. Harv Bus Rev. Disponible en https://hbr.org/resources/pdfs/comm/sap/18826-HBR-SAP-Healthcare-Aug-2014.pdf [Consultado 11 de Agosto 2015].
15. Moseley ET, Hsu DJ, Stone DJ, Celi LA (2014) Beyond open big data: addressing unreliable research. J Med Internet Res 16 (11): e259.

16. High value healthcare collaborative. Disponible en http://highvaluehealthcare.org/. [Consultado 14 de Agosto 2015].
17. Badawi O, Brennan T, Celi LA et al (2014) Making big data useful for health care: a summary of the inaugural MIT critical data conference. JMIR Med Inform 2 (2): e22.
18. APACHE Outcomes. Disponible en https://www.cerner.com/Solutions/Hospitals-and-Health-Systems/Critical-Care/APACHE-Outcomes/. [Consultado Nov 2014].
19. Saeed M, Villarroel M, Reisner AT et al (2011) Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database. Crit Care Med 39:952.
20. Ghassemi M, Celi LA, Stone DJ (2015) State of the art review: the data revolution in critical care. Crit Care 19:118.
21. Mills EJ, Thorlund K, Ioannidis J (2013) Demystifying trial networks and network meta-analysis. BMJ 346: f2914.
22. Angus DC (2007) Caring for the critically ill patient: challenges and opportunities. JAMA 298:456-458.
23. Ioannidis JPA, Haidich A-B, Pappa M et al (2001) Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA 286:7.
24. Anglemyer A, Horvath HT, Bero L (2014) Healthcare outcomes assess with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev 29:4.
25. Gayat E, Pirracchio R, Resche-Rigon M et al (2010) Propensity scores in intensive care and anesthesiology literature: a systematic review. Intensive Care Med 36:1993-2003.
26. Glynn RJ, Schneeweiss S, Stürmer T (2006) Indications for propensity scores and review of their use in pharmacoepidemiology. Basic Clin Pharmacol Toxicol 98:253-259.
27. Kitsios GD, Dahabreh IJ, Callahan S et al (2015) Can we trust observational studies using propensity scores in the critical care literature? A systematic comparison with randomized clinical trials. Crit Care Med 2015 Sep; 43 (9): 1870-9.28. Celi LA, Moseley E, Moses C et al (2014) from pharmacovigilance to clinical care optimization. Big Data 2 (3): 134-141.
29. Humphreys K, Maisel NC, Blodgett JC et al (2013) Extent and reporting of patient non enrollment in influential randomized clinical trials, 2001 to 2010. JAMA Intern Med 173:1029-1031.
30. Vioxx and Drug Safety. Statement of Sandra Kweder M.D. (Deputy Director, Office of New Drugs, US FDA) before the Senate Committee on Finance. Disponible en http://www.fda.gov/News Events/Testimony/ucm113235. htm. [Consultado Julio 2015].
31. Graham DJ, Campen D, Hui R et al (2005) Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. Lancet 365 (9458): 475-481.
32. Ghassemi M, Marshall J, Singh N et al (2014) Leveraging a critical care database: selective serotonin reuptake inhibition use prior to ICU admissión is associated with increased hospital

mortality. Chest 145 (4): 1-8.
33. Frankovich J, Longhurst CA, Sutherland SM (2011) Evidence-based medicine in the EMR era. New Engl J Med 365:19.
34. Celi LA, Zimolzak AJ, Stone DJ (2014) Dynamic clinical data mining: search engine-based decisión support. JMIR MedInform 2 (1): e13.

CAPÍTULO 4

CONECTÁNDOLO TODO: IMAGINANDO UN SISTEMA DE ATENCIÓN IDEAL BASADO EN DATOS

DAVID STONE, JUSTIN ROUSSEAU Y YUAN LAI

Puntos clave

- Un Sistema de Atención Ideal deberia incorporar elementos fundamentales de ingeniería de control como la detección, el cálculo, la acción y la retroalimentación, eficaces y basados en datos.
- Estos sistemas deben ser cuidadosa e intencionalmente diseñados con el fin de apoyar las decisiones clínicas, y no basados en las presiones del mercado y la conveniencia de los usuarios.

Este capítulo presenta ideas acerca de cómo los datos podrían ser empleados sistemáticamente de forma más efectiva en un sistema de salud diseñado con ese propósito. Previamente hemos escrito acerca de los componentes potenciales de tal sistema – por ejemplo, la minería de datos clínicos dinámica, cerrando el círculo de datos de la UCI, optimizando el sistema de datos en sí mismo, el *"crowdsourcing"* o colaboración abierta distribuida, etc., e intentaremos unirlo todo en este capítulo, que esperamos inspire y anime a otros a pensar y a movilizarse para crear un sistema de este tipo [1-10]. Dicho sistema, en teoría, apoyaría el flujo de trabajo clínico [1] aprovechando los datos para proporcionar una atención personalizada o de precisión a los individuos, al tiempo que se garantiza una atención óptima a nivel de la población; [2] proporcionando coordinación y comunicación entre los usuarios del sistema; y [3] definiendo, rastreando y mejorando la seguridad y la calidad. Si bien la atención médica es intrínsecamente heterogénea a nivel individual de los pacientes, consultas, especialidades y entornos clínicos, también proponemos algunas soluciones generales basadas en sistemas derivadas de casos de uso definidos por contextos. Este capítulo describe la infraestructura fundamental de un Sistema de Atención Ideal (SAI) logrado a través de la identificación, organización, captura, análisis, utilización e intercambio adecuado de los datos.

4.1 Ejemplos de Casos de Uso basados en la Inevitable Heterogeneidad Médica

La heterogeneidad intrínseca inherente a la atención de la salud a nivel individual de los pacientes, consultas, especialidades y entornos clínicos ha imposibilitado una solución única simple de sistemas. Anticipamos en un SAI los requisitos de identificar los elementos centrales comunes al cuidado médico de todos los pacientes (principios de seguridad, cuidados preventivos, cuidados efectivos de fin de la vida, listas de problemas actualizadas y precisas y manejo de listas de medicación) y posteriormente formular rutas basadas en contextos específicos. Debemos tener en cuenta que un paciente individual puede atravesar múltiples categorías. Cualquier paciente ambulatorio complejo también tendrá los requerimientos básicos de los objetivos de atención de un paciente ambulatorio en buen estado de salud y puede en algún momento tener un episodio de hospitalización. La Tabla 4.1 identifica una variedad de casos de uso práctico, incluyendo formas resumidas de los temas clínicos y de los datos pertinentes asociados con ellos.

Tabla 4.1 Casos de uso práctico con sus correspondientes objetivos clínicos y de datos

Casos de uso clínico	**Objetivos clínicos**	**Objetivos de datos**
Paciente ambulatorio en buen estado de salud	Proveer la atención preventiva necesaria; abordar las enfermedades agudas leves intermitentes	Documentación de mantenimiento preventivo de la salud: registros de vacunación, registros de exámenes de *screening* para cáncer, documentación de alergias, datos de tabaquismo y obesidad
Paciente ambulatorio con problemas crónicos complejos	Conectar y coordinar cuidados entre diversos sistemas y cuidadores	Garantizar información precisa y sincronizada en todos los ámbitos de la atención sin necesidad de supervisión por parte del paciente y/o familia; monitoreos específicos para prevenir la admisión y el reingreso
Paciente hospitalizado- cirugía electiva	Brindar un proceso quirúrgico y prequirúrgico seguro	Realizar un seguimiento de los procesos relevantes para la seguridad y la calidad; realizar un seguimiento de los resultados, complicaciones, incluidos los resultados relacionados con la seguridad
Paciente hospitalizado (departamento de	Identificar y predecir los pacientes del departamento de emergencias que requieren	Realizar un seguimiento de los resultados de los pacientes de urgencias, incluyendo los traslados a UCI y la mortalidad;

urgencias, sala de internación general, unidad de cuidados intensivos)	cuidados en UCI; seguridad y calidad de UCI; Identificar y predecir eventos adversos	seguimiento de los eventos adversos; monitorear mediciones habituales e innovadoras de la UCI
Paciente de centro de cuidados crónicos	Conectar y coordinar la atención entre diversos lugares y cuidadores de un paciente que tal vez no puede participar activamente del proceso	Asegurar información precisa y coordinada en todos los ámbitos de la atención sin necesidad de supervisión por parte del paciente y/o su familia
Egreso reciente del hospital	Prevenir el reingreso	Búsqueda de predictores asociados con los reingresos y las consecuentes intervenciones basadas en esas determinaciones; Realizar un seguimiento de los resultados clínicos y funcionales
Trabajo de Parto y alumbramiento	Decisión y momento para la cesárea; tasas más bajas de intervención y complicaciones	Obtención de datos para los predictores asociados con la cesárea u otras intervenciones; Seguimiento de las tasas de complicaciones y los resultados.
Cuidados paliativos y Final de la vida	Decisión y momento para iniciar cuidados paliativos; asegurar confort e integridad	Obtención de datos para determinar las características que indican la implementación de cuidados paliativos

4.2 Flujo de Trabajo Clínico, Documentación y Decisiones

La digitalización de la medicina ha avanzado con la amplia adopción de las historias clínicas electrónicas, en parte gracias a un uso significativo como parte de la ley de Tecnología de la Información en Salud Clínica y Económica (HITECH, por sus siglas en ingles) [11], pero ha generado variadas respuestas de parte de los médicos. Un amplio grado de digitalización es un elemento fundamental para crear un SAI. Definido al más alto nivel, un sistema es un conjunto de partes y funciones (también conocido como componentes y protocolos) que recibe entradas y produce salidas. En la atención médica, las entradas son los pacientes en distintos estados de salud y enfermedad, y las salidas son los resultados de estos pacientes. La Figura 4.1 muestra un circuito de control sencillo que describe la configuración de un sistema de salud basado en datos.

El ejercicio de la medicina tiene una larga historia de estar basada en datos, con la medicina diagnóstica que se remonta a tiempos antiguos. Los médicos recogen y organizan datos de las historias, examen físico y una gran variedad de exámenes para formular diagnósticos, pronósticos y los subsecuentes tratamientos. Sin embargo, este proceso no ha sido óptimo en

el sentido que estas decisiones, y las acciones posteriores basadas en esas decisiones, se han realizado en relativo aislamiento. Las decisiones dependen de la experiencia previa y el estado actual de conocimiento de los médicos involucrados, el cual puede o no estar apropiadamente basado en la evidencia. Además, estas decisiones, en su mayoría, no se han seguido ni medido para determinar su impacto en la seguridad y calidad. Por ende, hemos perdido mucho de lo bueno que se ha hecho y no hemos detectado mucho de lo que era malo.

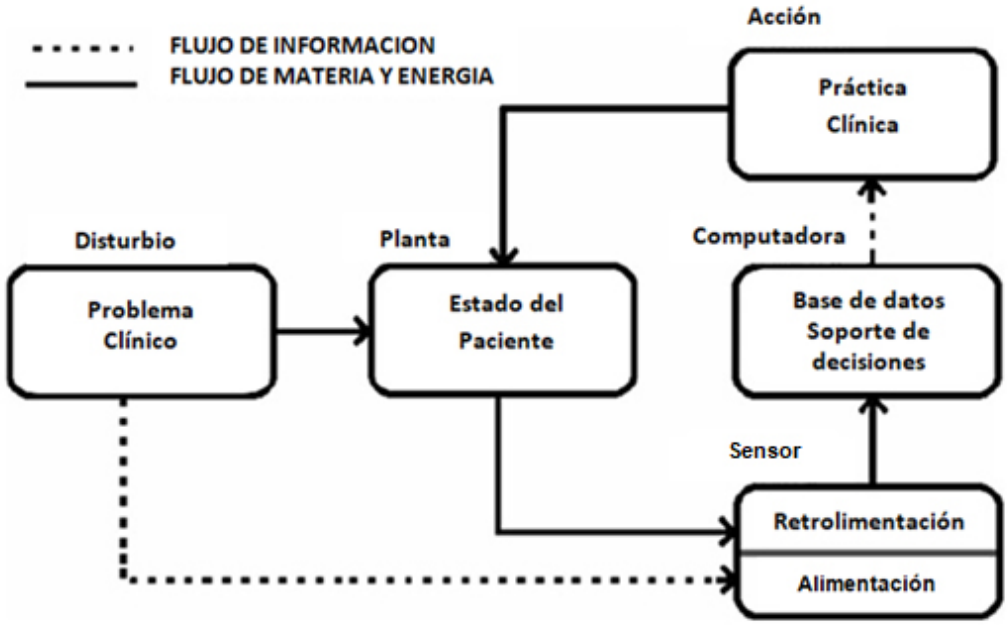

Fig. 4.1 Circuito de control que representa un sistema de atención basado en datos.
Un problema clínico como una infección u oclusión vascular afecta el estado del paciente. Posteriormente, el sensor del sistema detecta este cambio y envía los datos relevantes a la computadora para su almacenamiento y análisis. Esto puede dar lugar o no, a una intervención clínica que afecta aún más al estado del paciente, y que retroalimenta al sistema para su posterior análisis. El control de la retroalimentación implica la transmisión de alteraciones directamente al sensor sin que antes se vea afectado el estado del paciente. La detección de un factor de riesgo para trombosis venosa que desencadena la profilaxis basada en protocolos representa un ejemplo clínico del control de retroalimentación [3].

La digitalización de la medicina provee la oportunidad de remediar estas situaciones. A pesar de la deficiente usabilidad de la tradicional documentación en papel, las notas de los médicos en lenguaje natural

constituyen los datos fundamentales necesarios para alimentar un sistema de atención ideal. Si bien datos tales como los valores de laboratorio y los signos vitales fisiológicos pueden ser bastante confiables y cuantitativos, en general no reflejan la toma de decisiones y los diagnósticos que se establecen o consideran, los cuales derivan del análisis y síntesis de los datos disponibles (la evaluación con diagnósticos diferenciales), así como de los datos que se adquirirán en el plan diagnóstico.

La digitalización de la medicina se ha encontrado con dos cuestiones clave: [1] ¿Cómo puede desarrollarse un flujo de trabajo digital que soporte una documentación rápida y precisa para que el médico se sienta informado y no agobiado por el proceso? [2] ¿Cómo puede el proceso de entrada de datos apoyar y mejorar el proceso de toma de decisiones médicas? La primera iteración de las historias clínicas electrónicas (HCEs) ha intentado simplemente replicar la tradicional documentación en papel a un formato digital. Con el fin de abordar la primera cuestión, un soporte más eficaz del proceso de documentación requerirá rediseños innovadores para mejorar la HCE a medida que evoluciona. En lugar de requerir que el médico se siente frente a un teclado lejos de un paciente, el proceso necesita captar la información en tiempo real del encuentro con el paciente, en modalidades como el reconocimiento visual o de voz. Esto debe hacerse para que los detalles importantes sean capturados sin una excesiva interferencia con las interacciones personales o sin entradas erróneas producto del retraso en el registro. El sistema receptor debe tomar en cuenta la información previa del paciente en la interpretación de nuevas entradas para reconocer y asimilar con precisión la información esencial de la consulta en curso. Por otra parte, los datos que son recolectados no deben perderse a medida que el paciente avanza en el tiempo y se mueve entre distintas localizaciones geográficas. Un tema crucial es el que se ha perpetuado en la práctica actual de la medicina, de una consulta a otra– el médico y el paciente no deberían tener que "reinventar la rueda de la información" en cada consulta. Si bien cada médico debe proporcionar un enfoque nuevo a cada paciente, esto no debe requerir refrescar la historia médica completa del paciente en todas las consultas, desperdiciando tiempo y esfuerzo. Por otro lado, lo que se documenta debe ser transparente para el paciente, en contraste con el modelo de "beneficencia médica" que ha sido practicado en la mayor parte de la historia de la medicina donde era considerado beneficioso restringir el

acceso de los pacientes a sus propios registros. Se están dando pasos para alcanzar el objetivo de transparencia con el paciente con el movimiento de OpenNotes comenzado en el año 2010. Los efectos de este movimiento están siendo reconocidos a nivel nacional, con beneficios potenciales significativos en muchas áreas relacionadas con la seguridad del paciente y la calidad de la atención médica [13].

En cuanto a la segunda cuestión, hemos escrito acerca de cómo la entrada de datos de calidad puede apoyar la toma de decisiones médicas [14]. Las futuras iteraciones de una HCE rediseñada de forma innovadora en un SAI deberían ayudar en el ensamblaje y presentación inteligente de los datos, así como en la presentación del apoyo en la toma de decisiones en forma de evidencia y educación. Entonces, quien toma las decisiones es capaz de abordar cada consulta con la ventaja de un conocimiento previo y evidencia de respaldo longitudinal para el paciente individual, así como con comparaciones de su estado de salud con el de aquellos de pacientes con datos y diagnósticos similares (Fig. 4.2).

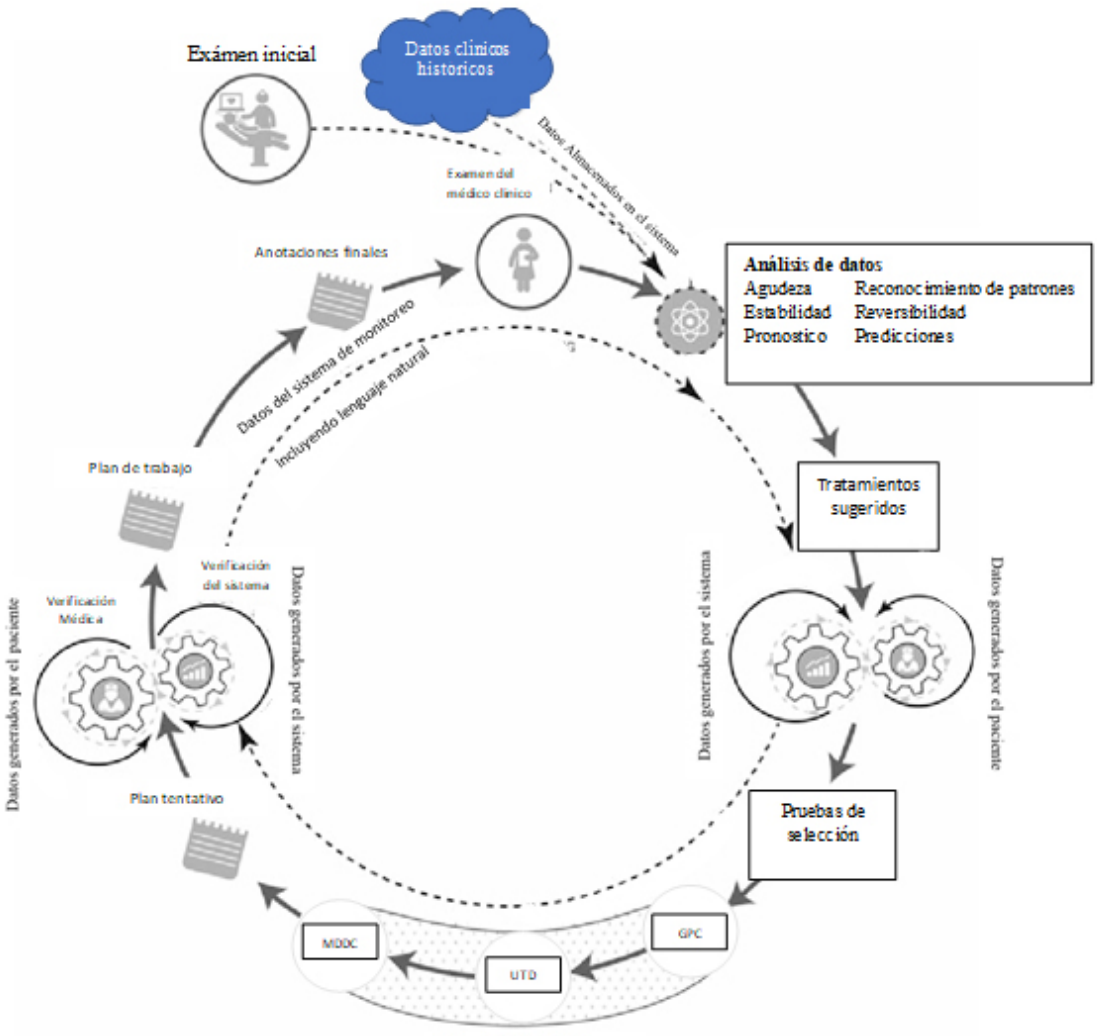

Fig. 4.2 Documentación clínica con un soporte de sistemas de datos totalmente integrado.
Las notas y datos previos son entradas para futuras notas y decisiones. El sistema analiza las entradas y muestra los diagnósticos sugeridos y la lista de problemas, y luego recomendaciones de estudios complementarios y tratamientos ordenados en forma jerárquica según varios niveles de evidencia: GPC– guías de práctica clínica, UTD-Up to Date, MDDC-minería dinámica de datos clínicos [14].

Pueden reconocerse patrones y tendencias en los datos, particularmente en el contexto de la historia médica previa de ese paciente y en su estado actual (Fig. 4.3).

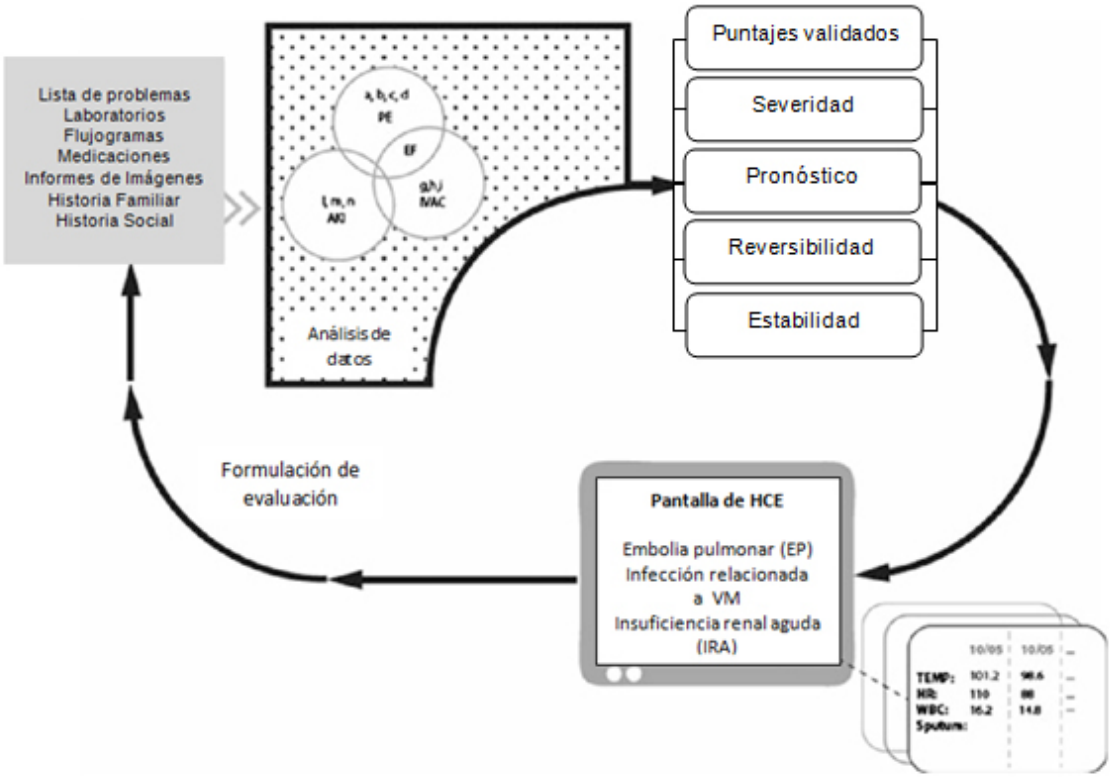

Fig. 4.3 Modelo de pantalla de Evaluación con ejemplos de datos de análisis.
Con base en los análisis constantemente realizados por el sistema y actualizados a medida que el usuario comienza a introducir notas, una serie de problemas son identificados y sugeridos al usuario por la pantalla de la HCE. Después de considerar estas sugerencias además de su propio análisis, el usuario puede seleccionar o editar los problemas que se sugieren o introducir problemas totalmente nuevos. La selección final de los problemas es considerada con el análisis en curso para futuras evaluaciones [14]

Los datos poblacionales deberían aprovecharse para optimizar las decisiones para los individuos, capturando, almacenando y utilizando la información de las consultas individuales para apoyar el cuidado de los otros, lo que hemos descrito como “minería dinámica de datos clínicos” [2]. Esto también es similar a lo que ha sido descrito como un “sistema de aprendizaje en salud” o un “código verde” para consultar los datos poblacionales para apoyar la toma de decisiones [15, 16].

En resumen, un SAI debe tener herramientas (por ejemplo, versiones mejoradas de las HCEs actuales) para capturar y utilizar los datos en formas

que permitan una documentación y toma de decisiones efectiva y eficiente en vez de solitaria y agobiante. Si bien notamos que los médicos individuales funcionan de manera brillante a pesar de los obstáculos e ineficiencias técnicas y de los sistemas a los que se enfrentan, hemos llegado a un punto de necesidad, reconocido por el Instituto de Medicina, que amenaza la calidad y seguridad de la atención de salud, y que requiere el desarrollo de herramientas digitales que faciliten la entrada de datos y la toma de decisiones, así como herramientas que puedan interactuar e incorporar otras características de un SAI integrado y basado en la tecnología digital. Esto requerirá interacciones y colaboraciones cercanas entre trabajadores de la salud, ingenieros incluyendo expertos en software y hardware, así como entre los pacientes, reguladores, responsables políticos, proveedores y administradores técnicos y de negocios del hospital [5].

4.3 Niveles de precisión y personalización

Muchas de las herramientas disponibles para los médicos se han vuelto increíblemente sofisticadas, incluidos los dispositivos tecnológicos y los conocimientos sobre biología molecular y bioquímica.

Sin embargo, otros elementos, incluidos aquellos que se utilizan cotidianamente, son más primitivos y serían familiares para los médicos de la antigüedad. Estos elementos incluyen por ejemplo datos clínicos como la frecuencia cardíaca y la presión arterial registrados en las planillas de enfermería. La monitorización de los pacientes no se aplica generalmente basándose en los datos, en particular las decisiones relativas a quiénes son monitoreados y con qué señales, la duración de la monitorización, y si los datos son almacenados, analizados, y utilizados más allá del momento actual. Además, cabe preguntarse si los umbrales numéricos comúnmente preestablecidos como valores anormalmente altos o bajos, extraen la máxima información clínica de esas señales. El reconocimiento de valores anormales se ha convertido en un problema importante, origen de un exceso de falsas alarmas que genera una consecuente fatiga de alarmas [18]. El análisis de los datos debería proporcionar a los médicos características personalizadas y contextualizadas de los signos vitales individuales (por ejemplo, patrones de variabilidad de frecuencia cardíaca y respiratoria, cambios sutiles de las formas de onda de ECG, etc.), de modo que los cambios realmente importantes puedan ser reconocidos de manera rápida y

efectiva sin abrumar la carga cognitiva del médico. Esto constituiría "monitorización personalizada basada en datos", en la que se analizan en tiempo real los datos de la pantalla del monitor para proporcionar información sobre el estado del paciente. Esta situación será más importante a medida que la monitorización sea más frecuente tanto en el hospital como en el ámbito ambulatorio, lo que no está lejos de ser una realidad con el desarrollo exponencial de monitores y aplicaciones de salud móviles.

Un enfoque potencial para esta cuestión consistiría en tratar a los monitores como un componente especializado de la HCE en lugar de dispositivos independientes que muestran la frecuencia cardíaca y emiten alarmas con frecuencia, a veces incluso injustificadas. De hecho, esto ha ocurrido en algunas situaciones en la medida en que los monitores se han conectado en red y en muchos casos pueden importar datos a la HCE. El circuito se cerrará cuando la información fluya bidireccionalmente, de modo que la HCE (y otros elementos como las bombas de infusión) puedan ayudar a proporcionar contextos clínicos e información personalizada para mejorar el rendimiento potencial de los monitores [14]. Mientras que en la actualidad, la interfaz de usuario del monitor permite únicamente ajustar los canales monitoreados y las alarmas, en un futuro la interfaz de usuario también será cada vez más rica, de modo que el usuario podrá, por ejemplo, con las adecuadas credenciales, acceder, editar y registrar en la HCE desde un monitor central o de cabecera, o añadir información directamente al monitor para ajustar el proceso de monitorización.

Los datos de los monitores están empezando a ser utilizados con fines analíticos prospectivos en términos de predicción de sepsis neonatal y problemas postoperatorios de cirugía cardíaca [19, 20].

La alerta neonatal HeRO se centra en la disminución de la variabilidad y el aumento de la desaceleración de la frecuencia cardíaca para identificar una posible sepsis, mientras que la alerta Etiometry emplea un análisis estadístico sofisticado de los elementos monitoreados que reflejan la función cardiaca para detectar y definir problemas antes de lo que los humanos podrían hacerlo normalmente.

El equipo de HeRO está trabajando ahora para desarrollar análisis predictivos de deterioro respiratorio, hemorragia significativa y sepsis en adultos [21]. El punto esencial es que los monitores que emplean esta clase de análisis predictivo, así como transmisión y analítica retrospectiva, pueden

aprovechar grandes cantidades de datos personales para mejorar el proceso de monitorización, así como la experiencia de los encuentros en salud, en particular en áreas de calidad y seguridad. Sin embargo, es esencial que estas aplicaciones individuales, exponencialmente crecientes en complejidad y sofisticación, no sean introducidas como bits no relacionados en un sistema de salud ya sobrecargado de datos y subdesarrollado en infraestructura. En el estado actual del sistema de salud, ya hay muchos datos. Sin embargo, no se están manejando, utilizando y aprovechando sistemáticamente. Es esencial que las nuevas aplicaciones se integren cuidadosamente en los flujos de trabajo. También deben estar sistemáticamente interconectadas e interoperables con el centro del sistema de atención, representado por la próxima generación de HCEs, de modo que la información pueda ser utilizada de forma coordinada, auditada en términos de su impacto en los flujos de trabajo, y monitorizada en términos de su impacto sobre los resultados, la calidad y la seguridad de los pacientes. La incorporación de nuevos elementos al sistema debe ser planificada, monitoreada y evaluada en un formato basado en datos. Los nuevos elementos deben contribuir al sistema que utiliza los datos en una forma dirigida y bien gestionada, en lugar de simplemente recolectando la información. La introducción de elementos fuera del núcleo de la HCE requiere comunicación y coordinación entre todos los elementos del sistema, del mismo modo que el uso eficaz de la HCE requiere comunicación y coordinación entre cuidadores y pacientes.

4.4 Coordinación, comunicación y orientación a través del laberinto clínico

La coordinación y la comunicación serían propiedades fundamentales de un SAI en contraste con los enormes esfuerzos individuales que se requieren para lograr estos objetivos en la actualidad.

Los pacientes y cuidadores deben ser capaces de asumir que el sistema captura, almacena y comparte su información dónde y cuándo se necesita. Cuando el paciente se traslada del centro de atención del paciente crónico para ser atendido en una sala de urgencias local o por un neurólogo, los médicos deberían tener toda la información previa necesaria disponible para tratarlo. Este también debe ser el caso cuando regrese al centro de pacientes crónicos; el sistema debería actualizar su registro clínico con los eventos

sucedidos en la consulta con el médico y también debería implementar las nuevas indicaciones que reflejen dicha consulta.

Esta comunicación y coordinación sin fisuras es especialmente importante para aquellos tipos de pacientes que no pueden proporcionar esta información por sí mismos: las personas de edad avanzada, las personas con problemas cognitivos, aquellos gravemente enfermos, etc. Desafortunadamente, el sistema actual fue desarrollado como una herramienta para ayudar en la facturación y el reembolso de las intervenciones. El reto al que nos enfrentamos para transformarlo y continuar desarrollándolo como un SAI es hacer una transición de su enfoque al cuidado del paciente. Actualmente, los pacientes y sus familiares deben luchar contra los implacables desafíos de la opacidad y la obstrucción, enfrentándose a la inmensa frustración y las amenazas que pesan sobre la seguridad del paciente y la calidad de la atención cuando tales riesgos no se tolerarían en ningún caso en otras industrias.

Los datos y la transmisión eficiente de la información dónde y cuándo sea necesario están en el centro de un SAI. Deben crearse redes de información que impregnen todos los lugares relevantes utilizando todas las características de interoperabilidad, privacidad y seguridad necesarias. El sistema debe mantener su foco en el paciente y debe actualizar, sincronizar y trasmitir la información instantáneamente (o lo suficientemente rápido como para satisfacer las necesidades clínicas) a todos aquellos que necesiten disponer de ella, incluyendo la familia.

Muchos médicos pueden ser malinterpretados como indiferentes, o incluso negligentes, en respuesta a su continua frustración generada por la dificultad de obtener información precisa y oportuna. El estado actual de los sistemas de salud en silos hace que la obtención de información de otros lugares conlleve un desafío prohibitivo, sin que se consiga ninguna recompensa por seguir luchando para obtener información pertinente para la continuidad de cuidado de los pacientes, provocando reacciones de los cuidadores, incluyendo grosería, negligencia, hostilidad o agotamiento. Este desafío de obtener información de fuentes externas también conduce a la repetición de pruebas diagnósticas que exponen a los pacientes a riesgos y exposiciones innecesarias, como las que se observan cuando un paciente es transferido de una institución a otra, pero las imágenes obtenidas en la primera institución no pueden ser transferidas apropiadamente [22].

Desafortunadamente, el Acta de Portabilidad y Responsabilidad del Seguro de Salud de 1996 (HIPAA, por sus siglas en inglés), la misma legislación diseñada para permitir la portabilidad de la información relativa a la atención al paciente, ha dificultado aún más esta transmisión de información. Un sistema eficaz de comunicación y coordinación beneficiaría la experiencia del cuidador, además de los pacientes, al proporcionarles las herramientas y la información que necesitan para llevar a cabo su trabajo.

El alcance de las personas afectadas por los retos inherentes a la atención de la salud actualmente es amplio. No sólo afecta a los que tienen problemas cognitivos, sino también a los que tienen educación o recursos limitados. Afecta tanto a aquellos cuyas historias médicas son complicadas así como los que no tienen antecedentes.

Incluso cuando los pacientes son capaces de contribuir a la gestión de sus propios datos clínicos, existe el potencial de que sean abrumados e incapacitados a través de las complejidades del sistema cuando se ven afectados por una enfermedad, sin importar su agudeza, gravedad o complejidad.

Las Historias Clínicas Electrónicas (HCE) interoperables centradas en los pacientes y no en su localización o marcas podrían proporcionar información necesaria y actualizada a medida que el paciente se traslada del consultorio A al hospital B, a su domicilio y de regreso a la sala de emergencia C. Cuando las personas están enfermas, ellos y sus cuidadores deben ser apoyados por el sistema en lugar de ser forzados a luchar contra él.

El intercambio de datos entre pacientes y cuidadores de una manera segura y eficiente no es principalmente un problema técnico en este momento, aunque hay muchos desafíos técnicos para lograr la interoperabilidad sin fisuras. Es tanto un negocio como un problema político. Esta compleja interacción puede verse en los esfuerzos para que la arquitectura sanitaria y las normas soporten la interoperabilidad descripta en el Informe JASON, “ Una infraestructura de datos de salud robusta” con respuestas de la industria y proveedores de HCE en el desarrollo y adopción de estándares de interoperabilidad HL7 FHIR (del inglés, Fast Healthcare Interoperability Resources) [23, 24].

En un SAI, todas las partes deben cooperar para interconectar las HCEs entre los cuidadores y la población, de modo que se obtenga la información precisa y confiable esencial para brindar atención en salud en forma

coordinada, sincronizada y permita la comunicación entre los distintos dominios de la práctica, pero dentro del dominio de cada paciente. Al igual que hemos visto a nivel de pacientes individuales, una sobreabundancia de datos no es útil si no son procesados, analizados, colocados en el contexto apropiado, y disponibles para las personas adecuadas en los lugares y momentos adecuados.

4.5 Seguridad y calidad en un SAI

Hay muchos ejemplos en salud, como el caso de la sangría con sanguijuelas, donde lo que se pensó que era la mejor práctica, basada en el conocimiento o la evidencia en un tiempo, se descubrió más tarde que era perjudicial para los pacientes. Nuestros conocimientos y su aplicación deben estar en un estado continuo de evaluación y reevaluación, de modo que los elementos no verdaderos puedan ser identificados y se puedan tomar medidas antes de que se produzca daño, o al menos cuando éste sea mínimo [4].

En la actualidad no existe un acuerdo sobre las métricas estándar a utilizar para medir la seguridad y la calidad en la atención sanitaria y no vamos a intentar establecer definiciones estándar en este capítulo [25].

Sin embargo, para poder discutir estos temas, es importante establecer una comprensión común de la terminología y su significado.

A nivel conceptual, concebimos la seguridad clínica como un problema de optimización estratégica, en el que debe tenerse en cuenta el nivel máximo de actuación admisible implementado en el contexto simultáneo de permitir el mínimo grado de daños relacionados con el cuidado. El objetivo es diseñar e implementar un sistema de atención que minimice los riesgos de seguridad para acercarse a la meta de cero. La digitalización de la medicina ofrece una posibilidad realista de alcanzar este objetivo de manera eficiente y eficaz. La aplicación de los principios de ingeniería de sistemas también proporciona herramientas para diseñar este tipo de sistemas.

La calidad global de la asistencia sanitaria es una sumatoria de la experiencia de los individuos, y para estos individuos, puede haber diferentes grados de calidad para diferentes períodos de su experiencia. Al igual que la seguridad, también pensamos en la calidad como un problema de optimización estratégica en la que los resultados y beneficios se

maximizan u optimizan, mientras que los costos y riesgos involucrados en los procesos necesarios para lograrlos son minimizados.

La provisión de calidad a través de la optimización de los resultados en la atención clínica es, en gran medida un problema de ingenieria de la confiabilidad y el flujo de la información, proporcionando la mejor evidencia en el momento adecuado para ayudar a tomar las mejores decisiones [3].

Los conceptos de las `mejores evidencias' y `mejores decisiones' dependen en sí mismos de las fuentes consideradas que van desde ensayos aleatorizados controlados hasta la opinión de expertos, pasando por las mejores prácticas locales. Para proporcionar una acción real, los flujos de información deben ser complementados por modalidades químicas (medicamentos), mecánicas (cirugía, fisioterapia, inyecciones, tacto humano) y electromagnéticas (imágenes, ultrasonido, radioterapia, habla humana), que pueden definir los procesos indicados por esos flujos de información.

Además, la calidad también puede definirse en función del grado de éxito del tratamiento de la enfermedad. Las enfermedades que se abordan en la medicina moderna son, sorprendemente en gran medida, aquellas de los problemas de control en bioingeniería [10]. Estas enfermedades pueden originarse en problemas de control que afectan a la inflamación, el metabolismo, la homeostasis fisiológica o el genoma. Sin embargo, todos ellos representan un fallo en un elemento o elementos de un sistema biológico normalmente bien controlado. La calidad de la respuesta clínica a estos fallos se optimiza comprendiéndolos de manera adecuada y lo suficientemente exhaustiva para que se puedan desarrollar tratamientos tolerables que controlen y/o eliminen la disfunción de los sistemas representados por la enfermedad clínica. Esto debe lograrse de una manera que minimice los costos indebidos de sufrimiento físico, mental o incluso espiritual.

En última instancia, la calidad médica se basa principalmente en los resultados, pero la naturaleza de los procesos que conducen a esos resultados debe ser considerada. Los resultados óptimos son deseables, pero no a cualquier precio, en la definición amplia del término. Por ejemplo, prolongar la vida indefinidamente no es un resultado óptimo en algunas circunstancias que son definidas contextualmente por preferencias individuales, familiares y culturales.

Una vez definida la seguridad y la calidad en nuestro contexto, el siguiente paso es desarrollar sistemas que capturen, rastreen y gestionen estos conceptos en retrospectiva, en tiempo real y en forma predictiva.

Sólo cuando sabemos con precisión que elementos estáticos y dinámicos de seguridad y calidad, queremos asegurarnos, es que podemos diseñar los sistemas para dar soporte a estos esfuerzos. Estos sistemas implicarán la integración de hardware y sistemas de software tales como monitores fisiológicos con la HCE (incluyendo la entrada de órdenes computarizadas, sistemas de archivo y comunicación de imágenes), y requerirá una variedad de análisis de datos especializada y específica para cada dominio así como innovaciones técnicas como las redes inalámbricas de sensores corporales para capturar el estado del paciente en tiempo real. El sistema conectará y comunicará información pertinente entre los cuidadores alimentando en forma estandarizada accesos esenciales y nodos de alerta con información oportuna y precisa. También es necesario que la información fluya en forma bidireccional (de los registros de los individuos a la población y del registro de la población a los individuos) para que ambos se puedan beneficiar de los datos [2,14]. Claramente, esto requerirá un sistema de información y monitoreo global que sea interoperable, interactivo tanto con sus propios componentes como con sus usuarios, informativo en forma activa pero selectiva. Las futuras generaciones de médicos clínicos se educarán en un ambiente en el que estos sistemas sean omnipresentes y modificables en forma selectiva de acuerdo a entradas como el "*crowdsourcing"*, e intrínseco a las tareas habituales, en contraste con las aplicaciones actuales estancas e impuestas en forma aparentemente arbitraria que los profesionales actuales pueden resistir y resentir [5,8]. Nosotros hemos observado la importancia de los problemas de control en la enfermedad y el control también representará un componente fundamental en el diseño de los futuros sistemas de seguridad y calidad. La detección y prevención de eventos adversos es un desafío significativo cuando dependemos de métodos de auto reporte o revisión de historias clínicas, este aspecto es de gran importancia en los Estados Unidos [26, 27]. Es posible desarrollar análisis predictivos como elementos de los sistemas, para informar a los usuarios en forma prospectiva de las amenazas de seguridad y calidad [19-21]. Los componentes cuidadosamente diseñados en forma prospectiva informarán a los participantes en tiempo real que está ocurriendo una actividad de alto

riesgo, de forma tal que pueda ser rectificada sin requerir un análisis retroactivo (Figura 4.4– Circuito de seguridad). Los análisis retrospectivos de datos rastrearán los factores que afectan la calidad y la seguridad de forma tal que la práctica, el flujo de trabajo y los sistemas tecnológicos se puedan modificar en forma acorde. Este SAI será capaz de monitorear errores médicos, eventos adversos, preocupaciones y métricas regulatorias y de seguridad y cumplimiento de las mejores prácticas así como su uso racional, en paralelo con costos y resultados.

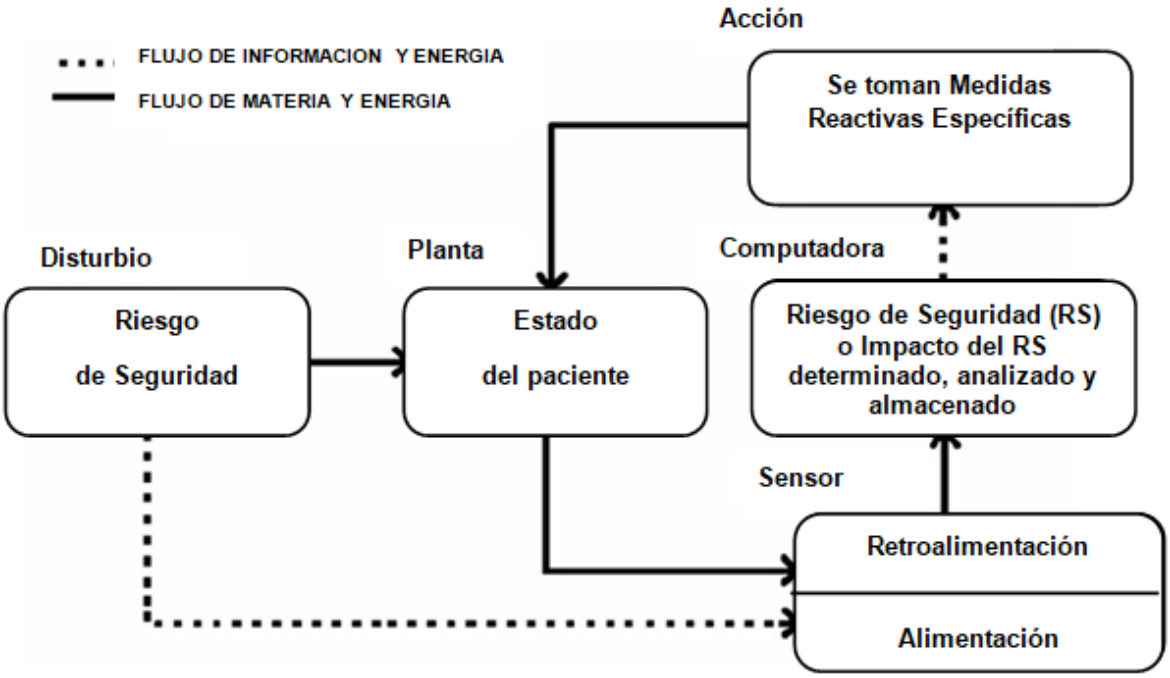

Fig. 4.4 Circuito de control que muestra un sistema de seguridad basado en datos.
Un problema de seguridad clínica afecta el estado del paciente. Posteriormente, el sensor del sistema detecta este cambio y envía los datos relevantes a la computadora para su almacenamiento y análisis. Esto puede o no dar lugar a la activación de una intervención compensatoria que afecta aún más al estado del paciente y que retroalimenta el sistema para un mayor análisis. El control de la retroalimentación implica la transmisión de las alteraciones directamente al sensor sin afectar primero el estado del paciente. Un ejemplo de tal retroalimentación incluye un dispositivo defectuoso o un riesgo biológico.

4.6 Conclusiones

Las soluciones básicas sistémicas para los problemas de datos en salud se basan en abordar en forma completa e inclusiva a los ejes del paciente, el

cuidador y las consideraciones de los sistemas de atención, que por momentos son aparentemente independientes pero que en última instancia son interactivos e interdependientes. Los sistemas de diseño necesarios se beneficiarán en gran medida de la incorporación básica de los elementos fundamentales de control de ingeniería como el sensado, la computación, acción y retroalimentación efectiva y basada en datos. Un SAI debe ser diseñado en forma cuidadosa e intencional antes que permitir su evolución basada en presiones de mercado y conveniencia del usuario. Los datos de los pacientes deben ser precisos, completos y actualizados. A medida que los pacientes progresan en el tiempo, sus registros deben ser actualizados en forma adecuada y oportuna con nuevos datos mientras que en forma concurrente los datos antiguos son modificados y/o borrados si se trasforman en irrelevantes o inexactos. Se deben planificar y tomar en consideración nuevas vías de entrada como datos generados por el paciente y datos generados en forma remota, como datos genómicos. Estos datos deben ser accesibles en forma segura, confiable y fácil para los usuarios apropiados, incluyendo al paciente. El cuidador debe tener acceso a estos datos por medio de una aplicación bien diseñada que apoye en forma positiva el proceso de documentación clínica e incluya modalidades de soporte de decisión clínica que reflejen la mejor evidencia, datos históricos de casos similares en la población así como los datos longitudinales del paciente. Todos deberían tener acceso a la información en la medida en que se utilice para construir los patrones actuales e históricos de seguridad y calidad. Además de los datos de los individuos, se necesita el acceso a los datos de las poblaciones para los propósitos mencionados y para brindar intervenciones efectivas en situación de emergencia como epidemias. La creación de este tipo de solución multimodal sistémica (Fig. 4.5-Arquitectura de un Sistema de Atención Ideal) requerirá del aporte de una gran variedad de expertos incluyendo aquellos en HCE, dispositivos de monitoreo, almacenamiento de datos e industria del análisis de datos junto con líderes en legislación en salud, decisores, reguladores de políticas y administradores. Existen muchas e importantes preguntas de ingeniería, económicas y políticas que no son consideradas en este capítulo. ¿Qué y quién proveerá la infraestructura y quién pagara por ella? ¿Esta clase de sistema funcionará con los hardware y software actuales o necesitará actualizaciones

fundamentales en función del grado requerido de confiabilidad y seguridad? ¿Cómo y dónde se incluirán los controles en este sistema?

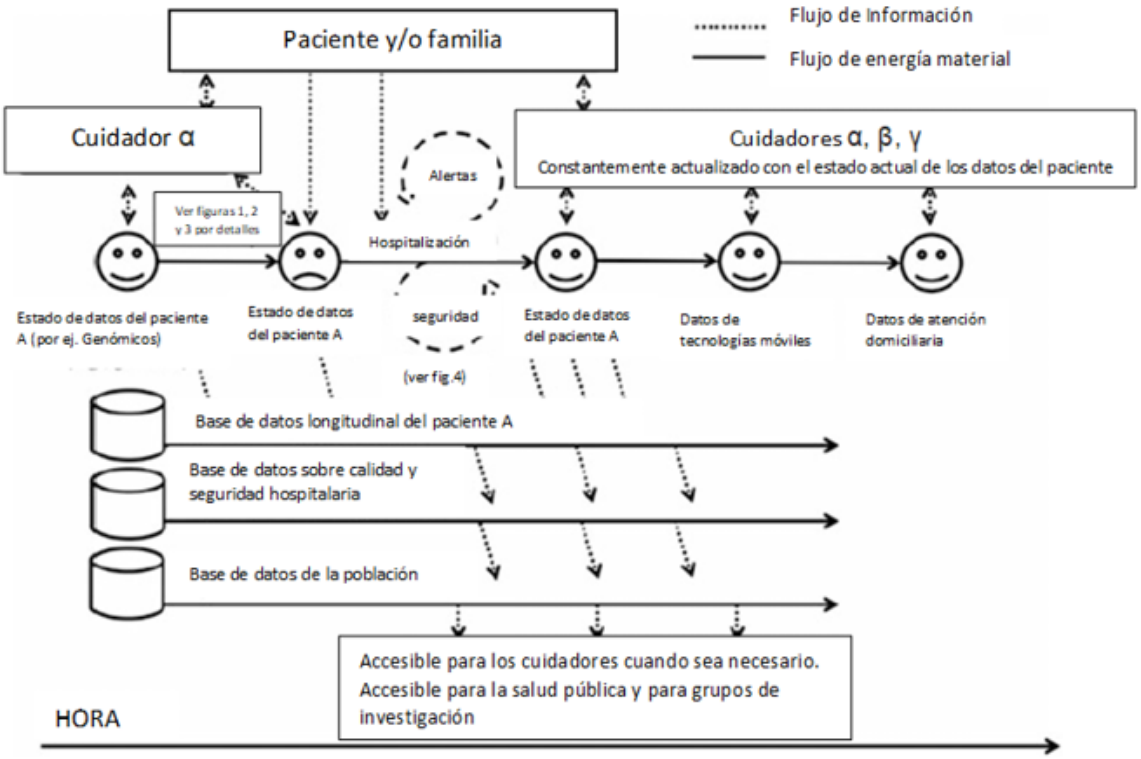

Fig. 4.5 Arquitectura de la información de un sistema de atención ideal (SAI).
Este diagrama integra los conceptos presentados en este capítulo que describen los sistemas de atención basados en datos, los sistemas de seguridad, junto la conexión y coordinación de los datos de los pacientes a través de múltiples modalidades para lograr un SAI. Los pacientes se mueven a través del tiempo e interactúan con el SAI en diferentes contextos. Bases de datos paralelas son integradas con la información de los pacientes en el tiempo, incluyendo los datos longitudinales de un paciente individual, bases de datos de calidad y seguridad de los hospitales y bases de datos poblacionales. Los datos del paciente, de las tecnologías móviles y de las empresas de atención domiciliaria mantienen a los cuidadores informados del estado actual de los datos del paciente.

Por ejemplo, estarán en un nivel individual de monitoreo inteligente o en un nivel más amplio de salud pública? ¿Cómo se manejarán los metadatos obtenidos para el bien de los individuos y de las poblaciones? Es crítico que la inclusión de nuevas modalidades y dispositivos se encuentre integrada completamente en el sistema antes que sumar componentes aislados que pueden contribuir con mayor complejidad y confusión que beneficio. Estos objetivos requerirán de una cooperación previamente no vista entre

competidores reales y potenciales y aquellos que previamente han podido trabajar en relativo aislamiento.

Nota: Las imágenes de este capítulo se encuentran disponibles a color en el ebook; disponible en www.hardineros.com

Referencias

1. Celi LA, Mark RG, Stone DJ, Montgomery R (2013) “Big data” in the ICU: closing the data loop. Am J Respir Crit Care Med 187 (11): 1157-1160.
2. Celi LA, Zimolzak AJ, Stone DJ (2014) Dynamic clinical data mining: search engine-based clinical decisión support. J Med Internet Res Med Inform 2 (1): e13. doi: 10.2196/medinform. 3110.
3. Celi LA, Csete M, Stone D (2014) Optimal data systems: the future of clinical predictions and decisión support. Curr Opin Crit Care 20:573-580.
4. Moseley ET, Hsu D, Stone DJ, Celi LA (2014) Beyond data liberation: addressing the problem of unreliable research. J Med Internet Res 16 (11): e259.
5. Celi LA, Ippolito A, Montgomery R, Moses C, Stone DJ (2014) Crowdsourcing knowledge discovery and innovations in medicine. J Med Internet Res 16 (9): e216. doi: 10.2196/jmir.3761.
6. Celi LA, Moseley E, Moses C, Ryan P, Somai M, Stone DJ, Tang K (2014) From pharmacovigilance to clinical care optimization. Big Data 2 (3): 134-141. doi: 10.1089/big. 2014.0008.
7. Badawi O, Brennan T, Celi LA, Feng M, Ghassemi M, Ippolito A, Johnson A, Mark RG, Mayaud L, Moody G, Moses C, Naumann T, Pimentel M, Pollard TJ, Santos M, Stone DJ, Zimolzak AJ (2014) Making big data useful for health care: a summary of the inaugural MIT critical data conference. J Med Internet Res Med Inform 2 (2): e22. doi: 10.2196/medinform. 3447.
8. Moskowitz A, McSparron J, Stone DJ, Celi LA (2015) Preparing a new generation of clinicians for the era of big data. Harvard Med Student Rev 2 (1): 24-27.

9. Ghassemi M, Celi LA, Stone DJ (2015) The data revolution in critical care. Ann Update Intensive Care Emerg Med 2015 (2015): 573-586.
10. Stone DJ, Csete ME, Celi LA (2015) Engineering control into medicine. J Crit Care. Published Online: January 29,2015. doi: 10.1016/j. jcrc.2015.01.019.
11. Health Information Technology for Economic and Clinical Health (HITECH) Act, Title XIII of División A and Title IV of División B of the American Recovery and Reinvestment Act of 2009 (ARRA), Pub.L. No. 111-5,123 Stat. 226 (Feb. 17,2009), codified at 42 U.S.C. §§300jj et seq.; §§17901 et seq.
12. Horstmanshoff HFJ, Stol M, Tilburg C (2004) Magic and rationality in ancient near Eastern and Graeco-Roman medicine, pp 97-99. Brill Publishers. ISBN 978-90-04-13666-3.
13. Bell SK, Folcarelli PH, Anselmo MK, Crotty BH, Flier LA, Walker J (2014) Connecting Patients and Clinicians: The Anticipated Effects of Open Notes on Patient Safety and Quality of Care. JtComm J Qual Patient Saf 41 (8): 378-384 (7).
14. Celi LA, Marshall JD, Lai Y, Stone DJ, Physician documentation and decisión making in the digital era. J Med Internet Res Med Inform (Forthcoming).
15. Longhurst CA, Harrington RA, Shah NH (2014) A 'green button' for using aggregate patient data at the point of care. Health Aff 33 (7): 1229-1235.
16. Friedman C, Rubin J, Brown J et al (2014) J Am Med Inform Assoc 0:1-6. doi: 10.1136/amiajnl-2014-002977.
17. Institute of Medicine (2012) Best care at lower cost: the path to continuously learning health care in America. Extraído de: http://iom. nationalacademies.org/Reports/2012/Best-Care-at-Lower-Cost-The-Path-to-Continuously-Learning-Health-Care-in-America.aspx.
18. The Joint Commissión (2014) National Patient safety goal on alarm management 2013.
19. www.etiometry.com
20. www.heroscore.com
21. Personal communication, Randall Moorman, MD.
22. Sodickson A, Opraseuth J, Ledbetter S (2011) Outside imaging in emergency departmenttransfer patients: CD import reduces rates of subsequent imaging utilization. Radiology 260 (2): 408-413.
23. A Robust Health Data Infrastructure. (Prepared by JASON at the MITRE Corporation under Contract No. JSR-13-700). Agency for Healthcare Research and Quality, Rockville, MD. April 2014. AHRQ Publication No. 14-0041-EF.
24. Health Level Seven® International. HL7 Launches Joint Argonaut Project to Advance FHIR.N.p. , 4 Dec. 2014. Web. 31 Aug. 2015. Disponible en http://www.hl7.org/documentcenter/public-temp-32560CB2-1C23-BA17-0CBD5D492A8F70CD/pressreleases/HL7-PRESS-20141204.pdf.
25. Austin JM et al (2015) National hospital ratings systems share few common scores and may generate confusión instead of clarity. Health Aff 34 (3): 423-430. doi: 10.1377/hlthaff.2014.0201.

26. Elton GEBM, Ripcsak GEH (2005) Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc 12:448-458.
27. Kohn LT, Corrigan JM, Donaldson MS (eds) (2000). To err is human: building a safer health system, vol 2. National Academy Press, Washington, DC.

CAPÍTULO 5

LA HISTORIA DE MIMIC

ROGER MARK

Puntos clave

- MIMIC (*Medical Information Mart for Intensive Care*) es un Data mart de Información Médica en Cuidados Intensivos compuesto por un gran conjunto de datos compartidos dentro del ambiente de la terapia intensiva, con un alto nivel de precisión y detalle, que soporta la implementación de algoritmos complejos de procesamiento de señales y consulta de datos, que facilitan la detección temprana de problemas complejos, proveen una orientación útil a la hora de realizar intervenciones terapéuticas, y de esta manera, conducen a mejoras en los resultados del paciente.
- Este complicado esfuerzo requirió de una comprometida y coordinada colaboración de instituciones académicas, de la industria y del sector salud, para proveer una base de datos de libre acceso para investigadores en todo el mundo.

5.1 La visión

Los pacientes en las unidades de cuidados intensivos (UCIs) son fisiológicamente frágiles e inestables, generalmente tienen enfermedades que amenazan la vida, y requieren un monitoreo cercano e intervenciones terapéuticas tempranas. Están conectados a un despliegue de equipamiento y monitores, y están cuidadosamente atendidos por el equipo médico. En forma diaria, se registra una cantidad abrumadora de información por cada paciente en una UCI: datos en forma de onda multicanal muestreada cientos de veces por segundo (señales fisiológicas), series de signos vitales actualizadas cada minuto, resultados de laboratorio, resultados de imágenes, registros de medicaciones y administración de fluidos, alarmas y alertas, notas del personal de salud y más. A principios del año 2000, nuestro grupo en el Laboratorio de Fisiología Computacional (LFC) en el MIT reconoció que la riqueza y el detalle de la información recolectada, generaba la oportunidad de crear una nueva generación de sistemas de

monitoreo para seguir el estado fisiológico del paciente, utilizando el poder del procesamiento moderno de señales, el reconocimiento de patrones, el modelado computacional y razonamiento clínico basado en el conocimiento. A largo plazo, deseábamos diseñar sistemas de monitoreo que no sólo sintetizaran y reportaran todas las mediciones relevantes a los clínicos, sino que además formaran hipótesis fisiopatológicas que explicaran mejor la información observada. Dichos sistemas permitirían la detección temprana de problemas complejos, aportarían una guía útil en intervenciones terapéuticas, y finalmente llevarían a mejores resultados en los pacientes.

También era claro que a pesar de los petabytes de información que son obtenidos diariamente durante la estadía en las UCI del país, la mayor parte de esta información no estaba siendo utilizada para generar evidencia o para descubrir nuevo conocimiento. Por lo tanto, el desafío era emplear la tecnología existente para recolectar, almacenar y organizar la información finamente detallada de la UCI, resultando en una fuente de información con un potencial enorme para crear nuevo conocimiento médico, nuevas herramientas de apoyo de decisiones, y nueva tecnología para la UCI. Propusimos desarrollar y hacer pública una base de datos “sustancial y representativa” generada a partir de pacientes con patología compleja médica y quirúrgica en la UCI.

5.2 Obtención de datos

En el año 2003, con colegas de la academia (Massachusetts Institute of Technology), industria (Philips Medical Systems), e instituciones médicas (Beth Israel Deaconess Medical Center, BIDMC) recibimos fondos del NIH (National Institutes of Health) para lanzar el proyecto “Integración de Señales, Modelos y Razonamientos en la Terapia Intensiva”, cuyo principal objetivo fue crear una base de datos masiva de investigación en terapia intensiva. El estudio fue aprobado por la Junta de Revisión Institucional del BIDMC (Boston, MA) y el MIT (Cambridge, MA). El requerimiento del consentimiento de cada paciente fue eximido ya que el estudio no afectaría la atención médica y toda la información médica protegida debía ser des-identificada.

Nos propusimos recopilar información clínica y fisiológica completa de todos los pacientes internados en las UCIs de adultos de nuestro hospital (BIDMC). El registro de cada paciente comenzó al momento de la admisión en la UCI y finalizó con su alta del hospital. El proceso de obtención de información fue continuo e invisible para el equipo médico tratante. Esto no afectó la atención de los pacientes ni su monitoreo. Se recogieron tres categorías diferentes de datos: *datos clínicos,* los cuales fueron obtenidos a partir del sistema de información de las UCIs y del hospital; *datos fisiológicos de alta resolución* (análisis de señales y series temporales de signos vitales y alarmas obtenidas por monitores clínicos); y *Datos de mortalidad* de los Registros de Mortalidad de la Administración del Seguro Social de los EE.UU. (Ver Fig. 5.1)

5.2.1 Información Clínica

La información clínica de los pacientes fue descargada a partir de archivos del Sistema de Información Clínica CareVue (Philips Healthcare, Andover, MA) usado en las UCIs. Los datos adicionales fueron obtenidos de extensos archivos digitales del hospital. Las clases de datos incluían

- **Demografía de los pacientes**
- **Datos administrativos hospitalarios:** admisión/alta/fechas de muerte, seguimiento en sala, datos de facturación
- **Datos fisiológicos:** signos vitales horarios, puntajes de severidad clínica, parámetros de asistencia respiratoria mecánica, etc.
- **Medicación:** medicación IV, órdenes médicas
- **Datos de laboratorio:** química, hematológicos, gasometría arterial, microbiología, etc.
- **Balance hídrico**
- **Notas e informes:** resúmenes de alta, notas de evolución, ECG, reportes de ecografías e imágenes.

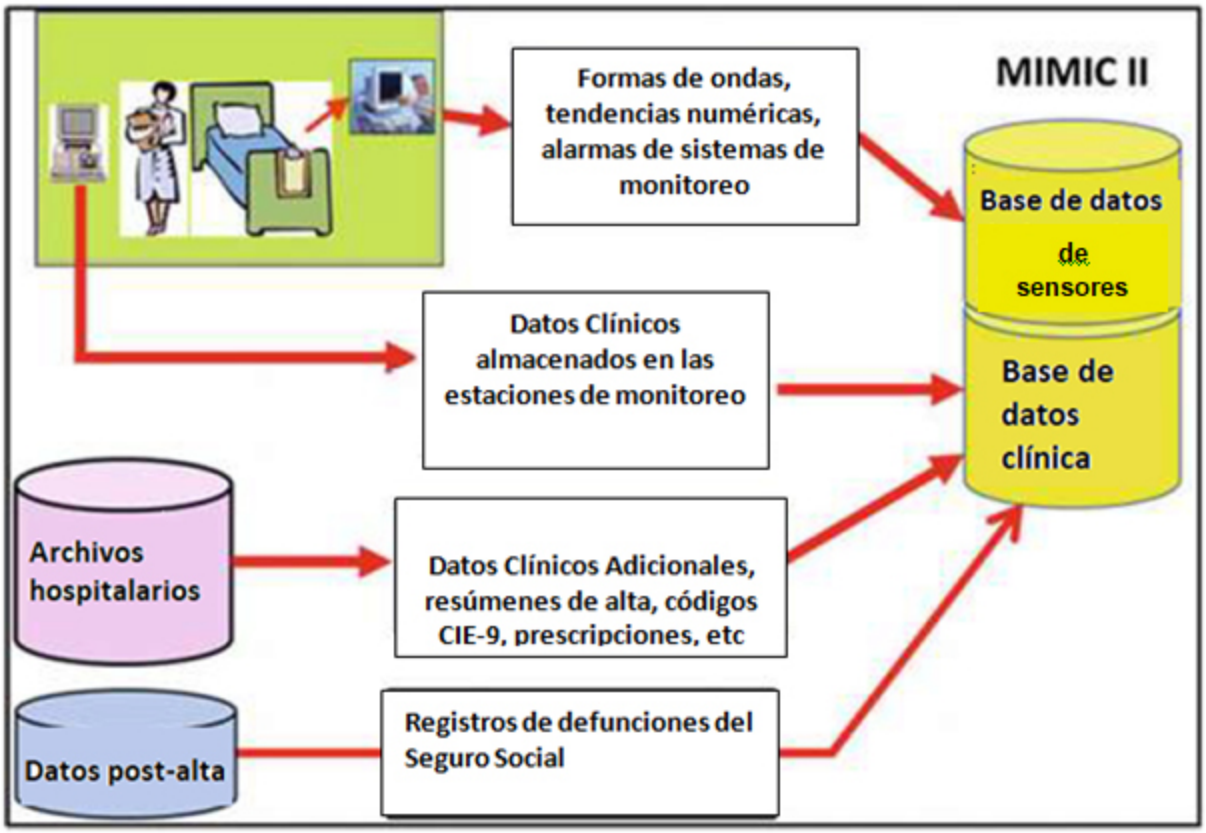

Figura 5.1 Fuentes de datos de MIMIC II

5.2.2 Datos fisiológicos

Los datos fisiológicos fueron obtenidos gracias a la asistencia técnica de los proveedores del sistema de monitoreo. Los equipos de monitoreo estaban ubicados en todas las camas de los pacientes que se encontraban en la UCI. Cada monitor recopilaba y digitalizaba información multiparamétrica de forma de onda o señales fisiológicas, procesaba las señales para derivar series temporales (tendencias) de mediciones clínicas como frecuencia cardíaca, presión arterial, saturación de oxígeno, etc., y también producía alarmas de monitoreo. Las formas de onda (como electrocardiogramas, presión arterial, curva pletismográfica por oximetría de pulso, respiración) fueron adquiridas a una frecuencia de muestreo de 125 Hz y los datos de tendencia obtenidos fueron actualizados a cada minuto. Posteriormente, los datos fueron almacenados en forma temporal en un servidor central que apoyaba varias UCIs. Un módulo de software creaba y almacenaba de forma permanente copias de los datos fisiológicos. La información era transportada físicamente desde el hospital a los

laboratorios cada 2-4 semanas donde era des-identificada, convertida a un formato de datos de código abierto e incorporada a la base de datos de señales de MIMIC II. Desafortunadamente, la capacidad de almacenamiento limitada y fallas esporádicas de los módulos de almacenamiento limitó la adquisición de datos fisiológicos a una fracción de las camas monitoreadas en las UCIs.

5.2.3 Datos de mortalidad

Los Registros de Mortalidad de la Administración del Seguro Social (ASS) fueron usados para documentar las fechas de la muerte de los pacientes que fueron dados de alta con vida del hospital. Esta información es importante para estudios de mortalidad a 28 días y 1 año.

5.3 Organización e integración de los datos

Se requirió un mayor esfuerzo para organizar la diversidad de información obtenida en una base de datos relacional bien documentada, y que contuviera todos los registros médicos integrados de cada paciente. A través de la base de datos de los hospitales, los pacientes eran identificados por su Número de Registro Médico y por su Número Fiscal (este último provee una identificación única de hospitalización para pacientes que podrían haber sido admitidos en múltiples oportunidades), lo que nos permitía integrar la información de varias fuentes hospitalarias. Finalmente, la información fue organizada en una base de datos relacional integral y exhaustiva. En el sitio web de MIMIC-II se encuentra disponible más información acerca de la integración de la base de datos, en particular cómo se aseguró la integridad de la base de datos [1]. La guía para el usuario también se encuentra disponible en forma online [2].

Una tarea adicional fue convertir los datos de los pacientes con forma de señales de un formato de propiedad de Philips a un formato de código abierto. Con la asistencia de los proveedores del equipamiento biomédico, se convirtieron las señales, tendencias y alarmas al formato de datos abierto, WFDB, que es usado para las bases de datos disponibles públicamente en el sitio web PhysioNet, financiado por el NIH (National Institute of Health) [3].

Todos los datos que fueron integrados en la base de datos MIMIC-II fueron des-identificados cumpliendo los estándares del Acta de Portabilidad y Responsabilidad del Seguro Médico (HIPAA) para facilitar el acceso público a MIMIC-II. La eliminación de información de salud protegida de las fuentes de datos fue directa (por ejemplo los campos de la base de datos que proveen el nombre del paciente, la fecha de nacimiento, etc.). También eliminamos la información de salud protegida de los resúmenes de alta, reportes diagnósticos, y aproximadamente 700.000 textos libres de enfermería y notas respiratorias en MIMIC-II usando un algoritmo automatizado que demostró un rendimiento superior para detectar información protegida de salud en comparación con el personal clínico [4]. Este algoritmo abarca el amplio espectro de estilos de escritura en nuestro set de datos, incluyendo variaciones personales en sintaxis, abreviaciones, y ortografía. Publicamos el algoritmo en forma de código abierto como una herramienta general para ser usada por otros para la de-identificación de notas de texto libre [5].

5.4 Intercambio de datos

MIMIC-II es un recurso sin precedentes e innovador de investigación abierta que garantiza, a investigadores de todo el mundo, acceso gratuito a información altamente desglosada de UCIs, y en el proceso, acelera sustancialmente la creación de conocimiento en el área de la medicina crítica. La base de señales MIMIC se encuentra disponible de manera gratuita a través del portal de PhysioNet, sin necesidad de registrarse. La base de datos clínicos MIMIC también está disponible sin costo. Sin embargo, para restringir su acceso a investigadores clínicos legítimos, se requiere completar un simple acuerdo de uso y probar que el investigador ha completado el entrenamiento en investigación en sujetos humanos [6].

La base de datos clínicos MIMIC-II está disponible en dos formas. En la primera forma, los investigadores interesados pueden obtener una versión de la base de datos en un archivo de texto, y el esquema asociado que les permite reconstruirla usando el sistema de gestión de bases de datos de su elección. En la segunda forma, los investigadores interesados pueden obtener acceso limitado a la base de datos a través de QueryBuilder, un servicio web protegido por contraseña. Las búsquedas en la base de datos

usando QueryBuilder permiten a los usuarios familiarizarse con las tablas y programar consultas utilizando el lenguaje SQL ("Structured Query Language"). Sin embargo, los resultados de las consultas están limitados a 1000 filas por los recursos limitados de nuestro laboratorio. Acceder y procesar información de MIMIC-II es complejo. Por lo tanto se recomienda que los estudios basados en la base de datos MIMIC-II sean conducidos en forma colaborativa por expertos en clínica, estadística y en bases de datos relacionales. En la página web se encuentra disponible documentación detallada y procedimientos para obtener acceso a MIMIC-II [1]. La versión actual de MIMIC-II es 2.6, y contiene aproximadamente 36.000 pacientes, incluyendo aproximadamente 7.000 neonatos, y cubriendo el período 2001-2008. En este momento aproximadamente 1.700 individuos alrededor del mundo en academia, industria y medicina fueron acreditados para acceder a MIMIC-II, y están produciendo resultados de investigación en procesamiento de señales fisiológicas, soporte en la toma de decisiones clínicas, algoritmos predictivos en cuidados críticos, farmacovigilancia, procesamiento de lenguaje natural y más.

5.5 Actualización

En el año 2008 el hospital realizó un gran cambio en la tecnología del sistema de información y en los procedimientos de documentación en la UCI. El sistema Philips CareVue fue reemplazado por tecnología iMDsoft's Meta Visión. En el año 2013 empezamos una gran actualización de MIMIC para incorporar información de UCI de adultos del período 2008-2012. El esfuerzo requirió aprender del esquema completamente nuevo de Meta Visión, y la fusión del nuevo formato de información con el diseño de MIMIC existente. La información de Meta Visión incluía nuevos elementos como notas de evolución clínica, registros de administración de medicación oral o intravenosa, etc. Los datos actualizados fueron extraídos de los archivos hospitalarios y de los archivos de defunción de ASS para los nuevos pacientes. Se invirtieron casi dos años de esfuerzo en adquirir, organizar, limpiar, normalizar y documentar la nueva base de datos antes de publicarla. MIMIC-III incluye 20.000 nuevas admisiones a UCI de adultos, y un total de aproximadamente 60.000. La nueva base de datos es conocida

como MIMIC-III, y el acrónimo fue renombrado como "**M**edical **I**nformation **M**art for **I**ntensive **C**are".

5.6 Soporte

El soporte de la base de datos MIMIC incluye: acreditación de nuevos usuarios, administración de la lista de usuarios autorizados (usuarios que hayan firmado el acuerdo de uso y hayan obtenido permiso para acceder a MIMIC-II), creación de cuenta de usuario, restauración de contraseñas y concesión/revocación de permisos. Los servidores que proveen MIMIC-II incluyen autenticación, aplicación, base de datos y servidores web. Todos los sistemas deben ser monitoreados, mantenidos, actualizados y resguardados; la carga del mantenimiento continúa aumentando a medida que crece el número de usuarios. El personal de ingeniería en LFC intenta contestar preguntas de usuarios a demanda. Las preguntas comunes son incluidas en una lista de preguntas frecuentes en el sitio web de MIMIC y regularmente actualizamos nuestra documentación online.

5.7 Lecciones Aprendidas

Crear y distribuir una base de datos como MIMIC es desafiante, complejo y requiere de la cooperación y ayuda de un gran número de personas e instituciones.

A continuación mencionamos una lista de algunos de los requerimientos más importantes (Tabla 5.1)

Tabla 5.1 Requisitos para los datos de salud

1- Accesibilidad de los datos de UCI y hospitalarios digitalizados, incluyendo datos clínicos estructurados y no estructurados, formas de onda de alta resolución y datos de signos vitales
2- Un departamento hospitalario de Tecnología de la Información cooperador y colaborador para asistir en la extracción de datos
3- Un Comité de Investigación Institucional y una administración hospitalaria colaboradores que aseguren la protección de la privacidad del paciente y la liberación de datos desidentificados a la comunidad investigadora

4- Adecuada ingeniería y ciencia de datos capacitada para diseñar e implementar el esquema de base de datos y desidentificar los datos (incluyendo los datos de texto desestructurados)
5- Experiencia sofisticada en procesamiento de señales para reformatear y manejar los flujos de datos con forma de onda
6- La cooperación y el apoyo de los vendedores de equipos
7- Facilidades computacionales adecuadas para el almacenamiento y la distribución de los datos
8- Personal técnico y administrativo adecuado para proveer la acreditación y el soporte del usuario
9- Adecuado soporte financiero

5.8 Direcciones futuras

La base de datos MIMIC-III es un potente y flexible recurso de investigación, pero la generalización de los estudios basados en MIMIC es algo limitada por el hecho de que los datos son obtenidos de una sola institución. Los datos multicéntricos podrían tener la ventaja de incluir una variabilidad de práctica médica más amplia, y por supuesto un mayor número de casos. Los datos de instituciones internacionales agregarían más robustez a la base de datos debido a una variación incluso mayor de prácticas y de población de pacientes.

Nuestro objetivo a largo plazo es crear un archivo de datos público, multicentrico e internacional para la investigación en cuidados críticos. Imaginamos una inmensa base de datos de UCI en alta resolución y detallada, que contengan registros médicos completos de pacientes de todo el mundo. La dificultad de semejante proyecto no puede ser subestimada; sin embargo nos proponemos sentar las bases para dicho sistema desarrollando un marco operativo que pueda incorporar fácilmente datos de múltiples instituciones, capaz de apoyar la investigación de cohortes de paciente críticamente enfermos en todo el mundo.

Agradecimientos El desarrollo y mantenimiento de los recursos de MIMIC y Physionet ha sido financiado por el *National Institute of Biomedical Imaging and Bioengineering* (NIBIB) y el *National Institute of General Medicine* (NIGMS) por el período de 2003 a la actualidad. Grants R01EB1659, R01EB017205, R01GM104987, y U01EB008577.

Nota: Las imágenes de este capítulo se encuentran disponibles a color en el ebook; disponible en www.hardineros.com

Referencias

1. MIMIC-II Web Site. https://archive. physionet.org/mimic2/.
2. MIMIC User Guide. https://archive. physionet.org/mimic2/UserGuide/.
3. Wave Form DataBase Data Format. https://archive. physionet.org/physiotools/wfdb.shtml.
4. Neamatullah I, Douglass M, Lehman LH, Reisner A, Villarroel M, Long WJ, Szolovits P, Moody GB, Mark RG, Clifford GD (2008) Automated de identification of free-text medical records. BMC Med Inform Decis Mak 8:32. doi: 10.1186/1472-6947-8-327.
5. Deidentification Software. http://www.physionet.org/physiotools/deid/.
6. Accessing MIMIC. https://archive. physionet.org/mimic2/mimic2-access.shtml.
7. MIMIC-III Website. http://mimic.physionet.org/.

Capítulo 6

INTEGRANDO DATOS NO CLÍNICOS CON HISTORIAS CLÍNICAS ELECTRÓNICAS

Yuan Lai, Edward Moseley,
Francisco Salgueiro y David Stone

Puntos clave

- Los factores no clínicos contribuyen en forma significativa con la salud de un individuo; proveer estos datos a los médicos clínicos podría dar información acerca del contexto, orientación y tratamientos.
- La gestión de los datos será esencial para proteger la información médica confidencial conservando al tiempo los beneficios de un sistema de salud integrado.

6.1 Introducción

La definición de datos "clínicos" se está expandiendo, ya que un dato se convierte en clínico una vez que tiene relación con un proceso de enfermedad. Por ejemplo: la accesibilidad al domicilio de un individuo clásicamente se definiría como un dato no clínico, pero en el contexto de un paciente con una discapacidad, este dato puede volverse clínicamente relevante, e ingresarse en el registro de la consulta al igual que la presión arterial y la temperatura corporal del paciente. Sin embargo, incluso con este simple ejemplo podemos vislumbrar algunos problemas con los datos tradicionalmente no clínicos cuando son reclasificados como clínicos, particularmente debido a su complejidad.

6.2 Factores no clínicos y determinantes de la salud

Los factores no clínicos ya se encuentran ligados significativamente a la salud. Muchas políticas de salud pública enfocadas en el transporte, la recreación, los sistemas de alimentación y desarrollo comunitario están basados en la relación entre la salud y sus determinantes no clínicos como factores conductuales, sociales y ambientales [1]. Los factores conductuales como la actividad física, la dieta, el tabaquismo y el consumo de alcohol están altamente relacionados con la epidemia de obesidad [2]. Parte de esta información, como el consumo de alcohol y tabaco, es documentada

habitualmente por los médicos clínicos. Otra información, como las conductas alimentarias y la actividad física, no es capturada típicamente, pero puede ser registrada por nuevas tecnologías (como computadoras portátiles) e integrada en las historias clínicas electrónicas (HCE). Esos esfuerzos pueden proveer a los clínicos un contexto adicional con el cual asesorar a los pacientes en un esfuerzo para aumentar su actividad física y alcanzar un resultado de salud deseado.

Desde una perspectiva de salud pública, los mismos datos obtenidos desde estos dispositivos pueden ser agregados y usados para guiar decisiones en políticas de salud pública. Continuando con el ejemplo anterior, cantidades adecuadas de actividad física contribuirán a disminuir los índices de mortalidad y enfermedades crónicas, incluyendo la enfermedad coronaria, la hipertensión, la diabetes, el cáncer de mama y la depresión en toda la población. Estos datos pueden ser utilizados para guiar intervenciones de salud pública costo efectivas y basadas en la evidencia.

Tanto los factores sociales como los ambientales están altamente relacionados con la salud. Los determinantes sociales de la salud (DSS) son factores no clínicos que afectan la salud y el estatus económico de los individuos y las comunidades, incluyendo el lugar de nacimiento, las condiciones de vida, las condiciones de trabajo y las características demográficas [3]. También están incluidos estresores sociales como el crimen, la violencia y los desordenes físicos, entre otros [4].

Los factores ambientales (por ejemplo, la polución del aire, los climas extremos, el ruido, pobre condiciones de vivienda) están muy relacionados con el estado de salud de un individuo. Las regiones urbanas densamente pobladas crean contaminación del aire, islas de calor y altos niveles de ruido, los cuales han sido implicados como causas o factores que empeoran una variedad de temas de salud. Por ejemplo, un estudio en la ciudad de Nueva York (NYC) mostró que, las admisiones en servicios de emergencias relacionadas al asma en los jóvenes entre 5 y 17 años, estaban altamente relacionadas con la exposición a ozono en el ambiente. La Encuesta Anual de Salud Comunitaria de NYC también revela que los problemas crónicos de salud autoreportados se relacionaban con temperaturas extremas, sugiriendo que la temperatura puede generar o exacerbar los síntomas de una enfermedad crónica en un individuo. Factores sociales como edad y niveles de pobreza también impactan en la salud. Un estudio en la ciudad de

Nueva York muestra que las hospitalizaciones por asma atribuible a finas partículas ($PM_{2.5}$, un subrogante de contaminación) son 4.5 veces más frecuentes en barrios con mayores niveles de pobreza [5].

Mientras las condiciones ambientales de los espacios abiertos ameritan la atención de la salud pública, el promedio de los estadounidenses pasa solo una hora de cada día en espacios abiertos; la mayoría de los individuos vive, trabaja y descansa en un ambiente cerrado, donde surgen otras preocupaciones. La pobre calidad de los espacios cerrados puede causar enfermedades relacionadas con la vivienda "el síndrome del edificio enfermo" (SBS, del inglés Sick Building Syndrome) – en el cual los ocupantes experimentan problemas de salud agudos y malestares a pesar de que no puede diagnosticarse ninguna enfermedad identificable [6]. Nuevamente, en la ciudad de Nueva York, varias agencias combinaron datos de vivienda en un esfuerzo para identificar problemas de contaminación en espacios cerrados. Usando análisis predictivo, la ciudad pudo aumentar el índice de detección de edificios considerados peligrosos, así como mejorar la oportunidad de ubicar departamentos con problemas de seguridad o peligros para la salud [7].

6.3 Aumento en la disponibilidad de datos

Por muchos años los científicos e investigadores tuvieron que lidiar con una cantidad de datos disponibles muy limitada para estudiar los factores conductuales, sociales y ambientales que existen en las ciudades, además de la dificultad de evaluar sus modelos con una muestra grande de datos urbanos [8]. La revolución en "big data" está aportando enormes volúmenes de datos y transformaciones paradigmáticas en muchas industrias dentro de los servicios y operaciones urbanas. Esto es particularmente cierto en el comercio, la seguridad y el cuidado de la salud, a medida que se recolectan, almacenan y analizan más datos en forma sistemática. El surgimiento de la informática urbana también coincide con una transición desde los sistemas de datos tradicionalmente cerrados y fragmentados a redes completamente conectadas y abiertas que incluyen comunicaciones en masa, participación de los ciudadanos (redes sociales) y flujo de información [9].

En el año 2008, 3.3 mil millones de los habitantes del mundo vivían en ciudades, representando por primera vez en la historia la mayoría de la población humana [10]. En el año 2014, el 54% de la población vivía en áreas

urbanas y está estimado que aumente hasta el 66% para el año 2050 [11]. Con el crecimiento de las ciudades, hay una creciente preocupación en el ámbito de la salud pública en relación al impacto de factores asociados como las poblaciones envejecidas, la alta densidad de la población, los servicios sanitarios inadecuados, la degradación del medioambiente, los factores de cambio climático, el aumento de la frecuencia de desastres naturales, además de la actual e inminente escasez de recursos. En forma concomitente se requiere una gran cantidad de información para planear y proveer a la salud pública de estas entidades urbanas, así como para prevenir y reaccionar a eventos adversos públicos de todos los tipos (por ejemplo: desastres epidemiológicos, naturales, criminales, y político-terroristas).

La naturaleza de la ciudad como una aglomeración de habitantes, objetos físicos y actividades, la hace una rica fuente de datos urbanos. Hoy, miles de millones de individuos están generando datos digitales a través de sus celulares y el uso de internet incluyendo redes sociales. Los Hardware como los sistemas de posicionamiento global (GPS, del inglés Global Positioning Systems) y otros sensores, también están transformándose en universales a medida que se vuelven más accesibles, resultando en diversos tipos de datos recolectados en nuevos y únicos formatos [12]. Esto es especialmente cierto en las ciudades debido a sus poblaciones masivas, creando puntos calientes de generación de datos y centros de flujo de información. Semejante disponibilidad de datos también puede proveer el sustrato para modelos estadísticamente más robustos a través de múltiples disciplinas.

Es esencial una mirada del volumen, variedad y formato de los datos urbanos abiertos para una mayor integración con las historias clínicas electrónicas. A medida que más ciudades empiezan a construir su infraestructura de la información, el volumen de datos de la ciudad aumenta rápidamente. La mayoría de los datos urbanos están en formato tabular con información basada en localización [8]. Las fuentes de datos y los procesos de recolección varían, basados en la naturaleza de los datos urbanos. Los sensores pasivos continuamente recolectan información ambiental como la temperatura, la calidad del aire, la radiación solar y el ruido, y construyen una infraestructura urbana de mediciones junto con una informatización universal [13]. También hay una gran cantidad de datos de las ciudades que son generados por los ciudadanos como pedidos de servicios y quejas.

Algunos datos preexistentes, como aquellos en un formato tabular adecuado, están inmediatamente listos para la integración, mientras que otros datos contenidos en tipos de archivos más complejos, como documentos en formato portable (PDF, del inglés Portable Document Format) u otros, son más difíciles de analizar. Este problema puede ser más complejo si los datos están codificados en letras de idiomas poco comunes.

El hecho de que muchos datos no clínicos, especialmente datos urbanos, sean geo-localizados permite a los médicos clínicos considerar la salud del paciente dentro de una mirada más amplia. Muchos factores ambientales, sociales y conductuales se conectan espacialmente, y esa correlación espacial es una medición clave en epidemiología, dado que permite facilitar la integración de los datos basada en la locación. Las conexiones y soluciones se vuelven más visibles uniendo datos no clínicos con HCEs en un nivel de planificación de la salud pública y la ciudad. Recientemente IBM anuncio, que asociando la informática cognitiva de la supercomputadora Watson con datos de la CVS Health (una cadena farmacéutica con locaciones a través de Estados Unidos), tendremos mejores predicciones referentes a la prevalencia de enfermedades crónicas como enfermedad cardiaca y diabetes en diferentes ciudades y localizaciones [14].

6.4 Integración, aplicación y calibración

En un resumen de todas las ciudades en Estados Unidos que publicaron set de datos abiertos desde el año 2013, se encontró que más del 75% de los set de datos estaban preparados en formato tabular [8]. Los datos tabulados son más amigables para la integración automatizada, dado que ya se encuentran en el formato final antes de ser integrados en bases de datos más relacionales (siempre y cuando el set de datos contenga un atributo significativo o una variable con la cual se pueda relacionar otras entradas de datos). Además, la integración de datos ocurre más fácilmente cuando el set de datos es “prolijo”, o sigue la regla de “una observación por fila y una variable por columna”. Cualquier proceso de manipulación de datos que resulte en un set de datos agregado o resumido puede quitar gran parte de la utilidad de esos datos [15].

Por ejemplo, una tabla que es familiar dentro de un ambiente de trabajo puede no ser fácilmente descifrable para otro individuo y puede ser casi imposible de analizar para una computadora sin un contexto adecuado dado

por el contenido de la tabla. Un ejemplo puede ser una tabla de presión arterial a lo largo del tiempo en diferentes lugares para un número de pacientes que puede verse como (Tabla 6.1) Aquí vemos dos pacientes, paciente 1 y paciente 2, presentándose en dos lugares, Random y Randomly, RA, en dos fechas diferentes. Mientras esta tabla puede ser leída por alguien familiarizado con el formato, de forma que un individuo podría entender que el paciente 1, el 1ro de enero de 2015 se presentó en un institución sanitaria en Random, RA con una presión sistólica de 130 mmHg y una presión diastólica de 75 mmHg, puede ser difícil manipular estos datos en un formato adecuado sin entender el contexto de la tabla.

Tabla 6.1 Ejemplo de una tabla que requiere una lectura adecuada al contexto

Registro de Presión Arterial del paciente	Random, RA		Randomly, RA	
	1 enero 2015	7 enero 2015	1 enero 2015	7 enero 2015
Paciente 1	130/75	139/83	141/77	146/82
Paciente 2	158/95	151/91	150/81	141/84

Si esta tabla debiera ser manipulada en una manera que fuera fácilmente analizada por una máquina (o por otros individuos sin requerir una explicación del contexto), debería seguir la regla de una columna por variable y una fila por observación como se observa debajo (Tabla 6.2).

Existen más limitaciones, impartidas debido a la resolución de los datos, que se refiere al nivel de detalle de los datos en el espacio, tiempo o tema, especialmente la dimensión espacial de los datos [16]. Los ejemplos incluyen: los formatos de tiempo MM/DD/YY comparado a YYYY; o código postal comparado con coordenadas geográficas. Incluso con estas limitaciones uno puede obtener información relevante de estos datos de espacio y tiempo.

Un método para proveer orientación espacial a un encuentro clínico fue recientemente adoptado por los administradores de la base de datos MIMIC (*Medical Information Mart for Intensive Care),* que actualmente contiene datos de más de 37000 admisiones en unidades de cuidados intensivos [17]. Los investigadores utilizan el sistema de código postal de Estados Unidos

para aproximar el área de residencia de los pacientes. Este método reporta los primeros tres dígitos del código postal del paciente, mientras omite los últimos dos [18]. Los primeros tres dígitos del código postal contienen dos piezas de información: el primer número del código se refiere al número de estado, y los otros dos se refieren al establecimiento del Centro Seccional del Servicio Postal de Estados Unidos, a través del cual se procesan los correos para los distintos condados de ese estado [19]. Los primeros tres dígitos del código postal son suficientes para encontrar todos los otros códigos postales atendidos por el establecimiento del Centro Seccional y los datos del nivel de población están disponibles por el código postal y por el censo del gobierno norteamericano [20]

Tabla 6.2 Un set de datos ordenados que contienen un formato sencillo de la tabla 6.1 legible por una máquina

ID del Paciente	Lugar	Fecha (MM/DD/AAAA)	Presión (mmHg)	Ciclo
1	Random, RA	1/1/2015	130	Sístole
1	Random, RA	1/1/2015	75	Diástole
1	Random, RA	1/1/2015	139	Sístole
1	Random, RA	1/1/2015	83	Diástole
1	Randomly, RA	1/1/2015	141	Sístole
1	Randomly, RA	1/1/2015	77	Diástole
1	Randomly, RA	1/1/2015	146	Sístole
1	Randomly, RA	1/1/2015	82	Diástole
2	Random, RA	1/1/2015	158	Sístole
2	Random, RA	1/1/2015	95	Diástole
2	Random, RA	1/1/2015	151	Sístole
2	Random, RA	1/1/2015	91	Diástole
2	Randomly, RA	1/1/2015	150	Sístole

2	Randomly, RA	1/1/2015	81	Diástole
2	Randomly, RA	1/1/2015	141	Sístole
2	Randomly, RA	1/1/2015	84	Diástole

Las conexiones y soluciones se vuelven más visibles uniendo datos no clínicos con HCE en un nivel de planificación de la salud pública y de la ciudad. Aunque muchos estudios previos muestran la correlación entre la contaminación del aire y el asma, solo recientemente se hizo posible la detección de $PM_{2.5}$, SO2, y níquel (Ni) en el aire que regresa a los generadores en edificios con sistemas de calefacción antiguos, en gran parte debido a la mayor recolección de datos y la integración entre múltiples agencias y disciplinas [21]. Mientras que los estudios revelan conexiones adicionales entre nuestro ambiente y los procesos patológicos, nuestra habilidad para identificar potenciales amenazas a la salud va a estar limitada por nuestra habilidad para medir estos factores ambientales con suficiente resolución para que puedan ser aplicables a nivel del paciente, creando medicina verdaderamente personalizada.

Por ejemplo, dos variables, comúnmente capturadas en muchas observaciones son la geo-espacialidad y la temporalidad. Como todas las acciones comparten estas condiciones, se hace posible la integración entre una variedad de datos que, de otra forma, estarían utilizados vagamente en una consulta médica. Cuando se involucran en una consulta, un medico puede determinar, a partir de los datos recolectados durante el examen clínico y confección de la historia, la locación precisa de ese paciente en un periodo particular de tiempo dentro de alguna resolución espacial. Como un caso de ejemplo, un paciente puede presentarse con un proceso inflamatorio de tracto respiratorio. El individuo puede vivir en Random, RA, y trabajar como un administrador en Randomly, RA. Uno puede graficar estas variables a lo largo del tiempo y separarlas para representar ambos, el ambiente laboral y el domicilio del individuo– así como también otro viaje (Fig. 6.1)

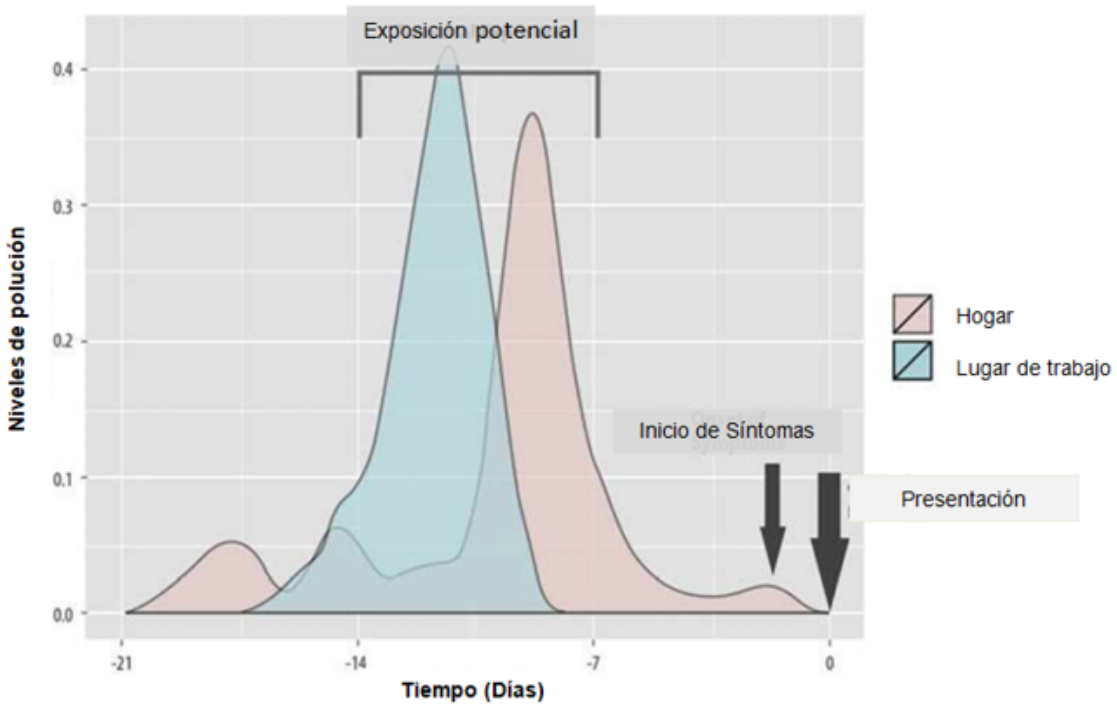

Fig. 6.1 Ejemplo de niveles de contaminación a lo largo del tiempo para el ambiente de "trabajo" y de "hogar" del paciente con etiquetas aproximadas que pueden proveer soporte de decisiones clínicas relevantes.

Este mismo método puede ser aplicado a otras variables que pudiera determinarse tienen correlatos estadísticos significativos durante un periodo de tiempo anterior al comienzo de los síntomas y luego la consulta médica.

Con el aumento de la disponibilidad de la tecnología de información, hay menos necesidad de redes de información centralizada y está abierta la oportunidad para que el individuo participe en la recolección de datos, creando redes de sensores virtuales de mediciones ambientales y de las enfermedades. Las redes móviles y sociales han creado oportunidades poderosas para la informática urbana y la planificación de desastres, particularmente en la vigilancia en salud pública y la respuesta a las crisis [13]. Hay aplicaciones móviles de geo-localización abiertas como, por ejemplo, Health Map's Outbreaks Near Me [22] y Sickweather [23] que recolectan datos en tiempo real en redes sociales.

En el año 2014 durante el brote de la enfermedad del virus del Ébola, el auto-reporte y el reporte de contacto cercano fue esencial para crear mapas precisos del brote de la enfermedad [24]. La aparición de los dispositivos móviles está empujando tanto a fabricantes de HCE a desarrollar

infraestructura que integra datos desde dispositivos móviles como a compañías tercerizadas a proveer almacenamiento en la nube e integración de datos desde diferentes dispositivos para un mayor poder analítico.

La atención y las inversiones en salud digital y ciudades digitales continúan creciendo rápidamente. En la prestación de servicios de salud digital, el financiamiento de los inversores creció desde 1.1 mil millones de dólares en el año 2011 a 5 mil millones en el año 2014, y el análisis de "big data" es el subsector más activo de los emprendimientos, tanto por la cantidad de inversiones como por el número de transacciones [25]. La integración será un largo proceso requiriendo capacidades digitales, nuevas políticas, colaboración entre los sectores público y privado, innovaciones entre líderes de la industria y centros de investigación [26]. Aun así, creemos que con más colaboración interdisciplinaria en minería y análisis de datos, ganaremos nuevos conocimientos en los factores no clínicos asociados a la salud y en indicadores de resultados de las enfermedades [27]. Además, esta integración crea un círculo de retroalimentación, empujando a las ciudades a recolectar mejor información y en mayor cantidad. Integrar información no clínica en registros de salud sigue siendo un desafío. Idealmente la información obtenida del paciente fluiría hacia una muestra urbana más grande y viceversa. Los desafíos continúan en proteger la confidencialidad a nivel del paciente y determinar la aplicabilidad de datos macroscópicos al paciente individual.

6.5 Un empoderamiento bien conectado.

Los procesos de enfermedad pueden ser provenir y ser modificados por interacciones entre el paciente y su ambiente. Entender este ambiente es importante para los médicos, los hospitales, para los decisores en políticas de salud pública y para los pacientes mismos. Con esta información podemos prevenir a los pacientes del riesgo de enfermedad (prevención primaria), actuar más tempranamente para minimizar morbilidad de enfermedades (prevención secundaria) y optimizar las intervenciones terapéuticas.

Un buen ejemplo del uso de datos no clínicos para la prevención de enfermedades es el uso de sistemas de información geográfica (SIG) para tamizaje preventivo de población en riesgo para enfermedades de transmisión sexual (ETS). Los Sistemas de información geográfica son usados para la vigilancia de ETS en alrededor del 50% de los programas de vigilancia

de ETS en Estados Unidos [28]. En Baltimore (Maryland, Estados Unidos) un estudio basado en SIG identificó el núcleo de grupos con infección repetida por gonococo (una ETS) que mostró agrupación geográfica [29]. Los autores sugirieron la posibilidad de mayor protección al dirigir la prevención a poblaciones restringidas geográficamente.

Un próximo escalón lógico es la interacción entre las autoridades de salud pública y las HCE. A medida que la información geográfica de-identificada de salud se torna disponible públicamente, una historia clínica electrónica podría descargar esta información desde la nube, aplicarla al código postal, sexo, edad y preferencia sexual (si está documentada) del paciente y advertir/señalar al médico que decidiría si es necesaria una intervención basado en un riesgo calculado de adquirir una ETS

6.6 Conclusión

Una buena gestión de datos será esencial para proteger la confidencialidad de la información de salud de la divulgación ilegal o accidental. Para los pacientes, la idea de aumentar el empoderamiento en su salud es esencial [8]. El aumento de la aplicación de sensores y de la visualización de datos hace nuestro comportamiento y nuestros alrededores más visibles y tangibles, y nos alerta sobre potenciales riesgos ambientales. Lo que es más importante, nos ayudará a entender mejor y ganar poder sobre nuestras propias vidas.

La dicotomía de abordar la salud poblacional versus la salud individual debe ser identificada. Los investigadores deberían preguntarse: ¿Qué información es relevante para el objetivo que estoy abordando, y cuáles datos debemos llevar desde los registros del paciente hacia el dominio de la salud pública? El corolario a esta pregunta es: ¿cómo podemos balancear el derecho de la privacidad con el beneficio de los datos no clínicos aplicables al individuo y a grandes poblaciones? Finalmente: ¿cómo podemos crear sistemas que seleccionen datos relevantes de un solo paciente y lo presenten al médico en un contexto de salud poblacional? En este capítulo, hemos intentado brindar un resumen del uso potencial de datos tradicionalmente no clínicos en historias clínicas electrónicas, sumado al mapeo de algunos obstáculos y estrategias para usar dicha información, además de resaltar ejemplos prácticos en el uso de estos datos en un entorno clínico.

Nota: Las imágenes de este capítulo se encuentran disponibles a color en el ebook; disponible en www.hardineros.com

Referencias

1. Barton H, Grant M (2013) Urban planning for health cities, a review of the progress of the european healthy cities program. J Urban Health Bull NY Acad Med 90:129-141.
2. Badland HM, Schofield GM, Witten K, Schluter PJ, Mavoa S, Kearns RA, Hinckson EA, Oliver M, Kaiwai H, Jensen VG, Ergler C, McGrath L, McPhee J (2009) Understanding the relationship between activity and neighborhoods (URBAN) study: research design and methodology. BMC Pub Health 9:244.
3. Osypuk TL, Joshi P, Geronimo K, Acevedo-Garcia D (2014) Do social and economic policies influence health? Rev Curr Epidemiol Rep 1:149-164.
4. Shmool JLC, Kubzansky LD, Newman OD, Spengler J, Shepard P, Clougherty JE (2014) Social stressors and air pollution across New York City communities: a spatial approach for assessing correlations among multiple exposures. Environ Health 13:91.
5. Kheirbek I, Wheeler K, Walters S, Kass D, Matte T (2013) PM2.5 and ozone health impactsand disparities in New York City: sensitivity to spatial and temporal resolution. Air Qual Atmos Health 6:473-486.
6. Indoor Air Facts No. 4 sick building syndrome. United States Environmental Protection Agency, Research and Development (MD-56) (1991).
7. Goldstein B, Dyson L (2013) Beyond transparency: open data and the future of civic innovation. Code for America Press, San Francisco.
8. Barbosa L, Pham K, Silva C, Vieira MR, Freire J (2014) Structured open urban data: understanding the landscape. Big Data 2:144-154.

9. Shane DG (2011) Urban design since 1945-a global perspective. Wiley, New York, p 284.
10. National Intelligence Council (2012) Global trends 2030: alternative worlds. National Intelligence Council.
11. World Urbanization Prospects, United Nations (2014).
12. Goldsmith S, Crawford S (2014) The responsive city: engaging communities through data-smart governance. Wiley, New York.
13. Boulos M, Resch B, Crowley D, Breslin J, Sohn G, Burtner R, Pike W, Jezierski E, Chuang K (2011) Crowdsourcing, citizen sensing and sensor web technologies for public andenvironmental health surveillance and crisis management: trends, OGC standards and application examples. Int J Health Geographic 10:67.
14. McMullan T. Dr Watson: IBM plans to use Big Data to manage diabetes and obesity. URL: http://www.alphr.com/life-culture/1001303/dr-watson-ibm-plans-to-use-big-data-to-manage-diabetes-and-obesity
15. Wickham H (2014) Tidy data. J Stat Softw 10:59.
16. Haining R (2004) Spatial data analysis-theory and practice. Cambridge University Press, Cambridge, p 67.
17. MIMIC II Databases. Available from: http://physionet.org/mimic2. Accessed 02 Aug 2015.
18. Massachusetts Institute of Technology, Laboratory of Computational Physiology. mimic2 v3.0 D-PATIENTS table. URL: https://github.com/mimic2/v3.0/blob/ad9c045a5a778c6eb283bdad310594484cca873c/-posts/2015-04-22-dpatients.md.Accessed 02 Aug 2015 (Archived by WebCite® at http://www.webcitation.org/6aUNzhW6g).
19. http://pe. usps.com/businessmail101/glossary.htm.
20. http://factfinder. census.gov/.
21. Jeffery N, McKelvey W, Matte T (2015) using tracking infrastructure to support public health programs, policies, and emergency response in New York City. Pub Health ManagPract 21 (2 Supp): S102-S106.
22. http://www.healthmap.org/outbreaksnearme/.
23. http://www.sickweather.com>.
24. Kouadio KI, Clement P, Bolongei J, Tamba A, Gasasira AN, Warsame A, Okeibunor JC, Ota MO, Tamba B, Gumede N, Shaba K, Poy A, Salla M, MihigoR, Nshimirimana D (2015) Epidemiological and surveillance response to Ebola virus disease outbreak in Lofa County, Liberia (Mar-Sept 2014); lessons learned, edn 1. PLOS Currents Outbreaks. 6 May 2015. doi: 10.1371/currents.outbreaks.9681514e450dc8d19d47e1724d2553a5.
25. The re-imagination of healthcare. StartUp Health Insights. www.startuphealth.com/insights.
26. Ericsson Networked Society City Index (2014).

27. Corti B, Badland H, Mavoa S, Turrell G, Bull F, Boruff B, Pettit C, Bauman A, Hooper P, Villanueva K, Burt T, Feng X, Learnihan V, Davey R, Grenfell R, Thackway S (2014) Reconnecting urban planning with health: a protocol for the development and validation of national livability indicators associated with non-communicable disease risk behaviors and health outcomes. Pub Health Res Pract 25 (1): e2511405.
28. Bissette JM, Stover JA, Newman LM, Delcher PC, Bernstein KT, Matthews L (2009) Assessment of geographic information systems and data confidentiality guidelines in STD programs. Pub Health Rep 124 (Suppl 2): 58-64.
29. Bernstein TK, Curriero FC, Jennings JM et al (2004) Defining core gonorrhea transmissión utilizing spatial data. Am J Epidemiol 160:51-58.

Capítulo 7

UTILIZANDO LA HISTORIA CLÍNICA ELECTRÓNICA PARA CONDUCIR INVESTIGACIONES DE RESULTADOS Y SERVICIOS DE SALUD

Laura Myers y Jennifer Stevens

Puntos clave

- Las Historias Clínicas Electrónicas (HCE) se han vuelto herramientas esenciales en la investigación clínica, tanto como complemento de los métodos existentes como en los crecientes dominios de investigación y análisis de resultados.
- Mientras que los datos de las HCE son cuantiosos y los métodos de análisis son poderosos, es esencial comprender cabalmente los sesgos y limitaciones introducidas cuando se usan en la investigación de los servicios de salud.

7.1 Introducción

Los datos de las HCE pueden ser una herramienta poderosa para investigación. Sin embargo, los investigadores deben estar atentos a la fiabilidad de los datos recolectados con fines clínicos y a los sesgos inherentes a la utilización de esos datos para conducir buenas investigaciones de resultados y de servicios de salud. En la actualidad, se están desarrollando métodos innovadores para mejorar la calidad de los datos y, con ello, nuestra capacidad para extraer conclusiones de estudios que utilicen datos de las HCE.

Los Estados Unidos de América dedican una gran cantidad de su producto bruto interno (17,6% en 2009) a la asistencia sanitaria [1]. Con esta gran inversión financiera y social, surgen preguntas fundamentales al momento de evaluar esta inversión:

- ¿Cómo conocemos qué tratamientos funcionan y para qué pacientes?
- ¿Cuánto debería costar la atención sanitaria? ¿Cuándo es demasiado para pagar? ¿En qué tipo de cuidados deberíamos invertir más o menos recursos?

- ¿Cómo funciona el sistema de salud y cómo podría funcionar mejor?

La investigación en servicios de salud es un campo que linda en la intersección entre las políticas de salud, la gestión y la asistencia y que busca responder estas preguntas. Fundamentalmente, la investigación en servicios de salud ubica al sistema de salud en el microscopio como organismo de estudio.

Para empezar a responder estas preguntas, los investigadores requieren de grandes volúmenes de datos que involucren múltiples pacientes a lo largo de diferentes estructuras de servicios de salud y a través del tiempo. El crecimiento de este campo de investigación en los últimos 15 años ha coincidido en forma simultánea con el desarrollo de las HCE y con un número creciente de proveedores que las utilizan en sus lugares de trabajo [2]. Las HCE proveen grandes cantidades de datos crudos para alimentar estas investigaciones, tanto a nivel granular del paciente y el proveedor, como en el nivel agregado del hospital, estado o nación.

Conducir investigaciones con datos de HCE conlleva múltiples desafíos. Estos datos suelen estar plagados de sesgos, dado que son recolectados con propósitos distintos a la investigación, son ingresados por diferentes usuarios para un mismo paciente y resultan difíciles de integrar a través de los sistemas de salud (ver el capítulo **"Factores confundidores por Indicación"**). En este capítulo se hará énfasis en los intentos por aprovechar la promesa que suponen las HCE para la investigación de los servicios de salud, poniendo especial consideración en los desafíos que deben afrontar los investigadores para poder obtener conclusiones significativas y válidas.

7.2 El surgimiento de las HCEs en la Investigación de los Servicios de Salud

7.2.1 Las HCEs en estudios observacionales y de resultados

Los estudios observacionales, sean retrospectivos o prospectivos, intentan realizar inferencias sobre los efectos de diferentes exposiciones. Dentro de la investigación de servicios de salud estas exposiciones incluyen tanto distintas exposiciones clínicas (por ejemplo: ¿el tratamiento de reemplazo hormonal ayuda o perjudica a los pacientes?) como exposiciones a los distintos tipos de servicios de salud (por ejemplo, ¿el ingreso para una

revascularización cardíaca en un hospital con mayor numero de admisiones anuales mejora la supervivencia del infarto de miocardio por sobre el ingreso en un hospital más pequeño?). La disponibilidad de una gran cantidad de datos en las historia clínicas electrónicas ha alimentado este tipo de investigación, a medida que la extracción y transcripción de datos de los registros en papel ha dejado de ser una barrera. Estos estudios sacan ventaja de los elementos demográficos y clínicos que son registrados de forma rutinaria como parte del encuentro con el sistema de salud (por ejemplo, edad, sexo, etnia, procedimientos realizados, duración de la estancia, recursos de cuidados críticos utilizados).

A continuación, hemos destacado ejemplos de estos tipos de investigación. Cada uno es un ejemplo de un estudio que ha utilizado datos de las HCEs, ya sea a nivel nacional como a nivel hospitalario, para realizar inferencias acerca del cuidado de la salud y cómo éste es provisto.

- **¿La atención de la salud es variable?** Los investigadores que compilaron y examinaron el Dartmouth Atlas, han demostrado una variación geográfica sustancial del cuidado. En su artículo original en *Science*, Wennberg y Gittlesohn identificaron una gran variación en el uso de servicios de salud en Vermont [3]. Estos autores utilizaron datos derivados de distintas clases de servicios médicos –servicios domiciliarios, altas de pacientes, etc– para realizar estas inferencias. Posteriormente otras investigaciones relativas a la variabilidad nacional de la asistencia médica, pudieron capitalizar la disponibilidad de este tipo de datos en formato electrónico.
- **¿Los hospitales con mayor experiencia en un área particular presentan mejor desempeño?** Birkmeyer y sus colaboradores estudiaron la relación entre el volumen de admisiones y los resultados quirúrgicos de un hospital, observando diferencias absolutas en la tasa de mortalidad ajustada entre hospitales de bajo volumen y alto volumen de ingresos, que van desde el 12% para la resección pancreática al 0,2% para la endarterectomía carotídea [5]. Kahn et al. también usaron datos disponibles de más de 20.000 pacientes para demostrar que la mortalidad asociada a la ventilación mecánica era 37% menor en hospitales de alto volumen de ingresos en comparación con hospitales de bajo volumen de ingresos [6]. Ambos grupos de investigación hicieron uso de grandes cantidades de datos clínicos y administrativos –

de Medicare en el caso de Birkmeyer y de la base de datos APACHE de Cerner en el de Kahn y col– para formular preguntas importantes sobre dónde los pacientes deben buscar diferentes tipos de cuidado.

- **¿Cómo podemos identificar daño a los pacientes a pesar del cuidado habitual?** Herzig y sus colaboradores utilizaron la HCE granular de una única institución y encontraron que los medicamentos ampliamente prescriptos para suprimir la producción de ácido estaban asociados con un riesgo incrementado de contraer neumonía [7]. Otros autores que condujeron investigaciones similares encontraron que este tipo de medicamentos suele continuarse luego del alta hospitalaria [8, 9].

Para facilitar el diseño apropiado y la identificación de confundidores en estudios observacionales, los investigadores han tenido que idear métodos para extraer marcadores diagnósticos, de severidad de enfermedad y de comorbilidades de los pacientes usando únicamente la huella digital. Post y col. [10] desarrollaron un algoritmo de búsqueda de pacientes con hipertensión refractaria a diuréticos, consultando por pacientes que tenían un diagnóstico de hipertensión pese a haber recibido un tratamiento con diuréticos durante 6 meses. Los métodos previamente validados para medir la severidad de la enfermedad de un paciente, como los puntajes APACHE o SAPS [11, 12], tienen elementos que no son fácilmente extraíbles sin realizar el ingreso manual de los datos. Para hacer frente a estos desafíos, investigadores como Escobar y Elixhayser han propuesto métodos alternativos derivados electrónicamente para medir severidad de enfermedad [13, 14] y para identificar comorbilidades [14]. El trabajo de Escobar, con una medida de severidad de enfermedad con un área bajo la curva de 0.88 hace uso de datos electrónicos de alta granularidad, incluyendo valores de laboratorio. La medida de comorbilidad de Elixhayser está disponible públicamente a través de la Agency for Healthcare Research and Quality con el solo requisito de información de facturación [15].

Finalmente, los investigadores deben desarrollar y emplear modelos matemáticos adecuados que corrijan las deficiencias de los datos electrónicos de salud, de lo contrario se arriesgan a que sus conclusiones sean inexactas. Los ejemplos de dichas técnicas de modelado estadístico son numerosos e incluyen el pareamiento por puntaje de propensión, métodos causales como modelos estructurales marginales y la ponderación

de probabilidad inversa y diseños de otros campos como análisis de variables instrumentales [16-19]. Los detalles de estos métodos son discutidos en otras secciones de este texto.

7.2.2 La Historia Clínica Electrónica como herramienta para facilitar el reclutamiento de pacientes en ensayos prospectivos

A pesar del poder de las HCE para conducir investigaciones de servicios de salud y de resultados en forma retrospectiva, el *gold standard* en investigación siguen siendo los ensayos prospectivos y aleatorizados. La HCE ha funcionado como una herramienta valiosa para tamizar a gran escala pacientes elegibles. En esta instancia, los investigadores utilizan los datos disponibles a través de los registros electrónicos como una técnica de tamizaje de grandes volúmenes para luego dirigir los esfuerzos del reclutamiento a los pacientes más adecuados. Los ensayos clínicos que desarrollan estrategias electrónicas de identificación y reclutamiento de pacientes se encuentran en un nivel superior de ventaja; sin embargo a pesar de ser robustos se han descrito como sensibles pero no específicos y suelen requerir la revisión manual de los registros individuales sumada a los esfuerzos del tamizaje [20]. Embi y col. [21] han propuesto usar las HCEs para generar simultáneamente Alertas de Ensayos Clínicos, en particular en HCEs comerciales como Epic, para posicionar a las historias como una estrategia en el lugar de cuidado. Esta estrategia podría acelerar el reclutamiento, aunque debe considerarse el riesgo de pérdida de confidencialidad del paciente y la continua tensión entre la atención del paciente y el reclutamiento en ensayos clínicos [22].

7.2.3 La HCE como herramienta para estudiar y mejorar los resultados de los pacientes

La calidad de la atención sanitaria también puede ser seguida e informada a través de HCEs, ya sea para mejoras internas o para evaluación comparativa nacional; el sistema de salud de Veterans' Affairs' (VA) resalta esta característica. Byrne y col. [23] reportaron que en los años 90, el VA gastó más dinero en infraestructura de tecnología de la información y logró mayores tasas de adopción en comparación con el sector privado. Su HCE de desarrollo propio, llamada VistA, brindaba una forma de rastrear los

procesos de prácticas preventivas como el tamizaje de diabetes o cáncer a través de mensajes en las ventanas emergentes. Entre el año 2004 y el año 2007, encontraron que el sistema de VA logró mejores controles de glucosa y lípidos para personas diabéticas en comparación con Medicare HMO benchmark [23]. Si bien la implementación inicial de VistA requirió una gran inversión de capital, se estima que dicha infraestructura permitió al sistema de VA ahorrar 3.090 millones de dólares en el largo plazo. Al mismo tiempo, continúa siendo una fuente de mejora de calidad a medida que las métricas evolucionan a lo largo del tiempo [23].

7.3 Cómo evitar errores comunes al utilizar la Historia Clínica Electrónica para realizar investigaciones en servicios de salud

Podríamos proponer el siguiente ejemplo hipotético de investigación como caso de estudio para destacar los principales desafíos que implica conducir investigaciones en servicios de salud con datos electrónicos:

–**Investigación propuesta:** las medicaciones antipsicóticas (por ejemplo, haloperidol) se prescriben frecuentemente en las unidades de cuidados intensivos para tratar pacientes con síntomas de delirio. Sin embargo, estas medicaciones se han asociado con su propio potencial riesgo de daño [24] en forma independiente del riesgo global de daño por el delirio. Los investigadores están interesados en saber si el tratamiento con antipsicóticos incrementa el riesgo de muerte intrahospitalaria y si eleva los costos de atención y el uso de recursos hospitalarios.

7.3.1 Paso 1: Reconocer la falibilidad de las Historias Clínicas Electrónicas

Las HCEs rara vez están completas o son totalmente correctas. Hogan y col. [25] intentaron estimar cuan completos y precisos eran los datos utilizados en estudios conducidos con una HCE y encontraron una variabilidad significativa en ambos atributos. La completitud osciló entre 31 y 100%, mientras que la precisión lo hizo entre 67 y 100% [25]. La tabla 7.1 resalta ejemplos de distintos diagnósticos y posibles fuentes de datos, que pueden o no estar disponibles para todos los pacientes.

–**Investigación propuesta:** los investigadores deberán determinar qué pacientes estuvieron expuestos a antipsicóticos y cuáles no. Se enfrentan con el problema de que es poco probable que exista un único lugar donde

esta información se encuentre almacenada. ¿Deberían utilizar los datos de administración de la farmacia? ¿O los de enfermería? ¿Deberían ver en qué pacientes se facturó esta medicación? ¿Y qué ocurre si necesitan esos datos de múltiples hospitales con distintas historias clínicas electrónicas?

Aún con una estrategia sólida de extracción de datos, la fidelidad de los diferentes tipos de datos es variable [26-33]. Por ejemplo, muchos sistemas de HCE tienen la opción de ingresar texto libre para describir una condición médica que puede contener errores de ortografía o estar escrito de manera no convencional. Otro ejemplo puede ser que el reembolso de algún código de facturación en particular puede influenciar la incidencia de ese código en la historia clínica electrónica, de manera que la facturación puede no reflejar la verdadera incidencia y prevalencia de una enfermedad [34,35].

7.3.2 Paso 2: Entender los factores de confusión, los sesgos y los datos faltantes utilizando la HCE para investigación

Revisaremos los siguientes aspectos metodológicos inherentes a la conducción de investigaciones con historias clínicas electrónicas: sesgo de selección, factores de confusión y datos faltantes. Estos son explorados con mayor profundidad en otros capítulos del texto.

Tabla 7.1 Ejemplos del espectro de datos que pueden utilizarse para identificar pacientes con enfermedad cardíaca isquémica o lesión pulmonar aguda en las historias clínicas electrónicas

Enfermedad	Fuente de datos	Ejemplo
Enfermedad cardiaca isquémica	Datos de facturación	Código CIE-9:410 [48]
	Datos de laboratorio	Troponina positiva al ingreso
	Documentación clínica	En la epicrisis: “se describe que el paciente tuvo elevación del ST en el ECG y fue ingresado en el servicio de hemodinamia
Lesión pulmonar aguda	Datos de facturación	Código CIE-9:518.5 y 518.82 con los códigos de procedimientos 96.70,96.71 y 96.72 para ventilación mecánica [49]
	Datos de radiología	Informes de Rx de tórax con “Bilateral” e “Infiltrados”
	Datos de laboratorio	PaO2 <300 mmHg

El sesgo de selección, o la incapacidad de representar poblaciones generalizables en la población estudiada, puede ocurrir si todos los pacientes, incluyendo los del grupo control, ya buscan cuidados médicos dentro de un sistema basado en HCE. Por ejemplo, estudios basados en HCE que comparen abordajes médicos contra quirúrgicos para una misma condición pueden no estar comparando poblaciones equivalentes en cada grupo; los pacientes que buscan una solución quirúrgica pueden diferir fundamentalmente de aquellos que buscan un tratamiento más conservador. Hripcsak y col. [36] utilizaron grandes volúmenes de datos clínicos de un centro universitario en 2007 para comparar la mortalidad por neumonía contra datos recolectados manualmente que habían sido publicados previamente; el diferente criterio de búsqueda alteró la población de pacientes y por lo tanto el riesgo de muerte. Si bien no se elimina completamente, el sesgo de selección se reduce cuando se utiliza la aleatorización prospectiva [37].

El sesgo de confusión representa la incapacidad de identificar adecuadamente una variable adicional que influencia tanto la variable dependiente como la independiente. En investigaciones con historias clínicas electrónicas, los factores de confusión representan un desafío singular ya que la identificación de todas las variables confusoras es casi imposible.

–**Investigación propuesta:** los investigadores de este estudio están interesados en los resultados de los pacientes expuestos a antipsicóticos durante su estadía hospitalaria. Pero es probable que los pacientes con cuadros de delirio en las unidades de cuidados intensivos se encuentren más enfermos que aquellos que no tienen cuadro clínico de delirio, y los pacientes más enfermos requieren mayores recursos hospitalarios. Como resultado, los antipsicóticos estarán en apariencia asociados con un mayor riesgo de mortalidad intrahospitalaria y uso de recursos hospitalarios, aunque no debido al efecto independiente de la droga sino como resultado de factores de confusión.

Los datos faltantes o las muestras de datos registrados de manera desigual como parte de las HCE crean sus propios desafíos complejos para la investigación de servicios de salud. Por ejemplo, restringir el análisis únicamente a pacientes que tengan un conjunto de datos completos puede producir inferencias muy diferentes (y escasamente generalizables). La

multidimensionalidad del problema suele subestimarse y subexaminarse. Casi todos los software analíticos convencionales presumen la completitud de la matriz de datos, llevando a que muchos investigadores no atiendan completamente estos problemas. Por ejemplo, los datos pueden estar desalineados debido a una falta de muestreo, a datos faltantes o simplemente porque están mal alineados. En otras palabras, los datos pueden no haberse medido en un período de tiempo de manera deliberada (por ejemplo, un paciente fue extubado y por ende no hay valores documentados para la ventilación mecánica) o pueden haberse medido pero no se ingresaron y por lo tanto no pueden ser utilizados. Rusanov y col estudiaron 10.000 pacientes ambulatorios en un centro universitario que recibieron anestesia general para procedimientos electivos. Los pacientes con mayor riesgo de resultados adversos que ingresaban al quirófano tenían más datos registrados, incluyendo valores de laboratorio, prescripciones de medicación y posibles órdenes de internación en relación con los pacientes menos enfermos [38]. Los métodos utilizados para manejar los datos faltantes incluían la omisión de los casos incompletos, la eliminación inteligente por pares, la sustitución por la media, la sustitución de regresión o técnicas de modelado para la máxima probabilidad y la imputación múltiple [39].

7.4 Direcciones futuras para las HCEs y la investigación de servicios de salud

7.4.1 Asegurando una adecuada protección de la privacidad del paciente

Hay controversia en torno a si utilizar la HCE para investigación va en contra de los estándares nacionales de privacidad. En grandes cohortes, muchos pacientes pueden presentar la misma información de salud por lo que los datos quedan lo suficientemente desidentificados. Es más, Ingelfinger y col. dan cuenta que aquellos países con registros de salud como los escandinavos presentan una ventaja para la investigación [40]. Sin embargo, la información sobre salud es un tipo de información protegida bajo Acta de Portabilidad y Responsabilidad del Seguro Médico (HIPAA) por lo que hay una gran preocupación entre los profesionales e investigadores de la salud norteamericanos respecto a su adecuado almacenamiento y difusión. Algunos argumentan que los pacientes deben consentir (en lugar

de simplemente ser notificados) que su información pueda ser utilizado para fines de investigación en el futuro. Ingelfinger y col. [40] recomiendan la aprobación de los registros por una Junta de Revisión Institucional (IRB, del inglés Institutional Review Board) así como un riguroso proceso de desidentificación.

La percepción pública acerca del uso secundario de HCE probablemente no sea tan prohibitiva como creyeron quienes diseñaron las políticas en la materia. En una encuesta realizada a 3.300 personas, estuvieron más dispuestos a que su información fuera utilizada por hospitales universitarios para realizar investigaciones en comparación con departamentos de Salud Pública o con propósito de mejoras de calidad [41]. Se mostraron mucho menos dispuestos a contribuir en investigaciones de marketing o a que su información fuera utilizada por empresas farmacéuticas [41].

Con la creciente cantidad de información que viene siendo ingresada en HCEs a lo largo del país, la *Asociación Americana de Informática Medica* convocó a un panel para realizar recomendaciones sobre cómo hacer el mejor uso de las HCE de manera segura para propósitos distintos al cuidado directo de los pacientes. En el año 2006 el panel solicitó un estándar nacional para lidiar con los problemas de privacidad. Describieron situaciones complejas donde una inadecuada desidentificación produjo violaciones a la seguridad y casos en los que los datos eran vendidos por médicos para su beneficio [42]. Si bien el panel demandó un marco nacional transparente, comprensivo y aceptado públicamente, no propuso ningún estándar en particular [42]. Otros grupos como el Instituto de Investigación de Resultados Centrados en el Paciente, (PCORI, por su nombre en inglés *Patient-Centered Outcomes Research Institute)* han atendido el mismo problema en un foro nacional en el año 2012 en el que se discutieron distintas visiones, pero no se realizaron recomendaciones explícitas. La controversia en el área continúa.

7.5 Colaboraciones multidimensionales

Continuando, el verdadero poder de los datos integrados solo puede ser aprovechado si se forman más colaboraciones, tanto dentro de las instituciones como entre ellas. La investigación a escala nacional en EE.UU. ha sido probada como plausible. La FDA implementó un programa piloto en

el año 2009 conocido como el programa Mini-Sentinel que reunió 31 organizaciones académicas y privadas para monitorear eventos de seguridad relacionados con medicamentos y dispositivos que actualmente están en el mercado [43]. Es cierto que la fusión de bases de datos puede requerir importantes recursos financieros, especialmente si las bases de datos deben ser codificadas y/o validadas, pero investigadores como Bradley y col. [44] creen que este es un buen uso costo-efectivo del dinero, debido al gran potencial que tiene en la realización de avances en los servicios de cuidado. La posibilidad de asegurar precisión en grandes volúmenes de datos y su integración a través de múltiples tecnologías y plataformas de registros de salud es fundamental para la factibilidad de la colaboración multidimensional. Los esfuerzos para asegurar la calidad y accesibilidad de los datos deben ser promovidos a la par que la privacidad del paciente.

7.6 Conclusión

Los investigadores continúan haciendo preguntas fundamentales sobre nuestro sistema de salud, sacando provecho de la avalancha de datos generada por las HCEs. Desafortunadamente, dicha avalancha es caótica y problemática. Mientras continúa el desarrollo en el campo de investigación de servicios de salud con HCEs, debemos exigir a los investigadores estándares rigurosos [45] y alentar a una mayor inversión en bases de datos clínicas que sean amigables para investigaciones y a una mayor colaboración entre instituciones. Solo entonces los descubrimientos en resultados y servicios de salud estarán cercanos [46,47]. Es momento que la asistencia sanitaria recoja los beneficios de la riqueza de datos que ya existe.

Referencias

1. Center for Medicare and Medicaid Services (2015) National health expenditure data fact sheet.
2. Jha AK, DesRoches CM, Campbell EG, Donelan K, Rao SR et al (2009) Use of electronic health records in U.S. hospitals. N Engl J Med 360:1628-1638.
3. Wennberg J, Gittelsohn (1973) Small area variations in health care delivery. Science 182:1102-1108.
4. Stevens JP, Nyweide D, Maresh S, Zaslavsky A, Shrank W et al (2015) Variation in inpatient consultation among older adults in the United States. J Gen Intern Med 30:992-999.
5. Birkmeyer JD (2000) Relation of surgical volume to outcome. Ann Surg 232:724-725.
6. Kahn JM, Goss CH, Heagerty PJ, Kramer AA, O'Brien CR et al (2006) Hospital volume and the outcomes of mechanical ventilation. N Engl J Med 355:41-50 68 7 Using EHR to Conduct Outcome and Health Services Research.
7. Herzig SJ, Howell MD, Ngo LH, Marcantonio ER (2009) Acid-suppressive medication use and the risk for hospital-acquired pneumonia. JAMA 301:2120-2128.
8. Murphy CE, Stevens AM, Ferrentino N, Crookes BA, Hebert JC et al (2008) Frequency of inappropriate continuation of acid suppressive therapy after discharge in patients who began therapy in the surgical intensive care unit. Pharmacotherapy 28:968-976.
9. Zink DA, Pohlman M, Barnes M, Cannon ME (2005) Long-term use of acid suppressión started inappropriately during hospitalization. Aliment PharmacolTher 21:1203-1209.
10. Post AR, Kurc T, Cholleti S, Gao J, Lin X et al (2013) The analytic information warehouse (AIW): a platform for analytics using electronic health record data. J Biomed Inform 46:410-424.
11. Zimmerman JE, Kramer AA, McNair DS, Malila FM (2006) Acute physiology and chronic health evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med 34:1297-1310.
12. Moreno RP, Metnitz PG, Almeida E, Jordan B, Bauer P et al (2005) SAPS 3-from evaluation of the patient to evaluation of the intensive care unit. Part 2: development of a prognostic model for hospital mortality at ICU admissión. Intensive Care Med 31:1345-1355.
13. Escobar GJ, Greene JD, Scheirer P, Gardner MN, Draper D et al (2008) Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases. Med Care 46:232-239.
14. Elixhauser A, Steiner C, Harris DR, Coffey RM (1998) Comorbidity measures for use with administrative data. Med Care 36:8-27.

15. Project HCaU (2015) Comorbidity software, Versión 3.7.
16. Rubin DB, Thomas N (1996) Matching using estimated propensity scores: relating theory to practice. Biometrics 52:249-264.
17. Rubin DB (1997) Estimating causal effects from large data sets using propensity scores. Ann Intern Med 127:757-763.
18. Howell MD, Novack V, Grgurich P, Soulliard D, Novack L et al (2010) Iatrogenic gastric acid suppressión and the risk of nosocomial Clostridium difficile infection. Arch Intern Med 170:784-790.
19. Hernan MA, Brumback B, Robins JM (2000) Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology 11:561-570.
20. Thadani SR, Weng C, Bigger JT, Ennever JF, Wajngurt D (2009) Electronic screening improves efficiency in clinical trial recruitment. J Am Med Inform Assoc 16:869-873.
21. Embi PJ, Jain A, Clark J, Harris CM (2005) Development of an electronic health record-based clinical trial alert system to enhance recruitment at the point of care. AMIA AnnuSymp Proc, 231-235.
22. PCORnet (2015) Rethinking clinical trials: a living textbook of pragmatic clinical trials.
23. Byrne CM, Mercincavage LM, Pan EC, Vincent AG, Johnston DS et al (2010) The value from investments in health information technology at the U.S. Department of Veterans Affairs. Health Aff (Millwood) 29:629-638.
24. Ray WA, Chung CP, Murray KT, Hall K, Stein CM (2009) Atypical antipsychotic drugs and the risk of sudden cardiac death. N Engl J Med 360:225-235.
25. Hogan WR, Wagner MM (1997) Accuracy of data in computer-based patient records. J Am Med Inform Assoc 4:342-355.
26. Lee DS, Donovan L, Austin PC, Gong Y, Liu PP et al (2005) Comparison of coding of heart failure and comorbidities in administrative and clinical data for use in outcomes research. Med Care 43:182-188.
27. Iwashyna TJ, Odden A, Rohde J, Bonham C, Kuhn L et al (2014) Identifying patients with severe sepsis using administrative claims: patient-level validation of the angus implementation of the international consensus conference definition of severe sepsis. Med Care 52: e39-e43.
28. Jones G, Tarightt N, Boelle PY, Marty J, Lalande V et al (2012) Accuracy of ICD-10 codes for surveillance of clostridium difficile infections, France. Emerg Infect Dis 18:979-981.
29. Kramer JR, Davila JA, Miller ED, Richardson P, Giordano TP et al (2008) The validity of viral hepatitis and chronic liver disease diagnoses in Veterans Affairs Administrative databases. Aliment PharmacolTher 27:274-282.
30. van de Garde EM, Oosterheert JJ, Bonten M, Kaplan RC, Leufkens HG (2007) International classification of diseases codes showed modest sensitivity for detecting community-acquired

pneumonia. J Clin Epidemiol 60:834-838.

31. Movig KL, Leufkens HG, Lenderink AW, Egberts AC (2003) Validity of hospital discharge International classification of diseases (ICD) codes for identifying patients with hyponatremia. J Clin Epidemiol 56:530-535.
32. Sickbert-Bennett EE, Weber DJ, Poole C, MacDonald PD, Maillard JM (2010) Utility of international classification of diseases, ninth revisión, clinical modification codes for communicable disease surveillance. Am J Epidemiol 172:1299-1305.
33. Jhung MA, Banerjee SN (2009) Administrative coding data and health care-associated infections. Clin Infect Dis 49:949-955.
34. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF et al (2005) Measuring diagnoses: ICD code accuracy. Health Serv Res 40:1620-1639.
35. Richesson RL, Rusincovitch SA, Wixted D, Batch BC, Feinglos MN et al (2013) A comparison of phenotype definitions for diabetes mellitus. J Am Med Inform Assoc 20: e319-e326.
36. Hripcsak G, Knirsch C, Zhou L, Wilcox A, Melton G (2011) Bias associated with mining electronic health records. J Biomed Discov Collab 6:48-52.
37. Hernan MA, Alonso A, Logan R, Grodstein F, Michels KB et al (2008) Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology 19:766-779.
38. Rusanov A, Weiskopf NG, Wang S, Weng C (2014) Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. BMC Med Inform DecisMak 14:51.
39. Allison PD (2001) Missing data. Sage Publishers, Thousand Oaks.
40. Ingelfinger JR, Drazen JM (2004) Registry research and medical privacy. N Engl J Med 350:1452-1453.
41. Grande D, Mitra N, Shah A, Wan F, Asch DA (2013) Public preferences about secondary uses of electronic health information. JAMA Intern Med 173:1798-1806.
42. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S et al (2007) Toward a national framework for the secondary use of health data: an American medical informatics association white paper. J Am Med Inform Assoc 14:1-9.
43. Platt R, Carnahan RM, Brown JS, Chrischilles E, Curtis LH et al (2012) The U.S. food and drug administration's mini-sentinel program: status and direction. Pharmacoepidemiol Drug Saf 21 (Suppl 1): 1-8.
44. Bradley CJ, Penberthy L, Devers KJ, Holden DJ (2010) Health services research and data linkages: issues, methods, and directions for the future. Health Serv Res 45:1468-1488.
45. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF (2012) A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med

Care 50 (Suppl): S21-S29.

46. Weber GM, Mandl KD, Kohane IS (2014) Finding the missing link for big biomedical data. JAMA 311:2479-2480.
47. Murdoch TB, Detsky AS (2013) The inevitable application of big data to health care. JAMA 309:1351-1352.
48. Birman-Deych E, Waterman AD, Yan Y, Nilasena DS, Radford MJ et al (2005) Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. Med Care 43:480-485.
49. Reynolds HN, McCunn M, Borg U, Habashi N, Cottingham C et al (1998) Acute respiratory distress syndrome: estimated incidence and mortality rate in a 5 million-person population base. Crit Care 2:29-34.
50. Herasevich V, Tsapenko M, Kojicic M, Ahmed A, Kashyap R et al (2011) Limiting ventilator-induced lung injury through individual electronic medical record surveillance. CritCare Med 39:34-39.

Capítulo 8

LA TRAMPA DE LA CONFUSIÓN RESIDUAL EN *BIG DATA*: UNA FUENTE DE ERROR

John Danziger y Andrew J. Zimolzak

Puntos clave

- Cualquier estudio observacional puede tener variables confusoras no identificadas que influencien los efectos de la exposición primaria, por lo tanto debemos confiar en la transparencia de la investigación al igual que en un examen cuidadoso y concienzudo de las limitaciones para poder tener confianza en cualquier hipótesis.
- La fisiopatología es complicada y a menudo confunde los datos medidos con muchas observaciones que son meros *proxies* para un proceso fisiológico y muchos factores diferentes que progresan a una disfunción similar.

8.1 Introducción

Nada es más peligroso que una idea, cuando solo tienes una…
Emile Chartier

La *Big Data* está definida por su vastedad, generalmente con set de datos altamente granulares, que combinados con abordajes analíticos y estadísticos avanzados puede potenciar conclusiones muy convincentes [1]. En esto quizás radique el mayor desafío para utilizar *big data* apropiadamente: entender aquello que no está disponible. Para evitar falsas inferencias de causalidad, resulta crítico reconocer las influencias que pueden afectar el resultado de interés, aunque no sean fácilmente mensurables.

Dada la dificultad de realizar estudios clínicos prospectivos, aleatorizados y bien diseñados en medicina, las fuentes de *Big Data* como la base de datos *Medical Information Mart for Intensive Care (MIMIC)* [2] son muy atractivas. Estas proveen un recurso poderoso para examinar la fuerza de asociaciones potenciales y comprobar si los principios fisiológicos asumidos permanecen robustos en la medicina clínica. De todos modos, teniendo en cuenta su naturaleza observacional, no puede establecerse causalidad y debe tenerse

gran cuidado al utilizar datos observacionales para influenciar patrones de práctica clínica. Hay numerosos ejemplos [3, 4] en la medicina clínica en que se utilizaron datos observacionales para determinar la toma de decisiones clínicas, que luego fueron refutadas y, mientras tanto, potencialmente pudieron causar daño. A pesar de que las asociaciones pueden ser fuertes, no considerar conexiones no detectadas lleva a falsas inferencias. El efecto no reconocido de una asociación entre una variable adicional asociada con la exposición primaria que influencia el resultado de interés es conocido como confundidor o factor de confusión.

8.2 Variables confusoras en *Big Data*

A veces el efecto confusor es referido como "mezcla de efectos" [5] en el que los efectos de la exposición sobre un particular resultado están asociados con un factor adicional, distorsionando por lo tanto la verdadera relación. Así, la confusión puede sugerir falsamente una asociación aparente cuando en realidad no existe asociación. La presencia de factores de confusión es una particular amenaza en los datos observacionales, como es el caso del *Big Data*, debido a la imposibilidad de aleatorizar grupos frente a la exposición. El proceso de aleatorización mitiga la influencia de aquellas influencias no reconocidas ya que estas quedan distribuidas equitativamente en los grupos. De todos modos, los datos observacionales con frecuencia están compuestos por grupos de pacientes que han sido identificados sobre la base de factores clínicos. Por ejemplo, con datos observacionales de cuidados críticos como es el caso de MIMIC, tal "asignación no aleatoria" ocurrió simplemente por ingresar en la unidad de cuidados intensivos (UCI). Ha tenido lugar algún proceso de decisión realizado por el equipo tratante, posiblemente en el Departamento de Emergencias, que consideró que el paciente estaba lo suficientemente enfermo para la UCI. Ese proceso de decisión probablemente es influenciado por una serie de factores, algunos de los cuales resultan identificables como la presión arterial o la severidad de la enfermedad, y otros que no lo son como el hecho de que "el paciente luce enfermo" según la intuición del médico.

8.2.1 La paradoja de la obesidad

Como ejemplo de la sutileza de esta influencia confundidora, abordemos la cuestión de la obesidad como predictor de mortalidad. En la mayoría de

los estudios ecológicos [6, 7] la obesidad está asociada con peores resultados: los pacientes obesos tienen mayor riesgo de morir que aquellos con peso normal probablemente por la mayor incidencia de diabetes, hipertensión y enfermedad cardiovascular. Pese a esto, entre los pacientes ingresados en la UCI la obesidad beneficia fuertemente la supervivencia [8, 9] como muestran múltiples estudios en los que se vio mejores resultados en pacientes obesos críticamente enfermos que en los de peso normal.

Existen muchas explicaciones potenciales para esta asociación paradójica. Por un lado, es plausible que los pacientes obesos críticamente enfermos tengan mayores reservas nutricionales y puedan soportar el prolongado estado de caquexia asociado a enfermedades graves que los pacientes con peso normal. De todas formas, exploremos algunas otras posibilidades. Dado que la obesidad se define con el índice de masa corporal (IMC) al momento de la admisión en la UCI, es posible que influencias no reconocidas sobre el peso corporal previo a la hospitalización que afectan en forma independiente el resultado y que podrían ser la razón verdadera de esta asociación paradójica. Por ejemplo, la retención de fluidos que podría ocurrir en una insuficiencia cardíaca congestiva elevaría el peso corporal, pero no la masa grasa, resultando en una elevación inapropiada del IMC. Cuando esta retención de líquidos deriva en un edema pulmonar suele considerarse como un marcador de gravedad de enfermedad y requiere un mayor nivel de cuidado, como el que se provee en la UCI. Por lo tanto, esta acumulación de líquido llevaría al equipo de la sala de emergencia a ingresar al paciente en la UCI en lugar de una sala general. Ahora, la insuficiencia cardíaca es una enfermedad tratable. Los diuréticos constituyen un tratamiento efectivo y pueden resolver el factor específico, es decir, la sobrecarga hídrica, que llevó a la internación en UCI. Así, este paciente en apariencia obeso, pero que en realidad no lo es, tendría una gran posibilidad de supervivencia. Comparemos este caso con el de otro paciente que desarrolló caquexia por un cáncer metastásico y perdió 30 libras (13.6 kg) antes de presentarse en la sala de emergencia. Dicho paciente habría disminuido significativamente su IMC durante las semanas previas a la enfermedad y su mal estado llevaría a una internación en una UCI, donde tendría peor pronóstico. En el último escenario, concluir que un IMC bajo está asociado con malos resultados no sería estrictamente correcto ya que serían las complicaciones del cáncer subyacente las que conducen a la muerte.

8.2.2 Sesgo de selección

Exploremos una última posibilidad con relación a cómo pueden presentarse confundidores en la paradoja de la obesidad en cuidados intensivos. Imagine dos hermanos gemelos con genética idéntica y las mismas comorbilidades y exposiciones, que presentan celulitis, debilidad y diarrea por lo que requerirán higiene y cambio de ropa frecuente. La única diferencia entre ellos es que uno tiene peso normal, mientras que el otro tiene obesidad mórbida. El equipo de emergencias debe decidir qué nivel de cuidado requieren estos pacientes. Dados los desafíos que conlleva cuidar pacientes con obesidad mórbida (elevar una pierna pesada, rotar para cambiar), es plausible que la obesidad en sí misma influencie la decisión. En este caso habría un gran sesgo de selección. El paciente obeso, que hubiera estado lo suficientemente sano para internarse en una sala general se interna en la UCI (donde comienzan a registrarse los datos observacionales) solamente por la obesidad. En forma no sorprendente, el paciente tendrá mejores resultados que otros internados en la UCI dado que estaba más sano en un primer lugar y fue ingresado allí únicamente porque era obeso.

Este sesgo de selección, que puede ser sutil, es un problema desafiante en estudios con asignación no aleatoria. Los grupos de pacientes se suelen diferenciar por la severidad de enfermedad y, por lo tanto, cualquier estudio observacional que evalúe los efectos de tratamientos puede fallar en abordar los factores asociados subyacentes. Por ejemplo, un estudio observacional reciente con *Big Data* intentó evaluar si la exposición a inhibidores de la bomba de protones (IBP) se encontraba asociada a hipomagnesemia [10]. En efecto, varios miles de los pacientes examinados usuarios de IBP tenían niveles más bajos de magnesio sérico al momento de la admisión. Sin embargo, la causa primaria de la prescripción del medicamento era desconocida. Es plausible que los pacientes que tuvieran dispepsia u otros síntomas gastrointestinales relacionados, que son las principales indicaciones de prescripción de IBP, presentaran menor ingesta de alimentos que contengan magnesio. Por lo tanto, la conclusión de que los IBP eran responsables de la hipomagnesemia sería una conjetura ya que la menor ingesta de magnesio podría ser una explicación igual de razonable.

8.2.3 Fisiopatología incierta

Además del sesgo de selección, como se ve en los ejemplos de la paradoja de la obesidad y en la asociación entre IBP e hipomagnesemia, existe otro importante factor de confusión en estudios de cuidados críticos. Dado que la fisiología y la fisiopatología son fuertes determinantes de los resultados en enfermedades críticas, la posibilidad de dar cuenta de los procesos fisiopatológicos subyacentes es muy importante, pero a la vez notoriamente difícil. Es necesario considerar que aún los clínicos que tratan en forma directa a los pacientes, examinando todos los detalles, a veces no pueden dar cuenta del proceso fisiológico. Reconocer una insuficiencia cardíaca diastólica sigue siendo un desafío. La caracterización precisa del funcionamiento de un órgano no es lineal. Entonces, si el médico no puede describir el proceso, ¿cómo podrían hacerlo los datos observacionales aislados del paciente? No pueden, y esto es una gran fuente de potenciales equivocaciones. Consideremos algunos ejemplos.

En cuidados intensivos, la frecuente cantidad de estudios de laboratorio fácilmente medibles y reproducibles constituyen un objeto atractivo para los estudios transversales. En la literatura casi todos los valores anormales de laboratorio han sido asociados con resultados adversos, incluyendo alteraciones del sodio, potasio, cloro, bicarbonato, uremia, nitrógeno, creatinina, glucosa, hemoglobina, etc. Muchos de estos estudios transversales derivaron en la confección de guías de tratamiento. De cualquier manera, la cuestión central es determinar si el valor de laboratorio anormal por sí solo es el que lleva a malos resultados o, si en cambio, es el proceso fisiopatológico subyacente que conduce a la anormalidad del valor.

Tomemos como ejemplo la hiponatremia. Existen múltiples datos observacionales que vinculan a la hiponatremia con mortalidad. Consecuentemente, surgieron numerosas guías de tratamiento sobre cómo corregir la hiponatremia a través de una combinación de restricción hídrica y administración de sodio [11]. De todos modos, la explicación del mecanismo de acción a través del cual la hiponatremia leve y/o crónica podría causar malos resultados no resulta totalmente convincente. Algunos datos sugieren que el potencial edema cerebral agudo podría conducir a desequilibrios y caídas, pero esta no resulta una explicación completamente convincente para la asociación entre hiponatremia al momento del ingreso y muerte intrahospitalaria.

Fig. 8.1 Mapa conceptual de la asociación entre función renal, determinada por la tasa de filtración glomerular, como determinante de la mortalidad cardiovascular

Muchos estudios transversales no han abordado la causa primaria de la hiponatremia. Frecuentemente la hiponatremia está causada por censado de depleción de volumen como puede ocurrir en la enfermedad hepática y cardíaca. El volumen censado es un concepto que describe una medición interna del volumen intravascular, que afecta directamente la avidez del cuerpo por el sodio y que bajo ciertas condiciones afecta la avidez por el agua. Es bastante difícil determinar clínicamente este volumen censado y no existen códigos de facturación o diagnósticos para describirlo. Por lo tanto, pese a que el volumen censado es el principal determinante de la natremia, en grandes poblaciones no es una variable capturable y no puede incluirse como covariable en un análisis ajustado. Su ausencia probablemente lleve a conclusiones falsas. Al momento, aunque existe una gran cantidad de estudios que muestran que la hiponatremia se asocia con malos resultados, no podemos concluir si tiene mayor importancia el exceso de agua por sí mismo o las alteraciones fisiopatológicas cardíacas o hepática subyacentes que causan la hiponatremia.

Consideremos otro ejemplo importante. Hay una gran cantidad de estudios en cuidados intensivos que vinculan la función renal con una miríada de resultados [12, 13]. Una conclusión indiscutible es que la función renal dañada está asociada con un incremento en la mortalidad cardiovascular, como se ilustra en la Fig. 8.1.

No obstante, esta asociación es sumamente compleja e involucra un número importante de factores de confusión que socavan la conclusión. El primer elemento es con qué precisión la creatinina sérica refleja la tasa de filtración glomerular (TFG). Ecuaciones como la fórmula MDRD (del inglés, *Modificaction of Diet in Renal Disease)* fueron desarrolladas como

herramientas epidemiológicas para estimar la TFG [14], pero no definen en forma precisa la fisiología renal subyacente. Además, aún considerando la creatinina sérica como indicador de la TFG, quedan múltiples aspectos de la función renal sin evaluar, incluyendo el balance de sodio y líquido, la producción de eritropoyetina y vitamina D activada y la función tubular, ninguno de los cuales es fácilmente mensurable y, por ende, no puede ser valorado.

Además, sumado a la confusión debida a la incapacidad de caracterizar con precisión la "función renal", resulta igual de problemática la confusión residual que se debe a la fisiopatología no registrada. En relación con la asociación de la función renal con mortalidad cardiovascular, se encuentran múltiples determinantes de la función cardíaca que simultánea e independientemente influencian tanto la concentración de la creatinina sérica como los resultados cardiovasculares. Por ejemplo, el incremento de la presión venosa yugular es un importante determinante del resultado cardíaco e influencia la función renal a través de la congestión de la vena renal. El gasto cardíaco, la presión arterial pulmonar y la activación del sistema renina-angiotensina-aldosterona probablemente también influencien tanto la función renal como los resultados cardíacos. El mapa conceptual se parece más al de la Fig. 8.2.

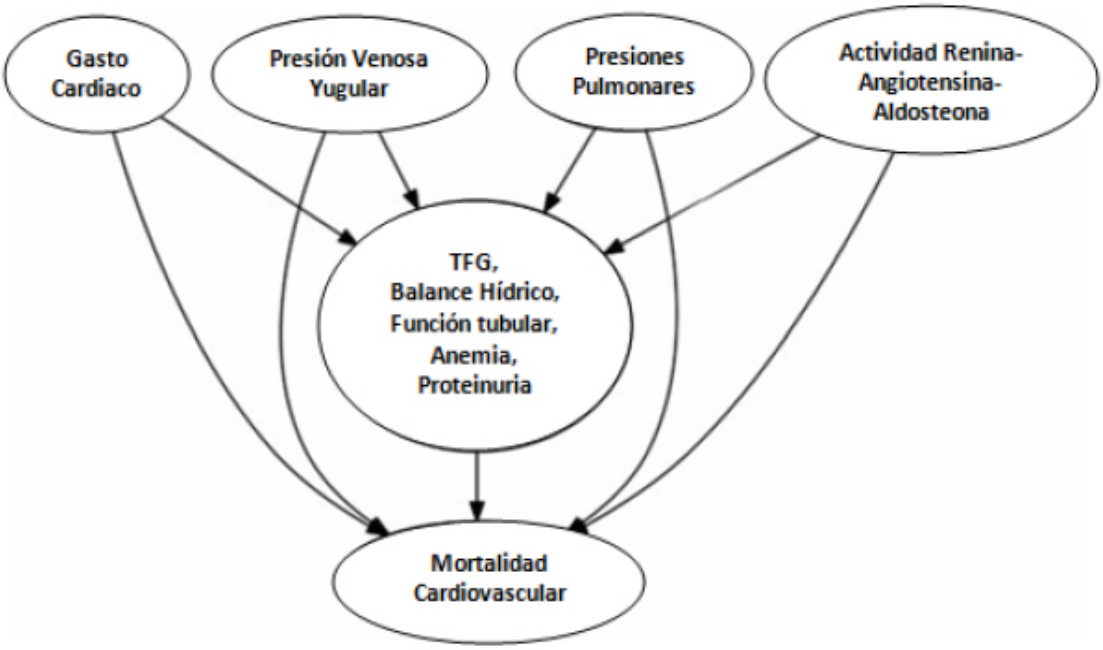

Fig. 8.2 Mapa conceptual de la asociación entre función renal y mortalidad cardiovascular incluyendo más factores de confusión

Dado que muchas de estas variables raramente se miden o cuantifican en grandes estudios epidemiológicos, seguramente exista confusión residual significativa y sesgos potenciales por la incapacidad de comprender la complejidad de los procesos fisiopatológicos subyacentes.

Se han desarrollado muchas técnicas estadísticas para dar cuenta de la confusión residual debida a la no aleatorización y a la severidad de enfermedad en cuidados intensivos. Las puntuaciones de propensión, que intentan captar los factores que llevan a la asignación no aleatorizada (por ejemplo, los factores que influencian la decisión de ingresar en la UCI o de exponer a IBP) se utilizan con frecuencia para minimizar el sesgo de selección [15]. El ajuste utilizando variables que intentan determinar la severidad de enfermedad, como los puntajes *Simplified Acute Physiology Score (SAPS)* [16] o el *Sequential/Sepsis-related Organ Failure Assessment (SOFA) score* [17] o los puntajes de ajuste de comorbilidades, como Charlson o Elixhauser [18, 19], sigue siendo impreciso de la misma manera que continúan siéndolo ajustes de riesgo con el Área Bajo la Curva de la Caracterísitca Operativa del Receptor (*AUCROC, del inglés Area under the reciever operating characteristic curve).* En última instancia, el efecto confundidor significativo no puede ajustarse aún con técnicas estadísticas sofisticadas, por lo que el examen cuidadoso y cauteloso de las limitaciones de cualquier estudio observacional debe ser transparente.

8.3 Conclusiones

En resumen, podemos decir que es necesario ser cuidadosos al cosechar los frutos del "*Big Data*" ya que es justamente aquello que no puede ser visto lo que puede resultar de mayor interés. Se debe ser claro sobre las limitaciones de usar datos observacionales y sugerir que la mayoría de los estudios observacionales que los utilizan son generadores de hipótesis, que luego requieren estudios mejor diseñados para responder la pregunta de investigación.

Referencias

1. Bourne PE (2014) What big data means to me. J Am Med Inf Assoc. 21 (2): 194-194.
2. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L-W, Moody G et al (2011) Multiparameter intelligent monitoring in intensive care II: a public-access intensive care unit database. Crit Care Med 39 (5): 952-960.
3. Patel CJ, Burford B, Ioannidis JPA (2015) Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. J Clin Epidemiol 68 (9): 1046-1058.
4. Tzoulaki I, Siontis KCM, Ioannidis JPA (2011) Prognostic effect size of cardiovascular biomarkers in datasets from observational studies versus randomised trials: meta-epidemiology study. BMJ 343: d6829.
5. Greenland S (2005) Confounding. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics, 2nd edn.
6. National Task Force on the Prevention and Treatment of Obesity (2000) Overweight, obesity, and health risk. Arch Intern Med 160 (7): 898-904.
7. Berrington de Gonzalez A, Hartge P, Cerhan JR, Flint AJ, Hannan L, MacInnis RJ et al (2010). Body-mass index and mortality among 1.46 million white adults. N Engl J Med 363 (23): 2211-2219.
8. Hutagalung R, Marques J, Kobylka K, Zeidan M, Kabisch B, Brunkhorst F et al (2011) The obesity paradox in surgical intensive care unit patients. Intensive Care Med 37 (11): 1793-1799.
9. Pickkers P, de Keizer N, Dusseljee J, Weerheijm D, van der Hoeven JG, Peek N (2013) Body mass index is associated with hospital mortality in critically ill patients: an observational cohort study. Crit Care Med 41 (8): 1878-1883.
10. Danziger J, William JH, Scott DJ, Lee J, Lehman L-W, Mark RG et al (2013) Proton-pump inhibitor use is associated with low serum magnesium concentrations. Kidney Int 83 (4): 692-699.
11. Verbalis JG, Goldsmith SR, Greenberg A, Korzelius C, Schrier RW, Sterns RH et al (2013) Diagnosis, evaluation, and treatment of hyponatremia: expert panel recommendations. Am J Med 126 (10 Suppl 1): S1-S42.

12. Apel M, Maia VPL, Zeidan M, Schinkoethe C, Wolf G, Reinhart K et al (2013) End-stage renal disease and outcome in a surgical intensive care unit. Crit Care 17 (6): R298.
13. Matsushita K, van der Velde M, Astor BC, Woodward M, Levey AS et al (2010) Chronic kidney disease prognosis consortium. Association of estimated glomerular filtration rate and albuminuria with all-cause and cardiovascular mortality in general population cohorts: a collaborative meta-analysis. Lancet 375 (9731): 2073-2081.
14. Levey AS, Bosch JP, Lewis JB, Greene T, Rogers N, Roth D (1999) A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Modification of diet in renal disease study group. Ann Intern Med 130 (6): 461-470.
15. Gayat E, Pirracchio R, Resche-Rigon M, Mebazaa A, Mary J-Y, Porcher R (2010) Propensity scores in intensive care and anaesthesiology literature: a systematic review. Intensive Care Med 36 (12): 1993-2003.
16. Le Gall JR, Lemeshow S, Saulnier F (1993) A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. JAMA 270 (24): 2957-2963.
17. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H et al (1996) The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. On behalf of the working group on sepsis-related problems of the European society of intensive care medicine. Intensive Care Med 22 (7): 707-710.
18. Charlson ME, Pompei P, Ales KL, MacKenzie CR (1987) A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 40 (5): 373-383.
19. Elixhauser A, Steiner C, Harris DR, Coffey RM (1998) Comorbidity measures for use with administrative data. Med Care 36 (1): 8-27.

PARTE II

UN LIBRO DE RECETAS:
DE FORMULAR LA PREGUNTA DE INVESTIGACIÓN
A LA VALIDACIÓN DE LOS HALLAZGOS

La primera parte de este libro de texto ha brindado al lector una perspectiva general de las historias clínicas electrónicas (HCEs), su potencial para la investigación médica y su uso para el análisis de datos retrospectivo. La parte II se enfoca en el uso de una HCE particular, la base de datos Medical Information Mart for Intensive Care (MIMIC), curada por el Laboratorio de Fisiología Computacional del Instituto de Tecnología de Massachusetts (MIT). Los lectores tendrán la oportunidad de desarrollar sus habilidades analíticas para la minería de datos clínicos mientras que siguen un proyecto de investigación completo, desde la definición inicial de la pregunta de investigación hasta la evaluación de la robustez de los resultados. Esta parte está diseñada como un libro de cocina; cada capítulo está integrado por algunos conceptos teóricos, seguidos de ejemplos trabajados usando la base MIMIC. La parte III de este libro estará dedicada a una variedad de diferentes casos de estudio para favorecer la comprensión de métodos de análisis más avanzados. Esta parte está subdividida en 9 capítulos que siguen el proceso común de generación de nueva evidencia médica utilizando minería de datos clínicos. En el capítulo 9, el lector aprenderá como transformar una pregunta clínica en una pregunta de investigación pertinente, que incluye definir un apropiado diseño del estudio y seleccionar la exposición y los resultados de interés. En el capítulo 10, el investigador aprenderá a definir qué población de pacientes es más relevante para dar respuesta a la pregunta de investigación. El análisis de las HCEs es con frecuencia un aspecto esencial y desafiante, por ende será descripto en forma detallada en los siguientes 4 capítulos. Los capítulos 11 y 12 tratan las tareas esenciales de preparación de datos y pre-procesamiento, obligatorias antes de que cualquier dato pueda alimentar una herramienta de análisis estadístico. El capítulo 11 explica cómo se estructura una base de datos, qué tipo de datos pueden contener y como extraer las variables de interés usando consultas. El capítulo 12 presenta algunos métodos comunes de pre-procesamiento de datos, que habitualmente implican la limpieza, integración y posterior reducción de

datos, el capítulo 13 brinda varios métodos de tratar los datos faltantes, el capítulo 14 discute las técnicas para identificar y tratar los valores *ouliers*. En el capítulo 15 se presentan los métodos comunes para explorar datos, tanto numéricos como gráficos. El análisis exploratorio de datos da al investigador una visión invalorable de las características y potenciales aspectos del set de datos y puede ayudar a generar más hipótesis. El capítulo 16, análisis de datos, presenta la teoría y métodos para desarrollo de modelos (sección 16.1) así como técnicas comunes de análisis de datos en estudios clínicos, llamados regresión lineal (sección 16.2), regresión logística (sección 16.3) y análisis de sobrevida (sección 16.4). Finalmente, el capítulo 17 discute los principios de la validación de modelos y análisis de sensibilidad, que prueban la robustez de los resultados de una investigación particular de acuerdo a los distintos supuestos del modelo.

Cada capítulo incluye elaborados ejemplos inspirados en un único estudio, publicado en Chest en el año 2015 por Hsu y col, que aborda una pregunta clave en la práctica clínica en cuidados intensivos: “¿la colocación de un catéter arterial invasivo (CAI) está asociado con disminución de la mortalidad en pacientes en ventilación mecánica pero que no requieren soporte vasopresor?”. Los accesos arteriales son ampliamente utilizados en la unidad de cuidados intensivos para el monitoreo continuo de la presión arterial y se piensa que es más preciso y confiable que el monitoreo no invasivo estándar. También tienen el beneficio adicional de permitir la extracción de muestras de sangre arterial de forma más sencilla, pudiendo reducir la necesidad de punciones arteriales repetidas. Sin embargo, teniendo en cuenta su naturaleza invasiva, los CAI conllevan el riesgo de infección del torrente sanguíneo y lesión vascular, por lo cual se requiere evaluar la evidencia de un beneficio potencial.

El principal resultado de interés seleccionado fue la mortalidad a los 28 días; los resultados secundarios incluían la duración de la estancia en la UCI y en el hospital, la duración de la ventilación mecánica, y el promedio de mediciones de gas en sangre realizadas.

Los autores identificaron la “colocación de catéteres arteriales” como su exposición de interés y llevaron a cabo un análisis de puntaje de propensión para probar la relación entre la exposición y los resultados usando MIMIC.

El resultado en este conjunto de datos en particular (alerta de spoiler) es que la presencia de un CAI no se asocia con una diferencia en la mortalidad

a los 28 días, en pacientes hemodinámicamente estables en ventilación mecánica. Este caso de estudio proporciona una base para aplicar la teoría anterior a un ejemplo de trabajo, y dará al lector una perspectiva de primera mano sobre varios aspectos de la minería de datos y técnicas analíticas. Esto no es de ninguna manera una exploración exhaustiva de la analítica de las HCEs y, cuando el caso carece de los detalles necesarios, hemos intentado incluir información adicional relevante para las técnicas analíticas comunes.

Para el lector interesado, se proporcionan referencias para lecturas más detalladas.

Capítulo 9

FORMULANDO LA PREGUNTA DE INVESTIGACIÓN

Anuj Metha, Brian Malley y Allan Walkey

Objetivos de Aprendizaje

- Comprender cómo convertir una pregunta clínica en una pregunta de investigación.
- Principios de elección de la muestra.
- Enfoques y dificultades potenciales.
- Principios para definir la exposición de interés.
- Principios para definir el resultado.
- Selección del apropiado diseño de estudio

9.1 Introducción

La pregunta clínica que surge en el momento de tomar la mayoría de las decisiones clínicas es: "¿Esto ayudará a mi paciente?". Antes de embarcarse en una investigación que proporcione datos que puedan ser utilizados para explicar la pregunta clínica, ésta debe ser transformada en una pregunta de investigación. El proceso de desarrollar una pregunta de investigación implica definir distintos componentes del estudio y también el tipo de estudio más adecuado para utilizar dichos componentes de forma que permita obtener resultados válidos y fiables. Estos componentes incluyen: ¿En quién es relevante esta pregunta de investigación? La población de sujetos definida por el investigador se denomina muestra. El fármaco, la maniobra, evento o característica en la que basamos nuestra hipótesis alternativa se llama exposición de interés. Por último, debe definirse el resultado de interés. Con estos componentes en mente, el investigador debe decidir cuál es el mejor diseño de estudio o el más factible para responder la pregunta. Si se elige un estudio observacional, entonces la elección de una base de datos también es crucial.

En este capítulo, exploraremos cómo los investigadores pueden trabajar en convertir una pregunta clínica en una pregunta de investigación utilizando el escenario clínico del uso de catéteres arteriales invasivos (CAI) durante la ventilación mecánica (VM). Además, discutiremos las fortalezas y

debilidades de los diseños comunes de estudios incluidos los ensayos controlados, aleatorizados, así como los estudios observacionales.

9.2 El Escenario Clínico: el Impacto de los Catéteres Arteriales Invasivos

Los pacientes que requieren VM porque no pueden mantener una respiración adecuada por sí mismos (ej., por una neumonía severa o una crisis asmática) son a menudo los pacientes más enfermos en el hospital, con índices de mortalidad que superan el 30% [1-3]. Existen múltiples opciones para monitorear la adecuación del soporte ventilatorio para los pacientes críticamente enfermos que requieren VM, que van desde mediciones transcutáneas no invasivas hasta sistemas de monitoreo invasivos continuo. Los catéteres arteriales invasivos son dispositivos de monitoreo invasivos que permiten el monitoreo continuo de la presión arterial en tiempo real y facilitan el acceso a muestras de sangre arterial para evaluar el pH, niveles de oxígeno y dióxido de carbono, entre otros [4-6]. Mientras que el monitoreo más cercano de los pacientes que requieren VM con CAIs puede parecer a primera vista beneficioso, los CAIs pueden dar lugar a graves eventos adversos, incluyendo la pérdida de flujo de sangre en la mano e infecciones [7,8]. Actualmente, faltan datos para definir si los beneficios pueden superar los riesgos de un monitoreo invasivo usando CAIs. Examinar los factores asociados con la decisión de utilizar CAIs, y los resultados en pacientes en los que se utilizaron catéteres arteriales invasivos continuos en comparación con los que tuvieron controles no invasivos solamente, puede proveer información útil para los médicos que se enfrentan a la decisión de colocar o no un CAI.

9.3 Convirtiendo Preguntas Clínicas en Preguntas de Investigación

El primer paso en el proceso de transformar una pregunta clínica en una de investigación es definir cuidadosamente la muestra de estudio (o cohorte de pacientes), la exposición de interés, y el resultado de interés. Estos 3 componentes –muestra, variable y resultado– son partes esenciales de cualquier pregunta de investigación. Pequeñas variaciones de cada componente pueden afectar drásticamente las conclusiones derivadas a partir de cualquier estudio de investigación, y también si la investigación abordará en forma adecuada la pregunta clínica general.

9.3.1 Muestra del Estudio

En el caso del uso del catéter arterial invasivo, podrían imaginarse muchas potenciales muestras de estudio de interés: por ejemplo, se podría incluir a todos los pacientes de UCI, todos los pacientes que reciben VM, todos los pacientes que reciben medicación intravenosa que actúa sobre la presión arterial, sólo adultos, únicamente niños, etc. Como alternativa, se podrían definir las muestras basadas en enfermedades o síndromes específicos, como el shock (donde los CAI pueden ser usados para un monitoreo cercano de la presión arterial) o asma severo (donde los CAI pueden ser usados para monitorear niveles de oxígeno o dióxido de carbono).

La elección de la muestra afectará tanto la validez interna como la externa (posibilidad de generalización) del estudio. Un estudio enfocado sólo en una población pediátrica puede no aplicarse en una población adulta. De igual forma, un estudio enfocado en pacientes que reciben VM puede no ser aplicable en pacientes no ventilados. Más aún, un estudio que incluye pacientes con diferentes indicaciones de uso de un CAI, con diferentes resultados relacionados con la indicación de uso del catéter arterial invasivo, puede carecer de validez interna debido al sesgo llamado "*confundidor*". Un confundidor es un tipo de sesgo en el que una variable de exposición se asocia tanto con la exposición como con el resultado.

Por ejemplo, si se estudian los beneficios de los CAIs sobre la mortalidad en todos los pacientes que reciben VM, los investigadores deben tomar en cuenta el hecho de que la colocación de un CAI puede ser en realidad indicativo de una mayor severidad de enfermedad. Por ejemplo, imagine un estudio con una muestra de pacientes en VM, en el cual aquellos que tienen un shock séptico reciben un CAI para facilitar la medicación vasoactiva y permitir un monitoreo cercano de la presión arterial mientras que pacientes con asma no recibieron un CAI ya que se utilizaron otros métodos para monitorizar su ventilación (como el monitoreo de la presión parcial de CO_2 exhalado). Los pacientes con shock séptico tienden a tener una enfermedad mucho más severa comparados con los pacientes con asma independientemente de si fue colocado un CAI. En dicho estudio, los investigadores podrían concluir que los CAIs se asocian a mayor mortalidad sólo porque fueron utilizados en pacientes más enfermos con un mayor

riesgo de morir. La variable “diagnóstico” es por lo tanto un factor confundidor, asociado con la exposición (decisión de colocar un CAI) y el resultado (la muerte). Una selección cuidadosa de la muestra es un método para tratar de abordar los problemas de confusión relacionados con la severidad de la enfermedad. Restringir las muestras del estudio para excluir los grupos que pueden confundir fuertemente los resultados (ej., excluir pacientes con medicación vasoactiva) es una estrategia para reducir el sesgo. Sin embargo, la selección de muestras homogéneas para incrementar la validez interna debería ser equilibrada con el deseo de generalizar los hallazgos del estudio a poblaciones de pacientes más amplias. Estos principios se discuten más extensamente en el capítulo 10 – “Selección de Cohortes”.

9.3.2 Exposición

La exposición en nuestra pregunta de investigación parece ser bastante clara: la colocación de un CAI. Sin embargo, se debe prestar especial atención a la definición de cada exposición o variable de interés. La clasificación errónea de las exposiciones puede sesgar los resultados. ¿Cómo debe medirse la variable CAI? Por ejemplo, los investigadores pueden usar métodos que van desde la revisión directa de la historia clinica hasta el uso de datos administrativos (ej., Clasificación Internacional de Enfermedades, códigos CIE) para identificar el uso de CAI. Cada método para averiguar la exposición de interés puede tener beneficios (mejor precisión de la revisión de las historias clínicas) y contras (muchas horas-persona para realizar una revisión manual de fichas).

Definir la ventana de tiempo durante la cual se mide una exposición de interés puede también tener consecuencias importantes que deben ser consideradas cuando se interpretan los resultados de la investigación. Para el propósito de nuestro estudio de CAI, la presencia de un CAI fue definida como tener colocado un CAI luego de iniciar la VM. El carácter temporal de la exposición es esencial para responder la pregunta clínica; algunos de los CAI colocados antes de la VM son para el monitoreo de pacientes quirúrgicos de bajo riesgo en la sala de operaciones. La inclusión de todos los pacientes con CAI, independientemente del momento, puede sesgar los resultados hacia un beneficio para los CAI al incluir a muchos pacientes, por

lo demás sanos, a los que se les colocó un CAI para monitorización quirúrgica. Como alternativa, si el grupo de exposición es definido como los pacientes que tuvieron un CAI durante al menos 48 horas luego del inicio de la VM, el estudio está en riesgo para un tipo de confundidor llamado *"sesgo de tiempo inmortal":* sólo los pacientes que estuvieran vivos podrían haber tenido colocado un CAI, mientras que los pacientes que murieron antes de las 48 horas (en teoría más enfermos) no podrían haber tenido un CAI.

Es tan importante definir el grupo de pacientes que recibieron o experimentaron una exposición, como el grupo "no expuesto" o el grupo de control. Si bien no todas las investigaciones requieren un grupo de control (ej., estudios epidemiológicos), el grupo de control es necesario para evaluar la efectividad de las intervenciones médicas. En el caso del estudio de los CAI, el grupo de control es bastante sencillo: los pacientes en VM que no tuvieron colocado un CAI. Sin embargo, hay importantes matices al definir los grupos de control. En nuestro ejemplo de estudio, un grupo alternativo de control podrían ser todos los pacientes en UCI que no recibieron un CAI. Sin embargo, la inclusión de pacientes que no reciben VM resulta en un grupo de control con enfermedades de menor gravedad y mortalidad esperada que los pacientes que reciben VM; lo cual podría sesgar a favor de no utilizar CAIs. Se necesita una cuidadosa elección del grupo control para interpretar en forma apropiada cualquier conclusión de la investigación; definir un grupo control apropiado es tan importante como definir la exposición.

9.3.3 Resultados

Finalmente, el investigador necesita determinar los resultados de interés. Pueden considerarse varios tipos diferentes de resultados, incluyendo resultados intermedios o mecanísticos (informa las vías etiológicas, pero puede no tener un impacto inmediato en los pacientes), resultados centrados en el paciente (informa resultados importantes para los pacientes, pero puede carecer de conocimientos mecanísticos: ej., escalas de comodidad, índices de calidad de vida o mortalidad), o resultados centrados en el sistema de salud (ej., utilización de recursos, o costos). En nuestro ejemplo del uso de CAI, pueden considerarse varios resultados incluyendo resultados intermedios (ej., número de extracciones de sangre

arterial, cambios en los parámetros de ventilación, o cambios en la medicación vasoactiva), resultados centrados en el paciente (ej., mortalidad a los 28 ó 90 días, tasas de eventos adversos), o el uso del sistema sanitario (ej., costos de hospitalización, sobrecarga de trabajo del médico). Como se muestra en nuestro ejemplo, los resultados pueden complementarse entre sí para llegar a una constelación de hallazgos que proporcione un cuadro más completo para abordar la pregunta de interés clínico.

Luego de definir claramente la muestra del estudio, la exposición y el resultado de interés, puede formularse la pregunta de investigación.

En nuestro ejemplo, la pregunta de investigación podría ser formulada como sigue:

"*En la población de interés (**cohorte de estudio**), ¿la exposición a la **variable de interés** está asociada con un resultado diferente que en el **grupo control**?,* lo cual se convierte en nuestro ejemplo:

"Entre los pacientes adultos de UCI en ventilación mecánica que no reciben medicación vasoactiva (ej., la muestra de estudio), ¿se asocia la colocación de un CAI luego de iniciar la VM (en comparación con quienes no reciben un CAI) (ej. pacientes expuestos y grupo control) con una mejora en las tasas de mortalidad a 28 días (resultado primario centrado en el paciente) y el número de mediciones de gases en sangre por día (apoyando el resultado secundario, intermedio/mecanísticos)?

9.4 Adecuando el Diseño del Estudio a la Pregunta de Investigación

Una vez que se ha definido la pregunta de investigación, el próximo paso es elegir el mejor diseño del estudio teniendo en cuenta la pregunta y los recursos disponibles. En la investigación biomédica, el *gold-standard* para los diseños de estudio sigue siendo el ensayo clínico doble ciego, aleatorizado placebo control (ECA) [9, 10]. En los ECA, los pacientes con una condición determinada (ej., todos los adultos que reciben VM) serían aleatorizados para recibir un fármaco o intervención de interés (ej., CAI) o aleatorizados para recibir el control (ej., no CAI), con una cuidadosa medición de los *resultados* predeterminados (ej., mortalidad a 28 días). En condiciones ideales, el proceso de aleatorización elimina todos los confundidores medidos y no medidos y permite hacer conclusiones causales que generalmente no pueden lograrse sin la aleatorización. Como

se muestra anteriormente, los confundidores son una amenaza para las inferencias válidas de los resultados de los estudios. Alternativamente en nuestro ejemplo de shock séptico versus asma, la gravedad de la enfermedad asociada con la condición subyacente puede constituir otro confundidor. La aleatorización basada exclusivamente en la exposición de interés intenta suprimir los factores de confusión. En nuestros ejemplos, la aleatorización adecuada en una gran muestra, teóricamente crearía distribuciones iguales de edad, igual número de pacientes con shock séptico y con asma en los grupos de exposición y control.

Sin embargo, los ECA tienen varias limitaciones. Aunque sus fundamentos teóricos son bastante simples, la compleja logística del reclutamiento y la permanencia de los pacientes, el consentimiento informado, la aleatorización, el seguimiento, y el cegamiento pueden resultar en ECAs que se desvíen de las "condiciones ideales" necesarias para una inferencia causal no sesgada. Además, los ECA tienen el mayor potencial de causar daño a los pacientes y requieren un monitoreo intensivo ya que el estudio dicta qué tipo de tratamiento recibe un paciente (en vez del médico) y puede desviarse del cuidado habitual. Dada la complejidad logística, los ECA suelen ser costosos y consumen mucho tiempo, llevando con frecuencia muchos años y millones de dólares para completarse. Incluso cuando la logística es factible, los ECA muchas veces 'eliminan' varios grupos de pacientes con el fin de minimizar los daños potenciales y maximizar la detección de asociaciones entre las intervenciones y los resultados de interés. Como resultado, los ECA pueden consistir en grupos de pacientes homogéneos que cumplen con criterios estrictos, lo cual puede reducir la validez externa de los hallazgos de los estudios. A pesar de muchos esfuerzos y costos, un ECA puede perder relevancia para la pregunta clínica de si la intervención de interés es de ayuda para su paciente particular o no. Finalmente, algunas preguntas clínicas éticamente no pueden ser respondidas con un ECA. Por ejemplo, la asociación entre el tabaquismo y el cáncer de pulmón nunca se ha demostrado en un ECA, ya que no es ético aleatorizar a los pacientes para que empiecen a fumar en un grupo de intervención de fumadores, ¡o aleatorizar pacientes a un grupo control en un ensayo para investigar la efectividad de un paracaídas! [11]

La investigación observacional difiere de los ECA. Los estudios observacionales no son experimentales; los investigadores registran los

patrones de la práctica clínica de rutina y obtienen conclusiones basadas en correlaciones y asociaciones sin intervenciones activas [9, 12]. Los estudios observacionales pueden ser retrospectivos (basados en datos que ya han sido recolectados), prospectivos (los datos son recolectados activamente en el tiempo), o bidireccionales (mixtos). A diferencia de los ECA, en los estudios observacionales los investigadores no tienen un rol en la decisión de qué tipo de tratamiento o intervención reciben los pacientes. Los estudios observacionales suelen ser menos complicados logísticamente que los ECA ya que no hay intervenciones activas, no hay aleatorización ni comités de monitoreo de datos, y los datos con frecuencia son recolectados retrospectivamente. Así, los estudios observacionales conllevan menor riesgo de dañar a los pacientes (distintos de la pérdida de la confidencialidad de los datos que se han recogido) que los ECA, y tienden a ser menos costosos y consumir menos tiempo. Las bases de datos retrospectivas como MIMIC-II [13] o *National Inpatient Sample* [14] también pueden proporcionar muestras de estudio mucho más grandes (decenas de miles en algunos casos) que podrían ser inscriptos en un ECA, permitiendo un mayor poder estadístico. Además, en los estudios observacionales frecuentemente se incluyen muestras de estudio más amplias, lo que conduce a una mayor generalizabilidad de los resultados a un rango de pacientes más amplio (validez externa). Finalmente, algunas preguntas clínicas que no sería ético estudiar en un ECA pueden ser investigadas con estudios observacionales. Por ejemplo, la asociación entre el cáncer de pulmón y el consumo de tabaco ha sido demostrada con múltiples grandes estudios epidemiológicos prospectivos [15, 16] y los efectos salvavidas de los paracaídas han sido demostrados en su mayoría a través del poder de la observación.

Aunque logísticamente más simples que los ECA, los fundamentos teóricos de los estudios observacionales son generalmente más complejos que los ECA. La obtención de estimaciones causales del efecto de una exposición específica para un resultado específico depende del concepto filosófico de lo “contrafáctico” [17]. Lo contrafáctico es la situación en la cual, siendo todos iguales, el mismo sujeto de investigación al mismo tiempo recibiría la exposición de interés y (lo contrafáctico) no recibiría la exposición de interés, con el mismo resultado medido en el sujeto de investigación expuesto y no expuesto. Como en el mundo real no podemos

crear sujetos de investigación clonados, nos apoyamos en la creación de grupos de pacientes similares al grupo que recibe la intervención de interés. En el caso de un ECA ideal con un número suficientemente grande de sujetos, el proceso de aleatorización utilizado para seleccionar los grupos de intervención y control crea dos 'universos' alternativos de pacientes que serán similares excepto en lo relativo a la exposición de interés. Dado que los estudios observacionales no pueden intervenir en los sujetos de estudio, crean experimentos naturales en los que el grupo contrafáctico es definido por el investigador y por los procesos clínicos que ocurren en el mundo real. Mayormente, los procesos clínicos del mundo real ocurren a menudo por una razón, y estas razones pueden causar una desviación de los ideales contrafácticos en los que los sujetos de estudio expuestos y no expuestos difieren de manera importante. En resumen, los estudios observacionales pueden ser más propensos al sesgo (problemas con la validez interna) que los ECA debido a la dificultad de obtener el grupo de control contrafáctico.

Se han identificados varios tipos de sesgos en los estudios observacionales. Los sesgos de selección ocurren cuando el proceso de selección de los pacientes expuestos y no expuestos introduce un sesgo al estudio. Por ejemplo, el tiempo entre el comienzo de la VM y la colocación del CAI puede introducir un tipo de "sesgo de selección de tratamiento de sobrevivientes", ya que los pacientes que recibieron un CAI no pudieron haber muerto antes de recibir los CAI. Los sesgos de información provienen de la medición o clasificación errónea de ciertas variables. En los estudios retrospectivos, los datos ya han sido recolectados y muchas veces es difícil de evaluar errores en los datos. Otro sesgo importante en los estudios observacionales es la confusión. Como se indica, la confusión ocurre cuando una tercera variable se correlaciona tanto con la exposición como con el resultado. Si no se tiene en cuenta la tercera variable, puede inferirse una falsa relación entre la exposición y el resultado. Por ejemplo, el tabaquismo es un confundidor importante en muchos estudios observacionales ya que se asocia con varios otros comportamientos como el consumo de café y alcohol. Un estudio que investigue la relación entre el consumo de café y la incidencia de cáncer de pulmón puede concluir que los individuos que toman más café tienen mayores tasas de cáncer de pulmón. Sin embargo, como el tabaquismo se asocia tanto con el consumo de café como con el cáncer de pulmón, es un confundidor en la relación entre el consumo de

café y el cáncer de pulmón si no se mide y no se tiene en cuenta en el análisis. Se han desarrollado diversos métodos para intentar abordar la confusión en la investigación observacional tales como ajustar por el factor de confusión, si se conoce y se mide, en las ecuaciones de regresión; el emparejamiento de cohortes por confundidores conocidos y el uso de variables instrumentales – métodos que serán explicados en detalle en próximos capítulos. Como alternativa, se puede restringir la muestra de estudio (ej. excluyendo a los pacientes con shock de un estudio que evalúa la utilidad de los CAI). Por estas razones, aunque poderoso, un estudio observacional individual puede, en el mejor de los casos, demostrar asociaciones y correlaciones y no puede probar la causalidad. Con el tiempo, una suma acumulativa de múltiples estudios observacionales de alta calidad junto con otra evidencia mecanística, puede llevar a conclusiones causales, como es el caso de la relación causal aceptada actualmente entre el tabaquismo y el cáncer de pulmón, establecida por estudios observacionales en humanos y ensayos experimentales en animales.

9.5 Tipos de Investigación Observacional

Hay muchos tipos diferentes de preguntas que pueden ser respondidas con una investigación observacional (Tabla 9.1). Los estudios epidemiológicos son uno de los principales tipos de investigación observacional que se centra en la carga de enfermedad en poblaciones predefinidas. Estos tipos de estudio suelen intentar definir incidencia, prevalencia, y factores de riesgo de enfermedad. Además, los estudios epidemiológicos también pueden investigar cambios en el cuidado de la salud o enfermedad a lo largo del tiempo. Los estudios epidemiológicos son la piedra angular de la salud pública y pueden influenciar fuertemente las decisiones políticas, la asignación de recursos y la atención del paciente. En el caso del cáncer de pulmón, grupos predefinidos de pacientes sin cáncer de pulmón fueron monitoreados por años hasta que algunos pacientes desarrollaron cáncer de pulmón. Luego los investigadores compararon numerosos factores de riesgo, como el tabaquismo, entre aquellos que desarrollaron y que no desarrollaron cáncer de pulmón lo que llevó a la conclusión de que el tabaquismo incrementa el riesgo de cáncer de pulmón

[15, 16]. Existen otros tipos de estudios epidemiológicos que se basan en principios similares a los estudios observacionales pero difieren en el tipo de preguntas planteadas.

Tabla 9.1 Principales tipos de investigación observacional y su propósito

Tipo de Investigación observacional	**Propósito**
Epidemiológica	Definir incidencia, prevalencia y factores de riesgo de enfermedad
Modelado predictivo	Predecir resultados futuros
Efectividad comparativa	Identificar intervenciones asociadas con resultados superiores
Farmacovigilancia	Detectar efectos adversos a drogas raros que ocurren en el largo plazo

Los estudios de modelado predictivo desarrollan modelos que son capaces de predecir de forma precisa los resultados futuros en grupos específicos de pacientes. En estudios predictivos, los investigadores definen un resultado de interés (ej., mortalidad hospitalaria) y usan los datos recolectados de los pacientes, como estudios de laboratorio, signos vitales, y estados de enfermedad para determinar qué factores contribuyeron al resultado. Luego los investigadores validan los modelos desarrollados a partir de grupo de un pacientes en un grupo distinto. Los estudios de modelado predictivo desarrollaron muchas escalas predictivas comúnmente utilizadas en la práctica clínica tales como la Escala de Riesgo Cardiovascular de Framingham [18], APACHE IV [19], SAPS II [20] y SOFA [21].

La investigación de efectividad comparativa es otra forma de investigación observacional que involucra la comparación de las intervenciones médicas existentes con el fin de determinar métodos efectivos de prestar atención sanitaria. La investigación de efectividad comparativa, a diferencia de los estudios epidemiológicos descriptivos, compara resultados entre pacientes similares que recibieron diferentes tratamientos para evaluar cuál de las intervenciones puede estar asociada con mejores resultados en condiciones del mundo real. Esto podría implicar la comparación de un fármaco A con un fármaco B o podría suponer la comparación de una intervención con un grupo de control que no recibió

dicha intervención. Dado que a menudo existen razones subyacentes por las que un paciente recibe el tratamiento A versus el B, o recibe o no una intervención, los estudios de efectividad comparativa deben tener meticulosamente en cuenta los potenciales factores de confusión. En el caso de los CAI, la pregunta de investigación que compara los pacientes que tenían colocado un CAI con aquellos que no lo tenían, representaría un estudio de efectividad comparativa.

Los estudios de farmacovigilancia son otra forma de investigación observacional. Como muchos ensayos de fármacos y dispositivos terminan luego de 1 ó 2 años, los métodos observacionales son utilizados para evaluar si hay patrones de efectos adversos raros que ocurren a largo plazo. La fase IV de los estudios clínicos es una forma de estudios de farmacovigilancia en la cual la información relacionada con la eficacia y el daño a largo plazo es reunida luego de que la droga ha sido aprobada.

9.6 Eligiendo la Base de Datos Correcta

Una parte fundamental del proceso de investigación consiste en decidir qué tipos de datos son necesarios para responder la pregunta de investigación. Los datos administrativos/de reclamos, el uso secundario de datos de ensayos clínicos, los estudios epidemiológicos prospectivos, y los sistemas de historia clínica electrónica (HCE) (tanto de las instituciones individuales como aquellas agregadas de múltiples instituciones) son distintas fuentes a partir de las cuales se pueden construir bases de datos. Las bases de datos administrativas, tales como *National Inpatient Sample* y *State Inpatient Databases,* integrantes del Proyecto HCUP (por sus siglas en inglés, Healthcare Cost and Utilization Project) o la base de datos *Medicare*, contienen información sobre la demografía de los pacientes y hospitales así como los códigos de facturación y de procedimientos. Se han desarrollado varias técnicas para transformar estos códigos de facturación y procedimientos en descripciones de enfermedades más útiles clínicamente. Las bases de datos administrativas tienden a proporcionar muestras de gran tamaño y, en algunos casos, pueden ser representativas de una población entera. Sin embargo, carecen de datos detallados a nivel del paciente referentes a la hospitalización, como los signos vitales, datos de laboratorio y microbiológicos, datos de tiempo (como la duración de la VM o los días

con un CAI) o datos farmacológicos, los cuales a menudo son importantes para lidiar con posibles confundidores.

Otra fuente de datos común para la investigación observacional son los grandes estudios epidemiológicos como el Estudio de Framingham así como grandes ECA multicéntricos como la red NIH ARDS Network. Los datos ya existentes pueden ser analizados retrospectivamente con nuevas preguntas de investigación en mente. Como los datos originales fueron recolectados con fines de investigación, estos tipos de bases de datos suelen tener información detallada y pormenorizada no disponible en otras bases de datos clínicas. Sin embargo, muchas veces los investigadores están limitados por el alcance de la recopilación de datos del estudio de la investigación original, lo cual limita las preguntas que pueden ser planteadas. Es importante señalar que la generalización puede ser limitada en los datos de los ensayos.

El advenimiento de la Historia Clínica Electrónica (HCE) ha resultado en la digitalización de las historias clínicas desde su formato anterior en papel. Los registros médicos digitalizados resultantes presentan oportunidades para superar algunas de las deficiencias de los datos administrativos, brindando datos detallados con resultados de laboratorio, medicaciones y el tiempo de los eventos clínicos [13]. Estas “grandes bases de datos” sacan provecho del hecho de que muchas HCE recogen datos de diversas fuentes como los monitores de los pacientes, los sistemas de laboratorio, y los sistemas de farmacia y los fusionan en un sistema unificado para los médicos. Esta información puede luego trasladarse a bases de datos deidentificadas para fines de investigación que contienen datos demográficos de los pacientes, información sobre facturación y procedimientos, datos de tiempo, de resultados hospitalarios, así como datos detallados a nivel del paciente y notas de los profesionales que se pueden buscar utilizando herramientas de procesamiento de lenguaje natural. El enfoque del “big data” puede atenuar la presencia de confusión al proveer la información detallada necesaria para evaluar la gravedad de las enfermedades (como resultados de laboratorio y signos vitales). Además, la naturaleza granular de los datos puede proporcionar una visión de la razón por la cual un paciente recibió una intervención y otro no, lo que puede en parte resolver los confundidores por indicación. Así pues, la promesa del “big data” es que contiene datos con un importante nivel de detalle. Las

bases de datos de “big data”, tales como MIMIC-III, tienen el potencial de expandir el alcance de lo que fue posible anteriormente con la investigación observacional.

9.7 Preparación

Menos del 10% de las decisiones clínicas están sustentadas en un alto nivel de evidencia [22]. Las preguntas clínicas surgen aproximadamente en cada uno de los pacientes y proporcionan una gran cantidad de preguntas de investigación. Cuando se formula una pregunta de investigación, los investigadores deben seleccionar cuidadosamente la muestra de sujetos, la variable de exposición, la variable resultado y las variables confundidoras. Una vez que la pregunta de investigación es clara, el diseño del estudio se convierte en el siguiente paso fundamental. Mientras que los ECA son el *gold standard* para establecer la inferencia causal bajo condiciones ideales, no siempre son prácticos, costo-efectivos, éticos o incluso posibles para algunos tipos de preguntas. La investigación observacional presenta una alternativa a la realización de ECA, pero a menudo se ve limitada en la inferencia causal por los confundidores no medidos.

Nuestro escenario clínico dio origen a la pregunta de si los CAI mejoraban los resultados de los pacientes que recibían VM. Esto traducido a una pregunta de investigación es: “Entre pacientes adultos de UCI con ventilación mecánica que no reciben medicación vasoactiva (muestra del estudio), ¿se asocia el uso de un CAI luego de iniciar la VM (exposición) con una mejora en las tasas de mortalidad a 28 días (*resultado*)?” Si bien un ECA podría responder esta pregunta, podría ser complicado logísticamente, costoso y difícil. Utilizando técnicas de efectividad comparativa, se puede plantear la pregunta usando una base de datos detallada retrospectiva, comparando los pacientes que recibieron un CAI con pacientes medibles similares que no tuvieron un CAI colocado. Sin embargo, debe prestarse cuidadosa atención a los confundidores de indicación no medidos, como por qué algunos pacientes recibieron un CAI y otros no. Factores tales como la gravedad de la enfermedad, la etiología de la falla respiratoria, y la presencia de ciertas patologías que pueden dificultar la colocación de un CAI (como la enfermedad arterial periférica) pueden ser considerados como posibles confundidores de la asociación entre CAI y mortalidad. Si bien

podría utilizarse una base de datos administrativa, esta podría carecer de información importante relacionada con posibles confundidores. En este sentido, bases de datos de HCE como MIMIC-III, con datos granulares detallados a nivel paciente, pueden permitir la medición de un mayor número de variables confundidoras previamente no medibles y permitir una mayor atenuación del sesgo en la investigación observacional.

Puntos clave

- La mayoría de las preguntas de investigación surgen de escenarios clínicos en los que el curso apropiado del tratamiento no es claro o es desconocido.
- La definición de la pregunta de investigación requiere una cuidadosa consideración de la mejor muestra de estudio, exposición y resultado con el fin de responder la pregunta clínica de interés.
- Si bien los estudios de investigación observacionales pueden superar muchas de las limitaciones de los ensayos controlados aleatorizados, es necesaria una cuidadosa consideración del diseño del estudio y selección de la base de datos para enfrentar los sesgos y factores de confusión.

Referencias

1. Esteban A, Frutos-Vivar F, Muriel A, Ferguson ND, Peñuelas O, Abraira V, Raymondos K, Rios F, Nin N, Apezteguía C, Violi DA, Thille AW, Brochard L, González M, Villagomez AJ, Hurtado J, Davies AR, Du B, Maggiore SM, Pelosi P, Soto L, Tomicic V, D'Empaire G, Matamis D, Abroug F, Moreno RP, Soares MA, Arabi Y, Sandi F, Jibaja M, Amin P, Koh Y, Kuiper MA, Bülow H-H, Zeggwagh AA, Anzueto

A (2013) Evolution of mortality over time in patients receiving mechanical ventilation. Am J Respir Crit Care Med 188 (2): 220-230.

2. Mehta A, Syeda SN, Wiener RS, Walkey AJ (2014) Temporal trends in invasive mechanical ventilation: severe sepsis/pneumonia, heart failure and chronic obstructive pulmonary disease. In: B23. Clinical trials and resultados, vol 271. American Thoracic Society, pp. A2537-A2537.
3. Stefan MS, Shieh M-S, Pekow PS, Rothberg MB, Steingrub JS, Lagu T, LindenauerPK (2013) Epidemiology and resultados of acute respiratory failure in the United States, 2001 to 2009: a national survey. J Hosp Med 8 (2): 76-82.
4. Traoré O, Liotier J, Souweine B (2005) Prospective study of arterial and central venous catheter colonization and of arterial– and central venous catheter-related bacteremia in intensive care units. Crit Care Med 33 (6): 1276-1280.
5. Gershengorn HB, Garland A, Kramer A, Scales DC, Rubenfeld G, Wunsch H (2014) Variation of arterial and central venous catheter use in United States intensive care units. Anesthesiology 120 (3): 650-664.
6. Gershengorn HB, Wunsch H, Scales DC, Zarychanski R, Rubenfeld G, Garland A (2014) Association between arterial catheter use and hospital mortality in intensive care units. JAMA Intern Med 174 (11): 1746-1754.
7. Maki DG, Kluger DM, Crnich CJ (2006) The risk of bloodstream infection in adults with different intravascular devices: a systematic review of 200 published prospective studies. Mayo Clin Proc 81 (9): 1159-1171.
8. Scheer BV, Perel A, Pfeiffer UJ (2002) Clinical review: complications and risk factors of peripheral arterial catheters used for haemodynamic monitoring in anaesthesia and intensive care medicine. Crit Care 6 (3): 199-204.
9. Concato J, Shah N, Horwitz RI (2000) Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 342 (25): 1887-1892.
10. Ho PM, Peterson PN, Masoudi FA (2008) Evaluating the evidence is there a rigid hierarchy? Circulation 118 (16): 1675-1684.
11. Smith GCS, Pell JP (2003) Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. BMJ 327 (7429): 1459-1461.
12. Booth CM, Tannock IF (2014) Randomised controlled trials and population-based observational research: partners in the evolution of medical evidence. Br J Cancer 110 (3): 551-555.
13. Scott DJ, Lee J, Silva I, Park S, Moody GB, Celi LA, Mark RG (2013) Accessing the public MIMIC-II intensive care relational database for clinical research. BMC Med Inform DecisMak 13 (1): 9.
14. Healthcare Cost and Utilization Project and Agency for Healthcare Research and Quality. Overview of the National (Nationwide) Inpatient Sample (NIS).

15. Doll R, Hill AB (1954) The mortality of doctors in relation to their smoking habits; a preliminary report. Br Med J 1 (4877): 1451-1455
16. Alberg AJ, Samet JM (2003) Epidemiology of lung cancer. Chest 123 (1 Suppl): 21S-49S.
17. Maldonado G, Greenland S (2002) Estimating causal effects. Int J Epidemiol 31 (2): 422-429.
18. Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB (1998) Prediction of coronary heart disease using risk factor categories. Circulation 97 (18): 1837-1847.
19. Zimmerman JE, Kramer AA, McNair DS, Malila FM (2006) Acute physiology and chronic health evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med 34 (5): 1297-1310.
20. Le Gall JR, Lemeshow S, Saulnier F (1993) A new simplified acute physiology score (SAPSII) based on a European/North American multicenter study. JAMA 270 (24): 2957-2963.
21. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, Reinhart CK, Suter PM, Thijs LG (1996) The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. On behalf of the working group on sepsis-related problems of the European society of intensive care medicine. Intensive Care Med 22 (7): 707-710.
22. Tricoci P, Allen JM, Kramer JM, Califf RM, Smith SC (2009). Scientific evidence underlying the ACC/AHA clinical practice guidelines. JAMA 301 (8): 831-841.
23. Del Fiol G, Workman TE, Gorman PN (2014) Clinical questions raised by clinicians at the point of care: a systematic review. JAMA Intern Med 174 (5): 710-718.

Capítulo 10

DEFINIENDO LA COHORTE DE PACIENTES

Ari Moskowitz y Kenneth Chen

Objetivos de Aprendizaje

- Comprender el proceso de selección de cohortes utilizando grandes bases de datos retrospectivas.
- Aprender habilidades adicionales específicas en la construcción de la cohorte, incluyendo visualización de datos y procesamiento del lenguaje natural (PLN).

10.1 Introducción

Un primer paso fundamental en cualquier estudio observacional es la selección de una cohorte de pacientes apropiada para el análisis. La importancia de invertir un tiempo y esfuerzo considerable en la selección de la población de estudio no puede subestimarse. El fracaso en la identificación de áreas de sesgos potenciales, confundidores, y datos faltantes en la fase inicial puede llevar a ineficiencias posteriores considerables. Además, hay que tener cuidado de seleccionar una población de pacientes adaptada a la pregunta de investigación de interés con el fin de aprovechar adecuadamente la gran cantidad de datos capturados por la Historia Clínica Electrónica (HCE).

En el siguiente capítulo nos centraremos en la selección de la cohorte de estudio. Específicamente, revisaremos los fundamentos del diseño de estudios observacionales con especial atención en los tipos de datos encontrados habitualmente en las HCE. Destacaremos las variables instrumentales frecuentemente utilizadas – son variables usadas para controlar los confundidores y los errores de medición en los estudios observacionales. Además, discutiremos cómo utilizar una combinación de técnicas basadas en datos y el razonamiento clínico en la selección de cohortes. El capítulo concluirá con una continuación del ejemplo trabajado en la parte uno de esta sección, donde discutiremos cómo la cohorte de pacientes fue seleccionada para el estudio de colocación de una vía arterial en la unidad de cuidados intensivos [1].

10.2 Parte 1 - Conceptos Teóricos

10.2.1 Exposición y Resultados de Interés

Estos conceptos son discutidos en detalle en el Capítulo 9 –"Formulando la Pregunta de Investigación". La minería de datos en la investigación biomédica utiliza un enfoque retrospectivo donde la exposición y el resultado de interés ocurren antes de la selección de los pacientes. Es realmente importante adaptar la exposición de interés buscada a la pregunta clínica en cuestión. Seleccionar una exposición demasiado amplia puede permitir una gran cohorte de pacientes, pero en detrimento de la precisión de los datos. De igual manera, ser muy específico en la elección de la exposición puede permitir tener más precisión pero a expensas del tamaño de la muestra y generalización.

La selección de una exposición de interés es el primer paso para determinar la cohorte de pacientes. En general, la exposición de interés puede ser considerada como centrada en el paciente, centrada en el episodio, o centrada en la interacción con el paciente. Esta terminología fue desarrollada por la empresa de almacenamiento de datos Health Catalyst para su herramienta Cohort Builder y proporciona un marco de trabajo razonable para identificar una exposición de interés. Las exposiciones centradas en el paciente se centran en las características intrínsecas de un grupo de pacientes. Esto puede incluir características demográficas (ej., género) o comorbilidades médicas (ej., diabetes). En contraste, las exposiciones centradas en el episodio son condiciones transitorias que requieren un curso de tratamiento discreto (ej., sepsis). Las exposiciones centradas en la interacción se refieren a una única intervención (ej., colocación de una vía arterial) [2]. Si bien las exposiciones centradas en la interacción suelen ser más simples de identificar, la elección de la exposición debería elegirse en función de la hipótesis específica de la investigación que estamos realizando.

El resultado de interés debería ser identificado a priori. El resultado debería relacionarse de forma natural con la exposición de interés y ser tan específico como sea posible para responder la pregunta clínica en cuestión. Se debe tener cuidado de evitar la identificación de falsas correlaciones que carecen de fundamentos fisiopatológicos (ver los ejemplos de falsas correlaciones mostradas en http://tylervigen.com). La relación buscada

debe basarse en la plausibilidad biológica. Las medidas de resultados amplias, como la mortalidad y la duración de la estancia pueden ser atractivas a priori pero, realmente, están influenciadas por demasiadas variables confusoras. La medición de resultados subrogantes (ej., cambio en la presión sanguínea, duración de la ventilación mecánica) puede ser de particular ayuda en la medida en que se relacionan más estrechamente a la exposición de interés y están menos afectados por confundidores.

Dado que las HCE no se orientan con frecuencia hacia el análisis y minería de datos, puede ser desafiante identificar una exposición de interés. Los datos numéricos estructurados, como resultados de laboratorio y signos vitales, son fáciles de buscar con técnicas estándar de consulta. El uso de los datos no estructurados tales como notas explicativas e informes radiológicos puede ser más difícil y requiere, a menudo, el uso de herramientas de PLN. Para seleccionar un fenotipo especifico de un paciente de un grupo heterogéneo y amplio de pacientes, puede ser útil aprovechar, tanto datos con formato estructurado como no estructurado.

Una vez que la exposición de interés es seleccionada, el investigador debe evaluar cómo utilizar un tipo de datos a o una combinación de estos para identificar la cohorte de estudio deseada para el análisis. Esto puede hacerse utilizando una combinación de técnicas *basadas en datos* y el razonamiento clínico, como se verá luego en el capítulo.

10.2.2 Grupo de Comparación

Además de identificar a los pacientes mapeándolos con la exposición de interés, el investigador también debe identificar un grupo de comparación. Idealmente, este grupo debe estar compuesto por pacientes fenotípicamente similares a aquellos en la cohorte de estudio, pero que carezcan de la exposición de interés. La cohorte de comparación seleccionada debe presentar el mismo riesgo de desarrollar el *resultado* de estudio. En investigación observacional, esto puede lograrse en particular a través del desarrollo del puntaje de propensión (Capítulo 23– “Análisis del Puntaje de Propensión”). En general, el grupo de comparación debería ser tan grande o más que la cohorte de estudio para maximizar la potencia del mismo. Es posible seleccionar demasiadas características con las que parear las cohortes de estudio con las de comparación, lo que reduce el número de

pacientes disponibles para la cohorte de comparación. Es necesario tener cuidado de prevenir el sobrepareamiento.

En casos seleccionados, los investigadores pueden sacar provecho de experimentos naturales en los que circunstancias externas a la HCE establecen fácilmente una cohorte de estudio y un grupo de comparación. Las llamadas "variables instrumentales" pueden incluir variaciones en la práctica entre unidades de cuidado, hospitales e incluso regiones geográficas. Las relaciones temporales (ej., antes y después) relativas a iniciativas de mejora de la calidad o publicaciones de guías de expertos pueden también ser aprovechadas como variables instrumentales. Los investigadores deberían estar atentos a estas herramientas tan útiles.

10.2.3 Construcción de la Cohorte de Estudio

La identificación de fenotipos específicos de pacientes para incluir en las cohortes de estudio y comparación requiere una combinación de razonamiento clínico y técnicas *basadas en datos.* Es esencial una colaboración estrecha entre médicos y científicos de datos para la selección de pacientes en los estudios donde se utilizan datos de las HCE.

El médico está en la primera línea de la atención y tiene contacto directo con escenarios clínicos complejos que existen fuera del ámbito de la medicina basada en la evidencia. Según un reporte del año 2011 del comité del Instituto de Medicina, sólo el 10-20% de las decisiones clínicas se basan en la evidencia [3]. Cerca del 50% de las guías de práctica clínica se basan en la opinión de expertos antes que en datos experimentales [4]. En este "desierto de datos" es función del médico identificar nuevas preguntas de investigación importantes para orientar la atención médica [5]. Estas preguntas llevan naturalmente a la identificación de una exposición de interés.

Una vez que la pregunta clínica y la exposición de interés han sido identificadas, el médico y el científico de datos necesitarán establecer una cohorte de pacientes. La consulta de fenotipos de datos estructurados y no estructurados puede ser compleja y requiere, con frecuencia, un ajuste en el criterio de búsqueda. A menudo se requieren consultas múltiples y complementarias con el fin de identificar el grupo específico de interés. Además, el equipo de investigación debe considerar la "singularidad" del

paciente ya que algunos pacientes tienen múltiples admisiones en la UCI tanto durante una única hospitalización como en visitas repetidas al hospital. Si el mismo paciente está incluido más de una vez en una cohorte de estudio, se pierde la asunción de mediciones independientes.

Los investigadores deben prestar atención a la necesidad de excluir algunos pacientes por razón de su historial médico o estado patológico, como el embarazo por ejemplo. En caso contrario, podrían introducirse factores confundidores y alterar la relación causal de interés.

En un ejemplo de un estudio publicado de MIMIC-II, los investigadores intentaron determinar si el uso de inhibidores de la bomba de protones (IBP) se asociaba con hipomagnesemia en pacientes graves en la UCI [6]. La exposición de interés en este estudio era 'el uso de IBP'. Se identificaron un grupo de comparación de pacientes expuestos a un agente antiácido alternativo (antagonistas del receptor de histamina-2) y un grupo de comparación que no recibía ninguna medicación antiácida. El resultado de interés fue un bajo nivel de magnesio. Para determinar la cohorte de estudio en este caso, se tuvieron que desarrollar consultas para identificar:

1. Primera admisión en UCI para cada paciente
2. Uso de IBP identificada a través del análisis de PLN de la sección "Medicamentos" de la historia de ingreso y del examen físico
3. Patologías que puedan influir en el uso de IBP y/o los niveles de magnesio (ej., enfermedad diarreica, etapa final de enfermedad renal)
4. Pacientes que fueron transferidos desde otros hospitales dado que podría no contarse con la información de la medicación recibida (pacientes excluidos)
5. Pacientes que no tuvieron mediciones de magnesio dentro de las 36 hs de la admisión en UCI (pacientes excluidos)
6. Pacientes con datos de comorbilidad faltantes (pacientes excluidos)
7. Potenciales confundidores incluyendo el uso de diuréticos

Las consultas SQL correspondientes a este ejemplo se proporcionan bajo el nombre *"SQL_cohort_selection".*

Maximizar la eficiencia de las consultas de datos a partir de las HCE es un área de activa investigación y desarrollo. Por ejemplo, la red i2b2 (Informatics *for Integrating Biology and the Bedside)* es un programa financiado por NIH con sede en *Partner's Health Center* (Boston, MA) que

está desarrollando un marco de trabajo para simplificar la búsqueda y extracción de datos de las HCE. Las herramientas de software desarrolladas por i2b2 son de descarga gratuita y prometen simplificar la identificación de un fenotipo clínico a partir de datos en bruto de la HCE https://www.i2b2.org/about/index.html. Este proyecto y otros similares deberían colaborar a simplificar el gran número de consultas necesarias para desarrollar una cohorte de estudio [7].

10.2.4 Exposiciones Ocultas

No todas las exposiciones de interés pueden identificarse directamente a partir de los datos incluidos en las HCEs. En estas circunstancias, los investigadores deben ser creativos para identificar puntos de datos registrados que sigan de cerca la exposición de interés. El razonamiento clínico es importante en estas circunstancias.

Por ejemplo, un equipo de investigación utilizando la base de datos MIMIC II seleccionó como exposición de interés 'fibrilación auricular con respuesta ventricular rápida recibiendo un fármaco de control de frecuencia'. La fibrilación auricular es una taquiarritmia común en poblaciones de enfermos críticos que ha sido asociada con peores resultados clínicos. La fibrilación auricular con respuesta ventricular rápida es a menudo tratada con uno de los tres agentes de control de frecuencia: metoprolol, diltiazem o amiodarona. Desafortunadamente, 'la fibrilación auricular con respuesta ventricular rápida' no es una variable estructurada en el sistema de HCE conectada a la base de datos MIMIC II. Realizar una búsqueda de PLN para el término 'fibrilación auricular con respuesta ventricular rápida' en las notas clínicas y los resúmenes de alta es factible, sin embargo no proporcionaría la resolución temporal necesaria respecto a la administración de la medicación.

Para superar este obstáculo, los investigadores generaron un algoritmo para identificar indirectamente la exposición oculta. Se desarrolló una consulta para identificar la primera dosis de un agente de control de frecuencia intravenoso (metoprolol, diltiazem, o amiodarona) recibida por un único paciente en la UCI. A continuación, se determinó si la frecuencia cardíaca del paciente dentro del intervalo de una hora de la administración de medicación era mayor a 110 latidos por minuto. Finalmente, se usó un

algoritmo de PLN para buscar en la ficha clínica la mención de fibrilación auricular. Aquellos pacientes que reunieron las tres condiciones fueron incluidos en la cohorte del estudio final. Se proporcionan ejemplos del código Matlab utilizado para identificar la cohorte de interés (función "Afib"), así como el código Perl para PNL (función "NLP").

10.2.5 Visualización de Datos

La representación gráfica de los datos alfanuméricos de la HCE puede ser de particular ayuda para establecer la cohorte de estudio. La visualización de datos de la HCE los hace más accesibles y permite la rápida identificación de tendencias que de otra manera serían difíciles de identificar. Además fomenta una comunicación más efectiva tanto entre los miembros del equipo como entre el equipo de investigación y el público en general que no está acostumbrado a la investigación de 'Big Data'. Estos principios se discuten más ampliamente en el Capítulo 15 de este libro de texto "Análisis Exploratorio de Datos".

En el proyecto mencionado anteriormente que exploraba el uso de agentes de control de la frecuencia para la fibrilación auricular con respuesta ventricular rápida, un *resultado* de interés fue el tiempo hasta el control de la respuesta ventricular rápida. Desafortunadamente, la literatura existente no provee una orientación específica en este campo. Utilizando la visualización de datos, se llegó a un consenso de que el control del ritmo debería definirse como una frecuencia cardiaca<110 latidos por minuto por al menos el 90% del tiempo en un período de 4 horas. Aunque algunos aspectos de esta definición son arbitrarios, la visualización de datos permitió a todos los miembros del equipo llegar a un acuerdo acerca de qué definición era la más defendible estadística y clínicamente.

10.2.6 Fidelidad de la Cohorte de Estudio

Los algoritmos de consulta en general no pueden presumir de un 100% de exactitud para identificar al fenotipo de paciente buscado. Se esperan falsos positivos y falsos negativos. Con el objetivo de garantizar la fiabilidad de la cohorte de estudio, puede ser de ayuda revisar manualmente un subconjunto aleatorio de pacientes seleccionados. Basado en el tamaño de la cohorte de estudio, el 5-10% de las gráficas clínicas deberían revisarse

para asegurar la presencia o ausencia de la exposición de interés. Esta tarea debe ser realizada por un médico. Si los recursos lo permiten, se puede asignar esta tarea a dos médicos revisores y comparar sus resultados independientes usando el estadístico Kappa.

Finalmente, los investigadores pueden usar el *"gold standard"* de la revisión manual realizada para establecer una Característica Operativa del Receptor (ROC, del inglés Receiver Operating Characteristic). Un área bajo la curva ROC que es >0.80 indica una 'buena' precisión del algoritmo y debería utilizarse como un mínimo absoluto de fidelidad de algoritmo. Si el área bajo la curva ROC es <0.80, debería utilizarse una combinación de técnicas de visualización de datos y razonamiento clínico para ajustar el algoritmo de la consulta a la exposición de interés.

10.3 Parte 2 - Caso de estudio: Selección de la Cohorte

En el caso de estudio presentado, los autores analizaron el efecto de los catéteres arteriales invasivos (CAI) en pacientes hemodinámicamente estables con falla respiratoria usando datos multivariados. Identificaron a la 'colocación del catéter arterial' como su exposición de interés del tipo basado en la interacción. Los CAI son ampliamente utilizados en la unidad de cuidados intensivos para medir la presión arterial latido a latido y se cree que son más precisos y fiables que el monitoreo estándar no invasivo de la presión sanguínea. También tienen el beneficio adicional de permitir la extracción de gasometrías sanguíneas más sencilla que puede reducir la necesidad de punciones arteriales repetidas. Dada su naturaleza invasiva, sin embargo, los CAI acarrean riesgos de infección del torrente sanguíneo y daño vascular. El *resultado* primario de interés seleccionado fue la mortalidad a 28 días con *resultados* secundarios que incluyeron la duración de la estancia hospitalaria y en UCI, la duración de la ventilación mecánica, y el número medio de mediciones de gases sanguíneos realizadas.

Los autores eligieron centrar su estudio en pacientes que requerían ventilación mecánica, sin necesidad de vasopresores y que no fueron ingresados por sepsis. En pacientes que requieren ventilación mecánica, se considera particularmente importante la doble función de los CAI de permitir el monitoreo de la presión sanguínea latido a latido y simplificar la extracción de gases en sangre arterial. Los pacientes con vasopresores y/o

sepsis fueron excluidos ya que los catéteres arteriales son necesarios en esta población para facilitar la rápida titulación de agentes vasoactivos. Además, sería difícil identificar una cantidad suficiente de pacientes que requieran vasopresores o admitidos por sepsis, que no reciban un CAI.

Los autores comenzaron la selección de su cohorte con los 24.581 pacientes incluidos en la base de datos MIMIC II. Para los pacientes con múltiples ingresos en UCI, solo se utilizó la primera admisión en UCI para asegurar la independencia de las mediciones. La función *"cohort1"* contiene la consulta SQL correspondiente a este paso. Luego, se identificaron los pacientes que requirieron ventilación mecánica dentro de las primeras 24 horas de su admisión en UCI y recibieron ventilación mecánica por al menos 24 horas (función *"cohort2"*). Después de identificar la cohorte de pacientes que requirieron ventilación mecánica, los autores consultaron por la colocación de un CAI después del inicio de la ventilación mecánica (función *"cohort3"*). Como la mayoría de los pacientes en la unidad de recuperación de cirugía cardiovascular tenían un CAI colocado antes de la admisión a UCI, todos los pacientes de la UCI cardiovascular fueron excluidos del análisis (función *"cohort4"*). Con el fin de excluir a los pacientes admitidos en la UCI con sepsis, los autores utilizaron los criterios de Angus (función *"cohort5"*). Finalmente, los pacientes que requirieron vasopresores durante su admisión en UCI fueron excluidos (función *"cohort6"*).

Se identificó el grupo de comparación de pacientes que recibieron ventilación mecánica de al menos 24 horas dentro de las primeras 24 horas de su ingreso a UCI pero que no tuvieron colocado un CAI. Finalmente, hubo 984 pacientes en el grupo que recibió un CAI y 792 pacientes que no. Estos grupos fueron comparados utilizando técnicas de pareamiento de propensión descritas en el Capítulo 23 – "Análisis de Puntaje de Propensión".

En última instancia, esta cohorte consiste en identificadores exclusivos de pacientes que cumplen con los criterios de inclusión. Otros investigadores pueden estar interesados en acceder a esta cohorte en particular con el fin de replicar los resultados del estudio o abordar otras preguntas de investigación. El sitio web MIMIC proporcionará a los investigadores, en el futuro, la posibilidad de compartir cohortes de pacientes, permitiendo por ende a los equipos de investigación interactuar y construir sobre el trabajo de otros grupos.

Puntos clave

- Dedicar tiempo a caracterizar la exposición y los *resultados* de interés antes del estudio.
- Utilizar datos estructurados y no estructurados para determinar su exposición y *resultado* de interés. La PLN puede ser de particular ayuda en el análisis de datos no estructurados.
- La visualización de datos puede ser de mucha ayuda para facilitar la comunicación entre los miembros del equipo.

Referencias

1. Hsu DJ, Feng M, Kothari R, Zhou H, Chen KP, Celi LA (2015) The association between indwelling arterial catheters and mortality in hemodynamically stable patients with respiratory failure: a propensity score analysis. Chest 148 (6): 1470-1476.
2. Merkley K (2013) Defining patient populations using analytical tools: cohort builder and risk stratification. Health Catalyst, 21 Aug 2013.
3. Institute of Medicine (US) Committee on Standards for Developing Trustworthy Clinical Practice Guidelines (2011) Clinical practice guidelines we can trust. National Academies Press (US), Washington (DC).
4. Committee on the Learning Health Care System in America and Institute of Medicine (2013) Best care at lower cost: the path to continuously learning health care in America. National Academies Press (US), Washington (DC).
5. Moskowitz A, McSparron J, Stone DJ, Celi LA (2015) Preparing a new generation of clinicians for the era of big data. Harv Med Stud Rev 2 (1): 24-27.

6. Danziger J, William JH, Scott DJ, Lee J, Lehman L, Mark RG, Howell MD, Celi LA, Mukamal KJ (2013) Proton-pump inhibitor use is associated with low serum magnesium concentrations. Kidney Int 83 (4): 692-699.
7. Jensen PB, Jensen LJ, Brunak S (2012) Mining electronic health records: towards better research applications and clinical care. NatRevGenet 13 (6): 395-405.

Capítulo 11

PREPARACIÓN DE LOS DATOS

Tom Pollard, Franck Dernoncourt,
Samuel Finlayson y Adrián Velasquez

Objetivos de aprendizaje

- Familiarizarse con las categorías habituales de datos médicos
- Apreciar la importancia de la colaboración entre el personal de salud y los analistas de datos
- Aprender la terminología común asociada con las bases de datos relacionales y archivos de texto plano
- Entender los conceptos claves de la investigación reproducible
- Obtener experiencia práctica en consultar una base de datos médica

11.1 Introducción

Los datos son el centro de toda investigación, es por ello que son importantes las prácticas robustas en la gestión de datos, para que los estudios sean llevados a cabo de manera eficiente y segura. Lo mismo se puede decir de la gestión del software utilizado para procesar y analizar datos. Garantizar que se apliquen buenas prácticas al comienzo de un estudio, puede resultar en un ahorro significativo, en términos de tiempo y esfuerzo en una etapa posterior.

Mientras que hay beneficios bien reconocidos en herramientas y prácticas como el control de versiones, marcos de prueba, y flujos de trabajo reproducibles, aún existe un importante camino previo a recorrer, antes de que esto sea adoptado ampliamente por la comunidad académica. En este capítulo discutiremos algunos aspectos claves a considerar cuando se trabaja con datos médicos y resaltar algunos enfoques que pueden hacer que los estudios sean colaborativos y reproducibles.

11.2 Parte 1 - Conceptos Teóricos

11.2.1 Categorías de Datos Hospitalarios

Los datos son obtenidos en forma rutinaria a través de diferentes fuentes disponibles en los hospitales y generalmente son optimizados para apoyar

las actividades clínicas y de facturación en lugar de la investigación. Las categorías de datos que comúnmente se encuentran en la práctica se resumen en la Tabla 11.1 y se discuten a continuación.

- Los datos de facturación generalmente consisten en los códigos que los hospitales y el personal de salud usa para presentar reclamos con sus proveedores de seguros. Los dos sistemas de codificación más utilizados son: la Clasificación Estadística Internacional de Enfermedades y Problemas de Salud Relacionados, comúnmente abreviada como la Clasificación Internacional de Enfermedades (CIE) que es sostenida por la Organización Mundial de la Salud y los códigos de la Terminología Actualizada de Procedimientos Médicos (CPT, por sus siglas en inglés, Current Procedural Terminology), sostenida por la Asociación Americana de Medicina. Estas terminologías jerárquicas fueron diseñadas para proporcionar una estandarización a la clasificación y el reporte médico.

Tabla 11.1 Categorías comunes de datos hospitalarios y problemas comunes a considerar en su análisis

Categorías	**Ejemplos**	**Problemas comunes a considerar**
Demografía	Edad, sexo, raza, altura, peso	Datos altamente sensibles que requieren una deidentificación cuidadosa. Los datos existentes en campos tales como raza pueden ser pocos.
Laboratorio	Creatinina, lactato, recuento de glóbulos blancos, resultados de microbiología.	A menudo no hay mediciones acerca de la calidad de las muestras. Los métodos y reactivos utilizados en las pruebas pueden variar de una unidad a la otra y a través del tiempo.
Imágenes radiográficas e	Rayos-X, Tomografías computadas (TC), ecocardiogramas	Información protegida de la salud de los pacientes, como sus nombres, pueden ser

informes asociados		escritos en las imágenes. Las plantillas utilizadas para generar informes pueden influir en su contenido.
Datos fisiológicos	Signos vitales, datos de sensores de electrocardiografía (ECG), datos de sensores de electroencefalografía (EEG)	Los datos pueden ser pre procesados por los propietarios de los algoritmos. Las etiquetas pueden no ser precisas (por ejemplo, la “glucosa por punción digital” puede estar tomada de sangre venosa)
Medicaciones	Prescripciones, dosis, duración y hora de las dosis	Pueden ser listadas medicaciones que fueron prescriptas pero que no fueron dadas. Las marcas de tiempo pueden describir puntos de orden y no de administración.
Diagnósticos y códigos de procedimiento	Códigos de la Clasificación Internacional de Enfermedades (CIE), Códigos del Grupo Relacionado con el Diagnóstico (GRD), códigos de la Terminología Actualizada de Procedimientos (CPT)	Frecuentemente basadas en una revisión retrospectiva de notas. No se busca que indiquen el estado médico del paciente. Sujeto a sesgos del codificador. Limitado por la adecuación de los códigos.
Notas del personal de salud y de procedimiento	Notas de ingreso, notas de evolución diaria, epicrisis reportes operativos	Errores tipográficos. El contexto es importante (por ejemplo, las enfermedades pueden aparecer al describir las historias clínicas familiares), las abreviaciones y acrónimos son comunes.

- El trazado de datos fisiológicos, incluyendo la información como la frecuencia cardíaca, presión arterial y frecuencia respiratoria son recolectadas al pie de la cama del paciente. La frecuencia y la duración del monitoreo generalmente está relacionado con el nivel del cuidado. Los datos en general son almacenados a una menor velocidad de la que

se muestrean (por ejemplo, cada 5-10 min) usando algoritmos calculados que con frecuencia son patentados y no se encuentran disponibles.

- Las notas y reportes, creados para registrar la evolución del paciente, resumen la estadía de un paciente hasta el alta y proveen hallazgos de estudios por imagen como rayos-X y ecocardiogramas. Mientras que los campos son de texto libre, las notas usualmente son creadas con la ayuda de plantillas estandarizadas, por lo cual es probable que se encuentren parcialmente estructuradas.
- Imágenes, como las obtenidas por rayos X, tomografía axial computada (TAC), ecocardiografía y resonancia magnética.
- Datos de laboratorio y medicaciones. Las órdenes de medicaciones y de estudios de laboratorio son ingresadas por el personal de salud en un sistema de prescripciones médicas, que luego son efectuadas por el personal de enfermería o del laboratorio. Dependiendo del sistema, algunas marcas en el tiempo pueden referirse al momento en que el médico realizó la orden y otras pueden referirse a cuando la droga fue administrada o cuando fueron reportados los resultados de laboratorio. Algunas medicaciones quizás fueron administradas días o semanas posteriores a la primera prescripción mientras que algunas pudieron no haber sido administradas.

11.2.2 Contexto y Colaboración

Uno de los mayores desafíos de trabajar con datos médicos es adquirir el conocimiento del contexto en que los datos son obtenidos. Por esta razón, es muy importante enfatizar de la importancia que tiene la colaboración entre el plantel del hospital y los analistas de datos de la investigación. Algunos ejemplos de los aspectos a considerar cuando se trabaja con datos médicos están descriptos en la Tabla 11.1 y son discutidos a continuación:

- Los códigos de facturación no están destinados a documentar el estado de salud o el tratamiento, de un paciente, desde una perspectiva clínica y es posible que no sean fiables [3]. Las prácticas de codificación pueden estar condicionadas, deliberadamente o no, por diversos factores como la compensación financiera y trámites administrativos asociados.
- Las marcas en el tiempo pueden tener distinto significado para las diferentes categorías de datos. Por ejemplo, las marcas de tiempo pueden indicar el punto en que la medición fue realizada, cuando la

medición fue ingresada al sistema, cuando la muestra fue tomada, o cuando los resultados fueron informados por el laboratorio.

- Las abreviaturas y las palabras mal escritas aparecen frecuentemente en los campos de texto libre. La abreviatura "pad" en inglés, por ejemplo, puede referirse a *"peripheral artery disease"* (enfermedad arterial periférica) o "absorptive bed pad " (protector absorbente para colchón), o "diaper pad" (pañal). Además, las notas mencionan frecuentemente enfermedades que son halladas en la historia familiar del paciente, pero no necesariamente en el paciente, por ese motivo se debe tener cuidado cuando se utilizan búsquedas de texto simple.
- Las etiquetas que describen conceptos pueden no ser precisas. Por ejemplo, durante las investigaciones preliminares sobre un estudio no publicado para evaluar la precisión en la medición de glucemia en sangre capilar, se encontró que el personal de salud tomaría mediciones de glucemia capilar usando sangre venosa, si era fácilmente accesible, para evitar pinchar el dedo de un paciente.

Cada hospital tiene su propio sesgo de datos. Estos sesgos pueden estar relacionados con factores como el tipo de población de pacientes atendidos, las prácticas locales del personal de salud u otro tipo de servicios brindados. Por ejemplo:

- Los centros académicos a menudo atienden pacientes más complicados y algunos hospitales podrían tener una tendencia a atender pacientes de un origen étnico o un estatus socioeconómico específico.
- Las consultas de seguimiento pueden ser menos comunes en los centros de referencia y debido a ello puede ser menos probable detectar complicaciones a largo plazo.
- Los centros de investigación probablemente sean más propensos a administrar a sus pacientes drogas experimentales generalmente no usadas en la práctica.

11.2.3 Datos Cuantitativos y Cualitativos

Los datos son descritos a menudo como cuantitativos o cualitativos. Los datos cuantitativos son datos que pueden ser medidos, escritos con números y manipulados numéricamente. Los datos cuantitativos pueden ser discretos, teniendo solamente ciertos valores (por ejemplo, los números enteros 1, 2, 3), o continuos, tomando cualquier valor (por ejemplo, 1.23, 2.59). El

número de veces que un paciente es ingresado a un hospital es discreto (un paciente no puede ser admitido 0.7 veces), mientras que el peso de un paciente es continuo (el peso de un paciente puede tener cualquier valor dentro de un rango).

Los datos cualitativos son información que no puede ser expresada como un número y son a menudo usados con el término intercambiable de datos "categóricos". Cuando no hay un orden natural de las categorías (por ejemplo, la etnia de un paciente), los datos son llamados nominales. Cuando las categorías pueden ser ordenadas, estas son llamadas variables ordinales (por ejemplo, la intensidad de un dolor en una escala). Cada valor posible de una variable categórica es conocido comúnmente como un nivel.

11.2.4 Bases de Datos y Archivos de Datos

Los datos normalmente se hallan disponibles a través de una base de datos o de un archivo que pudo haber sido exportado de una base de datos. Mientras que hay diferentes tipos de bases de datos y archivos de datos, las bases de datos relacionales y los archivos de valores separados por comas (CSV) son probablemente los más usados.

Archivos de valores separados por comas (CSV)

Los archivos de valores separados por comas (CSV, por sus siglas en inglés comma separated values) son un formato de texto plano usado para el almacenamiento de datos en una estructura tabular, en formato de hoja de cálculo. Mientras que no hay una regla dura y rápida para estructurar datos tabulares, usualmente se considera de buena práctica incluir una fila de encabezado, para enlistar cada variable en columnas separadas y enlistar observaciones en filas [4].

Como no hay un estándar oficial para el formato CSV, se usa el término de manera vaga, lo que puede a veces ocasionar problemas cuando se busca cargar los datos dentro de un paquete de análisis de datos. Una recomendación general es seguir la definición de CSVs establecida por la Internet Engineering Task Force en el documento de especificación RFC 4180 [5]. Resumido brevemente, RFC 4180 especifica que:

- Los archivos pueden comenzar opcionalmente con una fila de encabezado, con cada campo separado por una coma.

- Los registros deberían listarse en filas contiguas. Los campos deberían separarse por comas, y cada fila debería terminar con un salto de línea.
- Los campos que contienen números pueden ser opcionalmente incluidos con comillas dobles.
- Los campos que contienen texto (**“cadenas”**) deberían incluirse con comillas dobles.
- Si las comillas dobles aparecen dentro de una cadena de texto, estas deben ser evitadas con una comilla doble precediéndolas.

El formato CSV es popular en gran medida por su simplicidad y versatilidad. Los archivos CVS pueden ser editados con un editor de texto, cargados en una hoja de cálculos en paquetes como Microsoft Excel, importados y procesados por la mayoría de los paquetes de análisis de datos. Usualmente los archivos CSV son un formato de datos intermedio usado para manejar datos que fueron extraídos de una base de datos relacional en preparación para analizar. La Figura 11.1 muestra un ejemplo de un archivo CSV formateado con las especificaciones del documento RFC 4180.

Fig. 11.1 Archivo de valores separados por comas (CSV) formateado con las especificaciones RFC 4180.

Base de datos relacionales

Hay varios estilos de bases de datos actualmente en uso, pero probablemente la más ampliamente implementadas son las”bases de datos relacionales”. Las bases de datos relacionales pueden ser pensadas como una colección de tablas que están conectadas entre sí mediante claves compartidas. La organización de datos a través de tablas puede ayudar a

mantener la integridad de los datos y permitir un análisis rápido y un almacenamiento más eficiente.

El modelo que define la estructura y las relaciones de las tablas es conocido como un “esquema de base de datos”. Damos un ejemplo simple de una base de datos de un hospital, con cuatro tablas, esta podría estar comprendida de: Tabla 1, una lista de todos los pacientes; Tabla 2, un registro de los ingresos del hospital; Tabla 3, una lista de mediciones de signos vitales; Tabla 4, un diccionario de códigos de signos vitales y etiquetas asociadas. La Figura 11.2 demuestra como las tablas pueden ser unidas con claves primarias y externas. Brevemente, una clave primaria es un identificador único dentro de una tabla. Por ejemplo, el ***subject-id*** es la clave primaria en la tabla ***pacientes***, porque cada paciente es enumerado solo una vez. Una clave externa en una tabla se refiere a una clave primaria en otra tabla. Por ejemplo, el ***subject-id*** en la tabla de ***ingresos*** es una clave externa, porque referencia la clave primaria en la tabla ***pacientes***.

La extracción de datos desde una base de datos es conocido como “consultar” la base de datos (“querying” en inglés). El lenguaje de programación comúnmente usado para crear una consulta es conocido como “Lenguaje de Consulta Estructurado” o SQL, por sus siglas en inglés. Mientras la sintaxis de SQL es directa, a veces es desafiante elaborar las consultas debido al razonamiento conceptual requerido para unir los datos a través de múltiples tablas.

Hay diferentes sistemas de base de datos de uso habitual. Algunos de esos sistemas como la base de datos Oracle o el Servidor de Microsoft SQL son comerciales y pueden tener costos de licencia. Otros sistemas como PostgreSQL y MySQL son de código abierto y de instalación gratuita. Si bien, el principio general detrás de las bases de datos es el mismo, es útil saber que la sintaxis de programación varía levemente entre sistemas.

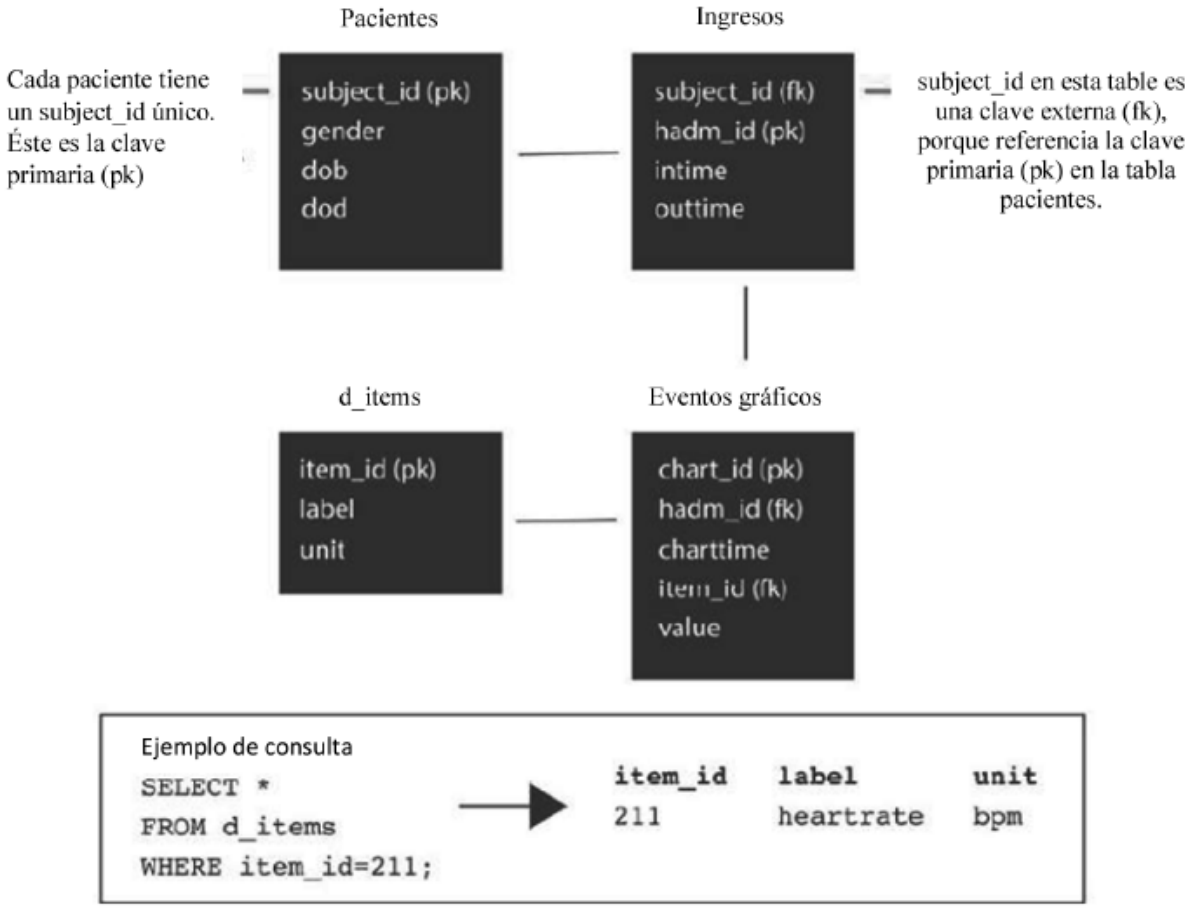

Fig. 11.2 Las bases de datos relacionales consisten en múltiples datos unidos por claves primarias y externas. La tabla de pacientes lista los pacientes únicos. La tabla de admisiones lista admisiones hospitalarias únicas. La tabla de eventos gráficos lista eventos como las mediciones de la frecuencia cardiaca. La tabla d-items es un diccionario que lista ítem-ids y etiquetas asociadas, como se mostró en el ejemplo buscado. p*k* es una clave primaria y. f*k* es una clave externa.

11.2.5 Reproducibilidad

Al mismo tiempo que un sistema de publicación que enfatiza la interpretación de resultados sobre la metodología detallada, los investigadores están bajo la presión de entregar regularmente artículos científicos de "alto impacto" para mantener sus carreras. Este ambiente puede contribuir a la "crisis de reproducibilidad" ampliamente reportada en la ciencia actual [6, 7].

Nuestra respuesta debería ser asegurar que los estudios son, en la medida de lo posible, reproducibles. Haciendo los datos y códigos accesibles,

podemos detectar y reparar más fácilmente errores inevitables, ayudar al otro a aprender de nuestros métodos, y promover mejor calidad de investigación.

Cuando se practica investigación reproducible, la fuente de datos no debería ser modificada. La edición de los datos crudos destruye la cadena de reproducibilidad. En cambio, se utilizan los códigos para procesar los datos de forma que todos los pasos que hacen un análisis desde la fuente al resultado puedan ser reproducidos.

Los códigos y los datos deben ser bien documentados y deben ser claras las condiciones de su utilización. Es típico proveer un archivo de texto plano "README" que da una introducción al paquete de análisis, junto con un archivo llamado "LICENSE" describiendo los condiciones de uso. Distintas herramientas como Júpiter Notebook, Sweave y Knitr pueden ser usados para introducir códigos y texto para producir documentos claros, estudios reproducibles por lo que están volviéndose cada vez más populares en la comunidad de investigadores (Fig. 11.3).

Los sistemas de control de versiones como Git pueden ser usados para seguir los cambios hechos en el código a lo largo del tiempo y también están volviéndose una herramienta cada vez más popular para los investigadores [8]. Cuando se trabaja con un sistema de control de versiones, una bitácora o *commit log,* provee un registro de los cambios al código por contribuidor, dando transparencia al proceso de desarrollo y actuando como una herramienta útil para descubrir y reparar errores.

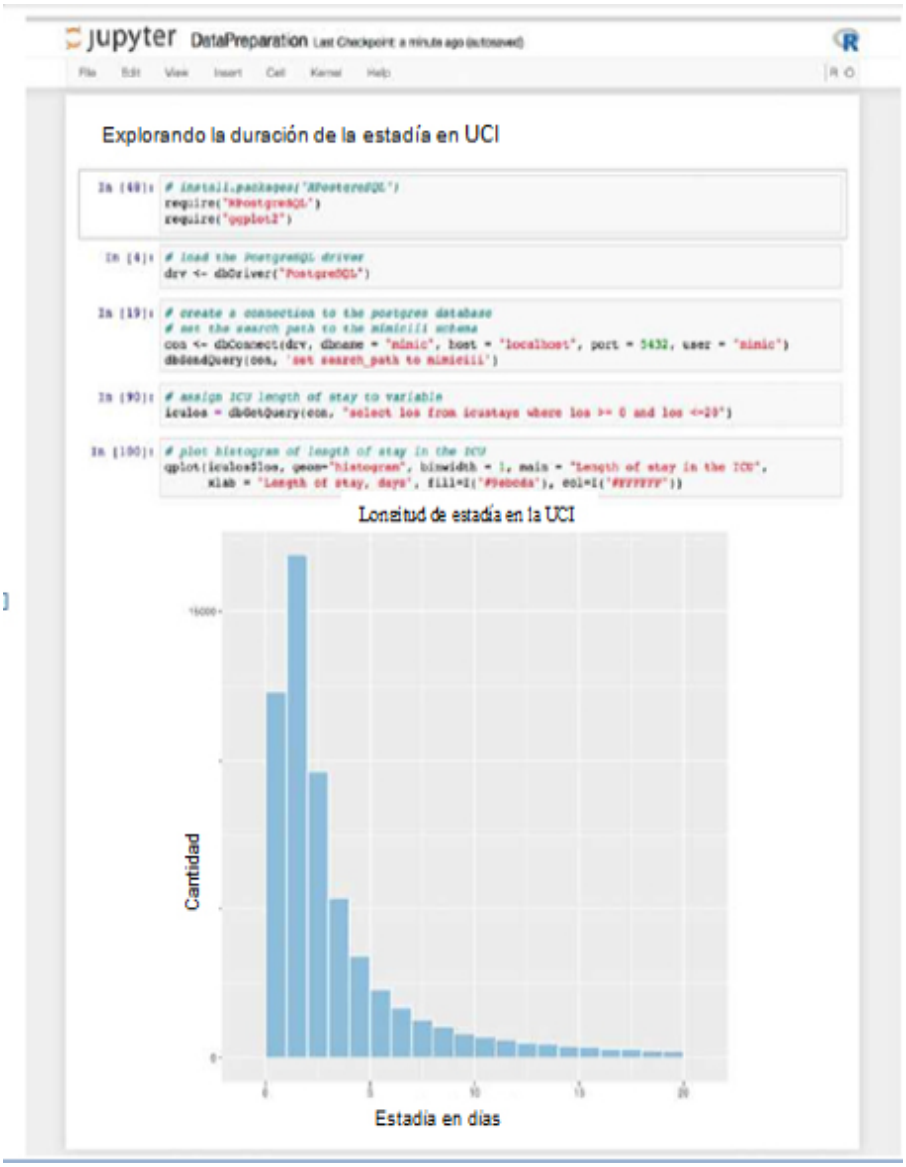

Fig. 11.3 Las Notebooks de Jupyter permiten la combinación de la documentación y el código en un análisis reproducible. En este ejemplo, la duración de la estadía en la UCI es cargada desde la base de datos MIMIC-III (v1.3) y graficada en un histograma [11].

La colaboración también es facilitada por el sistema de control de versiones. Git provee una gran funcionalidad que facilita la distribución del código y permite a múltiples personas trabajar conjuntamente en sincronía. La integración con servicios de hosting Git como Github proveen un mecanismo simple para obtener información, ayudando a minimizar el riesgo de la pérdida de datos, y también provee herramientas para el seguimiento de problemas y tareas [8, 9].

11.3 Parte 2 - Ejemplos prácticos de preparación de datos

11.3.1 Tablas de la base de datos MIMIC

Con el fin de llevar a cabo el estudio del efecto de los catéteres arteriales invasivos como fue descripto en el capítulo anterior, se usarán las siguientes tablas de la base de datos clínicos MIMIC-III:

- La tabla ***chartevents***, la tabla más larga de la base de datos. Contiene todos los datos graficados por el sistema de monitores de cuidados intensivos de cada cama, incluyendo registros fisiológicos como la frecuencia cardiaca y la presión arterial, como también la configuración utilizada por los catéteres arteriales invasivos.
- La tabla ***patients***, que contiene detalles demográficos de cada paciente ingresado en la unidad de cuidados intensivos, como también el género, fecha de nacimiento, y fecha de defunción.
- La tabla ***icustays***, que contiene detalles administrativos relacionados a la estadía en la UCI, como también el momento de admisión, momento del alta, y tipo de unidad de cuidado.

Antes de continuar con el siguiente ejercicio, recomendamos familiarizarse con la documentación de MIMIC y particularmente con las descripciones de la tabla, las cuales están disponibles en el sitio web de MIMIC [10].

11.3.2 Conceptos básicos de SQL

Una consulta SQL tiene el siguiente formato:

```
SELECT [columns]
FROM [table_name]
WHERE [conditions];
```

El resultado devuelto por la consulta es una lista de filas. La siguiente consulta lista el identificador de pacientes único (**subject-ids**) de todos los pacientes femeninos.

```
SELECT subject_id
FROM patients
WHERE gender = 'F';

-- returns:
 subject_id
------------
        654
        655
        656
        ...
```

Usualmente necesitamos especificar más de una condición. Por ejemplo, la siguiente consulta lista el **subject: ids** cuya primera o última unidad de cuidados fue una unidad coronaria (UCO):

```
SELECT subject_id
FROM icustays
WHERE first_careunit = 'CCU' OR last_careunit = 'CCU';

-- returns:
 subject_id
------------
        109
        109
        111
        ...
```

Dado que un paciente pudo estar en varias UCIs, el mismo ID del paciente a veces aparece varias veces en el resultado de la consulta previa. Para obtener solamente filas distintas, hay que usar la palabra clave **DISTINCT**.

```
SELECT DISTINCT subject_id
FROM icustays
WHERE first_careunit = 'CCU' OR last_careunit = 'CCU';

-- returns:
 subject_id
------------
      25949
       6158
      27223
      ...
```

Para contar cuantos pacientes hay en la tabla **icustays,** se debe combinar **DISTINCT** con la palabra clave **COUNT**. Como se puede apreciar, si no hay condición, simplemente no usamos la palabra clave **WHERE**.

```
SELECT COUNT(DISTINCT subject_id)
FROM icustays;

-- returns:
 count
-------
 46476
```

De la misma manera, podemos saber cuántos pacientes pasaron a través de la UCO usando la consulta:

```
SELECT COUNT(DISTINCT subject_id)
FROM icustays
WHERE first_careunit = 'CCU' OR last_careunit = 'CCU';

-- returns:
 count
-------
  7314
```

El operador * es usado para mostrar todas las columnas. La siguiente consulta muestra la tabla **icustays** entera:

```
SELECT *
FROM icustays;

-- returns
subject_id | hadm_id | icustay_id | ...
       109 |  139061 |     257358 | ...
       109 |  172335 |     262652 | ...
       109 |  126055 |     236124 | ...
...
```

Los resultados se pueden ordenar en una o varias columnas con **ORDER BY**. Para agregar un comentario en una consulta SQL, usar:

```
SELECT subject_id, hadm_id, icustay_id
FROM icustays
ORDER BY subject_id ASC; -- ASC sorts by ascending number

-- returns:
 subject_id | hadm_id | icustay_id
------------+---------+------------
          2 |  163353 |     243653
          3 |  145834 |     211552
          4 |  185777 |     294638
...
```

11.3.3 Uniones (JOINS)

Usualmente necesitamos información que proviene de múltiples tablas. Esta puede ser obtenida usando uniones SQL. Hay varios tipos de uniones, incluyendo **INNER JOIN, OUTER JOIN, LEFT JOIN, y RIGHT JOIN.** Es importante entender la diferencia entre estas uniones porque su uso puede impactar significativamente en las consultas. La guía detallada de uniones se encuentra disponible en la web, por ese motivo no entraremos en demasiados detalles. Sin embargo, daremos un ejemplo de una **INNER JOIN**, la cual selecciona todas las filas donde la clave de unión aparece en ambas tablas.

Usando la palabra clave **INNER JOIN**, vamos a contar cuantos pacientes adultos pasaron por la unidad coronaria. Para saber si un paciente es un adulto, necesitamos usar el atributo **dob** (año de nacimiento, por sus siglas en inglés) de la tabla **patients**. Podemos utilizar **INNER JOIN** para indicar que dos o más tablas deben estar combinadas basadas en una característica común, en nuestro caso el **subject-id**:

```
-- INNER JOIN will only return rows where subject_id
-- appears in the patients table and the icustays table
SELECT p.subject_id
FROM patients p
INNER JOIN icustays i
ON p.subject_id = i.subject_id
WHERE (i.first_careunit = 'CCU' OR i.last_careunit = 'CCU')
   AND (i.intime - p.dob) >= INTERVAL '18' year
ORDER BY subject_id ASC;

-- returns:
 subject_id
------------
         13
         18
         21
         ...
```

Tener en cuenta que:

- Asignamos un alias a una tabla para evitar anotar el nombre completo en toda la consulta. En nuestro () dimos el alias “p”.
- En la cláusula **SELECT**, anotamos **p. subject-id** en lugar de **subject-id** ya que ambas tablas **patients e icustays** contienen el atributo **subject-id**. Si

no especificamos de que tabla proviene **subject-id**, obtendríamos un error de “columna definida ambiguamente”.

- Para identificar si un paciente es un adulto, buscamos las diferencias de edad de 18 años o más entre **intime y dob** usando la palabra clave **INTERVAL**.

11.3.4 Rankear a través de filas usando una función de ventana

Nos centraremos ahora en el estudio del caso. Uno de los primeros pasos es identificar la primera admisión de cada paciente en la UCI. Para hacer esto, podemos usar la función **RANK ()** para ordenar las filas secuencialmente por tiempo. Usando la expresión **PARTITION BY** nos permite realizar el ranking a través de la ventana de **subject-id**:

```
SELECT subject_id, icustay_id, intime,
    RANK() OVER (PARTITION BY subject_id ORDER BY intime asc)
FROM icustays;

-- returns:
 subject_id | icustay_id |       intime        | rank
------------+------------+---------------------+------
          6 |     228232 | 2175-05-30 21:30:54 |    1
          7 |     278444 | 2121-05-23 15:35:29 |    1
          7 |     236754 | 2121-05-25 03:26:01 |    2
          ...
```

11.3.5 Hacer las consultas más administrables utilizando WITH

Para mantener las consultas SQL razonablemente cortas y simples, podemos usar la palabra clave **WITH**. **WITH** nos permite dividir la consulta en partes más pequeñas y manejables. La siguiente consulta crea temporalmente una tabla llamada “rankedstays” que lista el orden de estadía de cada paciente. Luego seleccionamos solamente las filas en esta tabla donde el ranking es igual a uno (por ejemplo, la primer estadía) y el paciente tiene 18 años o más:

```
WITH rankedstays AS (
    SELECT subject_id, icustay_id, intime,
        RANK() OVER (PARTITION BY subject_id ORDER BY intime asc)
    FROM icustays
)
SELECT r.subject_id, r.icustay_id, r.intime, r.rank
FROM rankedstays r
INNER JOIN patients p
ON r.subject_id = p.subject_id
WHERE r.rank = 1
AND (r.intime - p.dob) >= INTERVAL '18' year;

-- returns:
 subject_id | icustay_id |       intime        | rank
------------+------------+---------------------+------
          3 |     211552 | 2101-10-20 19:10:11 |    1
          4 |     294638 | 2191-03-16 00:29:31 |    1
          6 |     228232 | 2175-05-30 21:30:54 |    1
          ...
```


Referencias

1. Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, Guy RT et al (2014) Best practices for scientific computing. PLoS Biol 12 (1): e1001745. doi: 10.1371/journal. pbio.1001745. http://journals. plos.org/plosbiology/article?id=10.1371/journal. pbio.1001745.
2. Editorial (2012) Must try harder. Nature 483 (509). doi: 10.1038/483509a. http://www.nature.com/nature/journal/v483/n7391/full/483509a.html.
3. Misset B, Nakache D, Vesin A, Darmon M, Garrouste-Orgeas M, Mourvillier B et al (2008) Reliability of diagnostic coding in intensive care patients. Crit Care 12 (4): R95. http://www.ncbi. nlm.nih.gov/pmc/articles/PMC2575581/.
4. Wickham H (2014) Tidy data. J Stat Softw 59 (10): 1-23. doi: 10.18637/jss. v059. i10. https://www.jstatsoft.org/article/view/v059i10

5. Sustainability of Digital Formats Planning for Library of Congress Collections. [Consultado 24 Febrero 2016]. CSV, Comma Separated Values (RFC 4180). Disponible en http://www.digitalpreservation.gov/formats/fdd/fdd000323.shtml.
6. Editorial (2013) Unreliable research: trouble at the Lab. Economist. http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble.
7. Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, Crosas M, et al (2014) Ten simple rules for the care and feeding of scientific data. PLoS Comput Biol 10 (4): e1003542. doi: 10.1371/journal. pcbi.1003542. http://journals.plos.org/ploscompbiol/article?id=10.1371/ journal. pcbi.1003542.
8. Karthik R (2013) Git can facilitate greater reproducibility and increased transparency in science. Source Code Biol Med 28; 8 (1): 7. doi: 10.1186/1751-0473-8-7. http://scfbm.biomedcentral.com/articles/10.1186/1751-0473-8-7.
9. GitHub. https://github.com. [Consultado 24 Feb 2016].
10. MIMIC website. http://mimic.physionet.org. [Consultado24 Feb 2016]
11. MIMIC Code Repository. Disponible en https://github.com/MIT-LCP/mimic-code. [Consultado 24 Feb 2016].

Capítulo 12

PREPROCESAMIENTO DE DATOS

Bryan Malley, Daniele Ramazzotti
y Joy Tzung Wu

Objetivos de aprendizaje

- Entender los requerimientos de una base de datos "limpia", que se encuentre "ordenada" y lista para su uso en análisis estadístico.
- Entender los pasos para la limpieza de datos crudos, integración de datos, reducción y reestructuración de datos.
- Ser capaz de aplicar técnicas básicas para lidiar con problemas comunes con datos crudos, incluyendo datos ausentes, inconsistente y de múltiples fuentes.

12.1 Introducción

El preprocesamiento de datos consiste en una serie de pasos necesarios para transformar datos crudos derivados de la extracción (ver capítulo 11) en un set de datos "limpio" y "ordenado" previo al análisis estadístico. La investigación que utiliza historias clínicas electrónicas (HCEs) a menudo involucra el análisis secundario de registros de salud que fueron recolectados para fines clínicos y de facturación (no de estudio) e ingresados en una base de datos de estudio por medio de procesos automáticos. Por lo tanto, estas bases de datos pueden tener muchos problemas de control de calidad. El preprocesamiento apunta a evaluar y mejorar la calidad de los datos para permitir un análisis estadístico confiable.

El preprocesamiento de datos comprende varios pasos diferentes. Los siguientes son los pasos generales requeridos [1]:

- "Limpieza" de datos. Este paso trata con datos ausentes, ruido, valores *outliers* y registros incorrectos o duplicados mientras minimiza la introducción de sesgo en la base de datos. Estos métodos son explorados en detalle en los capítulos 13 y 14.
- "Integración de datos". Los datos crudos extraídos pueden provenir de fuentes heterogéneas o estar en sets de datos separados. Este paso

reorganiza los sets de datos crudos en un solo set que contiene toda la información requerida para el análisis estadístico deseado.

- "Transformación de datos". Este paso traduce y/o escala variables almacenadas en una variedad de formatos o unidades en los datos crudos, en formatos o unidades que son más útiles para los métodos estadísticos que el investigador quiere implementar.
- "Reducción de datos". Después de haber integrado y transformado el set de datos, este paso elimina registros y variables redundantes, así como reorganiza los datos en una manera eficiente y "ordenada" para el análisis.

El preprocesamiento a veces es iterativo y puede involucrar repetir esta serie de pasos hasta que los datos se encuentren organizados satisfactoriamente para el análisis estadístico propuesto. Durante el preprocesamiento, es necesario asegurarse de no introducir accidentalmente un sesgo al modificar un set de datos de manera tal que pueda impactar en el resultado del análisis estadístico. Asimismo, debemos evitar alcanzar resultados estadísticamente significativos a través de análisis de "prueba y error" en diferentes versiones de sets de datos preprocesados.

12.2 Parte 1 - Conceptos teóricos

12.2.1 Limpieza de Datos

Los datos del mundo real están generalmente "desordenados": pueden estar incompletos (por ejemplo, datos faltantes), pueden tener ruido (por ejemplo, errores aleatorios o valores atípicos que se desvían de la línea de base esperada), y pueden ser inconsistentes (por ejemplo, un paciente de 21 años admitido al servicio de terapia intensiva neonatal).

Las razones de esto son múltiples. Los datos pueden faltar debido a problemas técnicos fortuitos con los monitores de signos vitales, la dependencia en el ingreso de datos por parte del personal de salud, o porque algunas variables clínicas no son adquiridas de manera consistente dado que los datos de la HCE fueron recolectados con fines distintos al objetivo de nuestro estudio. De manera similar, el ruido en los datos puede ser por fallas o limitaciones tecnológicas de los instrumentos durante la recolección (por ejemplo, la disminución de los valores de presión arterial

medidos a través de una vía arterial), o por error humano en el registro. Todos los factores mencionados previamente también pueden llevar a inconsistencias en los datos. En conclusión, todas estas razones crean la necesidad de una limpieza meticulosa de los datos previa al análisis.

Datos faltantes

En el capítulo 13 presentaremos una discusión más detallada con respecto a los datos faltantes. Aquí describimos tres formas posibles de tratar con los datos faltantes [1]:

- Ignorar el registro. Este método no es muy efectivo, a menos que el registro (observación/fila) contenga muchas variables con valores faltantes. Este abordaje es especialmente problemático cuando el porcentaje de valores faltantes por variable varía considerablemente o cuando hay un patrón de datos faltantes relacionado con una causa subyacente no reconocida, como una enfermedad o condición particular de un paciente en la admisión.
- Determinar y completar el valor faltante manualmente. En general, este abordaje es el más certero pero a la vez es el que más tiempo consume y habitualmente no es factible en un set de datos grande con muchos valores faltantes.
- Usar un valor esperado. Los valores faltantes pueden ser suplidos con valores predichos (por ejemplo, usando la media de los datos disponibles o algún método de predicción). Se debe subrayar que este abordaje puede introducir sesgo en los datos, dado que los valores insertados pueden ser incorrectos. Este método también es útil para comparar y chequear la validez de los resultados obtenidos, ignorando registros faltantes.

Datos con ruido

Llamamos *"ruido"* a errores o diferencias aleatorias observadas en una variable –un problema común para el análisis secundario de datos de las HCEs. Por ejemplo, no es infrecuente que los pacientes hospitalizados tengan signos vitales o valores de laboratorio muy por fuera de los parámetros normales debido a muestras de sangre inadecuadas (hemolizadas), o cables de los monitores desconectados por movimientos del paciente. Los médicos por lo general conocen estas fuentes de errores y pueden repetir la medición

e ignorar el valor atípico incorrecto, al definir el tratamiento del paciente. Sin embargo, en muchos casos los médicos no pueden eliminar mediciones erróneas de la historia clínica, por lo tanto éstas serán capturadas en la base de datos. En el capítulo 14 se brinda una discusión detallada sobre cómo se deben tratar los datos con ruido y atípicos; por ahora limitamos la discusión a algunos lineamientos básicos [1].

- Métodos de *binning* (discretización, o cuantización). Los métodos de *binning*, suavizan los datos ordenados considerando sus "alrededores", o los valores cercanos. Este tipo de abordajes para reducir ruido, que solamente consideran los valores cercanos, se dice que hacen "suavizado local".
- *Clustering*. Los datos atípicos pueden ser detectados por *clustering*, que consiste agrupar un conjunto de valores de forma tal que los que están en el mismo grupo (es decir, en el mismo *cluster*) son más similares entre sí que con aquellos en otros grupos.
- Aprendizaje automático (*machine learning*). Los datos pueden ser suavizados por medio de varios abordajes de aprendizaje automático.

Uno de los métodos clásicos es el análisis de regresión, donde los datos son ajustados a una función específica (por lo general lineal).

Al igual que para los datos faltantes, la supervisión humana durante el proceso de suavizado de ruido o detección de datos atípicos puede ser efectiva pero también requiere mucho tiempo.

Datos inconsistentes

Puede haber inconsistencias o datos duplicados. Algunos de ellos pueden ser corregidos manualmente usando referencias externas. Este es el caso, por ejemplo, de errores cometidos durante el ingreso de datos. También es posible utilizar herramientas de ingeniería del conocimiento para detectar la violación de restricciones en los datos conocidas. Por ejemplo, se pueden usar las dependencias funcionales conocidas entre atributos para encontrar valores que contradigan las restricciones funcionales.

Las inconsistencias en las HCEs son el resultado de información ingresada en la base de datos por miles de médicos y miembros del personal de salud del hospital, además de información capturada de una variedad de interfaces automatizadas entre la HCE y el resto, desde los monitores de telemetría

hasta el laboratorio del hospital. La misma información frecuentemente es ingresada en formatos diferentes desde estas distintas fuentes. Tomemos, por ejemplo, la administración intravenosa de 1 g del antibiótico vancomicina contenido en 250 mL de solución de dextrosa. Este evento puede ser capturado en el set de datos de muchas formas diferentes. Para un paciente, este evento puede ser capturado desde la orden de medicación como desde el número de código (ITEMID en MIMIC) del formulario para el antibiótico vancomicina con una columna separada capturando la dosis, almacenada como una variable numérica. Sin embargo, en otro paciente el mismo evento podría ser encontrado en el ingreso de fluidos y en los registros de salida bajo el código para la solución de dextrosa IV con un texto libre asociado ingresado por el proveedor. Este texto sería capturado en la HCE como, por ejemplo, "vancomicina 1 g en 250 ml", guardado como una variable de texto (cadena de caracteres, vector de caracteres, etc.) con la posibilidad de errores de ortografía o el uso de abreviaciones no estandarizadas. Clínicamente, éstos indican exactamente el mismo evento, pero en la HCE y por ende en los datos crudos, son representados de manera diferente. Esto puede llevar a que un único evento clínico no sea capturado en el set de datos del estudio, siendo considerado incorrectamente como un evento diferente, o siendo capturado múltiples veces ocurriendo una sola vez.

Para producir un set de datos adecuado para el análisis, cada paciente debe tener el mismo evento representado de la misma manera. Para esto, el tratamiento de las inconsistencias debería suceder en la fase de ingreso o de extracción de datos. Es por ello, teniendo en cuenta que la extracción de datos es imperfecta, el preprocesamiento se vuelve importante. Frecuentemente, corregir estas inconsistencias implica cierta comprensión sobre cómo habrán sido capturados los datos de interés en el ámbito clínico y dónde estarán almacenados en la base de datos de la HCE.

12.2.2 Integración de Datos

La integración de datos es el proceso en el que se combinan datos derivados de varias fuentes de datos (como bases de datos, archivos planos, etc.) en un set de datos consistente. Hay una serie de problemas a considerar durante la integración de los datos, relacionados sobretodo con estándares posiblemente diferentes entre las fuentes de datos. Por ejemplo,

algunas variables pueden referirse por medio de identidades diferentes, en dos o más fuentes.

En la base de datos MIMIC, esto se convierte en un problema principalmente cuando alguna información es ingresada en la HCE durante diferentes momentos en la trayectoria de la atención del paciente, como antes de la admisión en el departamento de emergencias, o en registros externos. Por ejemplo, un paciente puede tener valores de laboratorio realizados en la sala de emergencias antes de ser admitidos en la unidad de cuidados intensivos (UCI). Para tener un set de datos completo, será necesario integrar todo el set de valores de laboratorio del paciente (incluyendo aquellos no asociados con el mismo identificador MIMIC *ICUSTAY*) con el registro de admisión a la UCI, sin repetir o perder registros. Se puede alcanzar la integración adecuada, usando valores compartidos entre los sets de datos (como en este ejemplo, un identificador de estadía hospitalaria o una marca de tiempo).

Una vez que se completa la limpieza e integración de los datos, obtenemos un set de datos donde las entradas son confiables.

12.2.3 Transformación de Datos

Hay muchas posibles transformaciones que uno podría hacer a los datos crudos, dependiendo del análisis estadístico planeado para el estudio. El objetivo es transformar los valores de los datos en un formato, escala o unidad que sea más adecuada para el análisis (por ejemplo, la transformación logarítmica para un modelo de regresión lineal). Estas son algunas opciones comunes posibles:

Normalización

Esto generalmente significa que los datos para una variable numérica son escalados para estar entre un set de valores específico, como 0-1. Por ejemplo, se puede escalar el puntaje de severidad de la enfermedad de cada paciente entre 0 y 1, usando el rango conocido del puntaje para comparar entre pacientes en un análisis de regresión múltiple.

Agregación

Dos o más valores de un mismo atributo son resumidos en uno. Un ejemplo común es la transformación de variables categóricas donde

múltiples categorías pueden ser agrupadas en una. Un ejemplo en MIMIC es definir todos los pacientes quirúrgicos asignándoles una nueva variable binaria a todos los pacientes con servicio de UCI indicado cómo "*SICU*" (UCI quirúrgica) o "*CSRU*" (UCI de cirugía cardíaca).

Generalización

Similar a la agregación, en este caso los atributos de bajo nivel son transformados en atributos de mayor nivel. Por ejemplo, en el análisis de pacientes con enfermedad renal crónica (ERC), en vez de usar variables numéricas continuas como los niveles de creatinina del paciente, uno podría usar una variable para los estadios de la ECR de acuerdo a su definición en guías aceptadas.

12.2.4 Reducción de los datos

El análisis complejo en sets de datos grandes puede llevar mucho tiempo o incluso ser inviable. El paso final del pre procesamiento de datos es la reducción de los datos, es decir, el proceso de reducción de los datos ingresados por medio de una representación más efectiva del set de datos sin comprometer la integridad de los datos originales. El objetivo de este paso es proveer una versión del set de datos en la cual el análisis estadístico subsecuente sea más efectivo. La reducción de datos puede estar o no libre de pérdidas. Esto significa que la base de datos final puede contener toda la información de la base de datos original en un formato más eficiente (como eliminando registros redundantes) o puede ser que la integridad se mantenga pero que se pierda alguna información en la transformación de los datos, y que estos estén por lo tanto solamente representados en la forma nueva (como múltiples valores representados como un valor promedio).

Un ejemplo habitual de la base de datos MIMIC es el colapso de códigos CIE9 en categorías clínicas amplias o variables de interés, y la asignación de pacientes a estas categorías. Esto reduce el set de datos, que pasan de tener múltiples entradas de códigos CIE9, en formato de texto, para un determinado paciente, a tener una sola entrada de una variable binaria para un área de interés para el estudio, como la historia de enfermedad coronaria. Otro ejemplo sería el caso de usar presión arterial como una variable en el análisis. Un paciente de ICU tendrá generalmente su presión sistólica y diastólica monitoreada continuamente con una vía arterial o

medida múltiples veces por hora con un manguito de presión arterial automatizado. Esto resulta en cientos de puntos de datos para cada uno de los quizás miles de pacientes en el estudio. Dependiendo de los objetivos del estudio, puede ser necesario calcular una nueva variable como el promedio de la presión arterial media durante el primer día de admisión a la UCI.

Finalmente, como parte de una organización más efectiva de los sets de datos, uno podría también tener como objetivo reformar sus columnas y filas, para que estén acorde con las siguientes 3 reglas de un set de datos "ordenados" [2, 3]:

1. Cada variable forma una columna
2. Cada observación forma una fila
3. Cada valor tiene su propia celda

Los sets de datos "ordenados" tienen la ventaja de ser fácilmente visualizados y manipulados para análisis estadístico posterior. Los sets de datos exportados desde MIMIC están por lo general un tanto "ordenados"; por lo tanto, es muy raro que no se cumpla la regla 2. Sin embargo, algunas veces puede haber todavía muchos valores categóricos dentro de una columna, incluso para el sets de datos MIMIC, lo cual rompe la regla 1. Por ejemplo, múltiples categorías de estatus marital o etnia bajo la misma columna. Para algunos análisis, es útil separar cada valor categórico de una variable en sus propias columnas. Afortunadamente, no tenemos que preocuparnos a menudo por la regla 3 para datos de MIMIC, dado que en general no se presentan múltiples valores en una celda. Estos conceptos se aclararán luego de los ejemplos de MIMIC en Sección 12.3.

12.3 Parte 2 - Ejemplos de preprocesamiento de datos en R

Hay muchas herramientas disponibles para hacer el preprocesamiento de datos, por ejemplo R, STATA, SAS y Python; cada una difiere en el nivel de experiencia en programación requerida.

R es una herramienta libre que es compatible con una variedad de paquetes estadísticos y de manipulación de datos. En esta sección del capítulo, vamos a repasar algunos ejemplos, demostrando varios pasos del preprocesamiento de datos en R, usando datos de varios set de datos de MIMIC (incluidos códigos de extracción SQL). Debido al importante contenido involucrado en el paso de la limpieza de datos de

preprocesamiento, este paso va a ser tratado por separado en los capítulos 13 y 14. Los ejemplos en esta sección se ocuparán de algunos conocimientos básicos de R así como de la integración, la reducción y la transformación de los datos.

12.3.1 R - Conocimientos básicos

La salida de datos más comunes desde una consulta en una base de datos MIMIC es en forma de archivo de "valores separados por coma", con los nombres de archivo finalizando en "c*sv".* Este formato de archivo de salida puede ser seleccionado cuando exportamos los resultados de la consulta SQL desde la base de datos MIMIC. Además de los archivos. "c*sv",* R también puede leer otros formatos de archivo, como Excel, SAS, etc., pero no entraremos aquí en este detalle.

Entendiendo los "Tipos de Datos" en R

Muchos de los que han usado otro software de análisis de datos o que tienen experiencia en programación, estarán familiarizados con el concepto de "tipos de datos". R estrictamente almacena datos en varios tipos de datos diferentes, llamados "clases":

- Numérica – ejemplo 3.1415, 1.618
- Entero – ejemplo -1, 0, 1, 2, 3
- Caracter – ejemplo "vancomicina", "metronidazol".
- Lógico – VERDADERO, FALSO (TRUE – FALSE)
- Factores/categóricos – ejemplo hombre o mujer para la variable género

R además generalmente no admite la combinación de tipos de datos para una variable, excepto en:

- List – como un vector dimensional, ejemplo ("vancomicina", 1.618, "rojo")
- Data-frame – como una tabla de 2 dimensiones con filas (observaciones) y columnas (variables)

Las listas y los data frames son tratados como su propia "clase" en R.

La salida de las consultas desde MIMIC normalmente será en la forma de tablas de datos con diferentes tipos en diferentes columnas. Por lo tanto, R

generalmente almacena esas tablas como *"data frames*" cuando son leídos en R.

Valores Especiales en R

- NA-"not available", generalmente un indicador por default para valores faltantes.
- NAN-"not a number", solo aplica a vectores numéricos.
- NULL-valor "empty" o set. En general devuelto por expresiones donde el valor es indefinido.
- Inf-valor para "infinity" y solo aplica a vectores numéricos.

Configurando un Directorio de Trabajo

Este paso le indica a R dónde leer en la fuente de archivos.

Comando: setwd ("directory-path")

Ejemplo: (Si todos los archivos de datos son guardados en el directorio "MIMIC-data-files" en el Escritorio (Desktop))

```
setwd("~/Desktop/MIMIC_data_files")

# List files in directory:
list.files()
## [1] "c_score_sicker.csv"          "comorbidity_scores.csv"
## [3] "demographics.csv"            "mean_arterial_pressure.csv"
## [5] "population.csv"
```

Leyendo Archivos. csv desde los Resultados de una Consulta MIMIC

A los datos leídos en R se les asigna un "nombre" para luego referenciarlos.

Ejemplo:

```
demo <- read.csv("demographics.csv")
```

Visualizando el Set de datos

Hay varios comandos en R que son muy útiles para obtener una "mirada" de los conjuntos de datos y ver como lucen antes de comenzar a manipularlos.

- Visualizar las dos primeras y las dos últimas filas. Por ejemplo:

```
head(demo, 2)

##   subject_id hadm_id marital_status_descr ethnicity_descr
## 1          4   17296               SINGLE           WHITE
## 2          6   23467              MARRIED           WHITE

tail(demo, 2)

##       subject_id hadm_id marital_status_descr  ethnicity_descr
## 27624      32807   32736              MARRIED UNABLE TO OBTAIN
## 27625      32805   34884             DIVORCED            WHITE
```

- Visualizar las estadísticas resumen. Por ejemplo:

```
summary(demo)

##    subject_id       hadm_id      marital_status_descr
##  Min.   :    3   Min.   :    1     MARRIED  :13447
##  1st Qu.: 8063   1st Qu.: 9204   SINGLE   : 6412
##  Median :16060   Median :18278   WIDOWED  : 4029
##  Mean   :16112   Mean   :18035   DIVORCED : 1623
##  3rd Qu.:24119   3rd Qu.:26762            : 1552
##  Max.   :32809   Max.   :36118   SEPARATED:  320
##                                    (Other)  :  242
##                ethnicity_descr
##  WHITE                  :19360
##  UNKNOWN/NOT SPECIFIED : 3446
##  BLACK/AFRICAN AMERICAN: 2251
##  …
```

- Visualizar la estructura del conjunto de datos (obs= número de filas). Por ejemplo:

```
str(demo)

## 'data.frame':    27625 obs. of  4 variables:
##  $ subject_id          : int  4 6 3 9 15 14 11 18 18 19 ...
##  $ hadm_id             : int  17296 23467 2075 8253 4819 23919 28128
24759 33481 25788 ...
##  $ marital_status_descr: Factor w/ 8 levels "","DIVORCED",..: 6 4 4
1 6 4 4 4 4 1 ...
##  $ ethnicity_descr     : Factor w/ 39 levels "AMERICAN INDIAN/ALASKA
NATIVE",..: 35 35 35 34 12 35 35 35 35 35 ...
```

- Descubrir la “clase” de una variable o un set de datos. Por ejemplo:

```
class(demo)

## [1] "data.frame"
```

- Visualizar el número de filas y columnas, o alternativamente, la dimensión del set de datos. Por ejemplo:

```
nrow(demo)

## [1] 27625

ncol(demo)

## [1] 4

dim(demo)

## [1] 27625     4
```

- Calcular la longitud de una variable. Por ejemplo:

```
length(x)
length(y)
```

```
[1] 1
[1] 5
```

Subconfigurando un Set de Datos y agregando Nuevas Variables/Columnas

Objetivo: A veces, puede ser útil ver solamente algunas columnas o filas en un set de datos/*data-frame*---a esto se llama subagrupar.

Vamos a crear un *data-frame* simple para demostrar conceptos básicos para subagrupar y otras funciones y comandos en R. Una forma simple para realizar esto es crear cada columna del *dataframe* de forma separada y luego combinarlas en un conjunto de datos. Notar los diferentes tipos de datos para las columnas/variables creadas, y tener en cuenta que R es sensible a mayúsculas y minúsculas.

Ejemplos: Observar que los comentarios que aparecen luego del signo (#) no serán evaluados.

```
subject_id <- c(1:6)                                    #integer
gender <- as.factor(c("F", "F", "M", "F", "M", "M"))#factor/categorical
height <- c(1.52, 1.65, 1.75, 1.72, 1.85, 1.78)     #numeric
weight <- c(56.7, 99.6, 90.4, 85.3, 71.4, 130.5)    #numeric
data <- data.frame(subject_id, gender, height, weight)

head(data, 4)                                  # View only the first 4 rows

##   subject_id gender height weight
## 1          1      F   1.52   56.7
## 2          2      F   1.65   99.6
## 3          3      M   1.75   90.4
## …

str(data)                    # Note the class of each variable/column

## 'data.frame':    6 obs. of  4 variables:
##  $ subject_id: int  1 2 3 4 5 6
##  $ gender    : Factor w/ 2 levels "F","M": 1 1 2 1 2 2
##  $ height    : num  1.52 1.65 1.75 1.72 1.85 1.78
##  $ weight    : num  56.7 99.6 90.4 85.3 71.4 ...
```

Para subagrupar o solamente extraer por ejemplo, el peso, podemos usar ya sea el signo dólar ($) después del set de datos, o usar corchetes, []. El $ selecciona la columna con el nombre de la columna (sin comillas en este caso). Los corchetes [] aquí seleccionaron la columna del peso por su número de columna:

```
w1 <- data$weight; w1

## [1]  56.7  99.6  90.4  85.3  71.4 130.5

w2 <- data[, 4]; w2

## [1]  56.7  99.6  90.4  85.3  71.4 130.5
```

Generalmente uno puede subagrupar un set de datos especificando las filas y la columna deseada de esta manera: datos [número de fila, número de columna]. Por ejemplo:

```
dat_sub <- data[2:4, 1:3]; dat_sub

##   subject_id gender height
## 2          2      F   1.65
## 3          3      M   1.75
## 4          4      F   1.72
```

Los corchetes son útiles para subagrupar múltiples columnas o filas. Tener en cuenta que es importante "concatenar", c (), si seleccionamos múltiples variables/columnas y usar comillas al seleccionar los nombres de las columnas.

```
h_w1 <- data[, c(3, 4)]; h_w1

##   height weight
## 1   1.52   56.7
## 2   1.65   99.6
## 3   1.75   90.4
## …

h_w2 <- data[, c("height", "weight")]; h_w2

##   height weight
## 1   1.52   56.7
## 2   1.65   99.6
## 3   1.75   90.4
## …
```

Hay diferentes formas de calcular el Indice de Masa Corporal (IMC) (peso/altura^2), en una nueva columna, pero aquí se muestra un método simple:

```
data$BMI <- data$weight/data$height^2
head(data, 4)

##   subject_id gender height weight      BMI
## 1          1      F   1.52   56.7 24.54120
## 2          2      F   1.65   99.6 36.58402
## 3          3      M   1.75   90.4 29.51837
## 4          4      F   1.72   85.3 28.83315
```

BMI, índice de masa corporal

Vamos a crear una nueva columna, obesidad, para un IMC>30, como VERDADERO o FALSO. Esto también demuestra el uso de "lógica" en R.

```
data$obese <- data$BMI > 30
head(data)

##   subject_id gender height weight      BMI obese
## 1          1      F   1.52   56.7 24.54120 FALSE
## 2          2      F   1.65   99.6 36.58402  TRUE
## 3          3      M   1.75   90.4 29.51837 FALSE
## …
```

Podemos también usar vectores lógicos para subagrupar un set de datos en R. Se crea un vector lógico, llamado aquí "ob", y luego lo colocamos entre corchetes para indicarle a R que seleccione solamente las filas donde el IMC>30 es VERDADERO:

```
ob <- data$BMI > 30
data_ob <- data[ob, ];data_ob

##   subject_id gender height weight      BMI obese
## 2          2      F   1.65   99.6 36.58402  TRUE
## 6          6      M   1.78  130.5 41.18798  TRUE
```

Combinando Set de datos (Llamados Data-Frames en R)

Objetivo: Usualmente diferentes variables (columnas) de interés en una pregunta de investigación pueden venir de tablas separadas en MIMIC y podrían haber sido exportadas como archivos. csv separados si estas no fueron fusionadas mediante consultas SQL. Para facilitar el análisis y la visualización, habitualmente es deseable fusionar estos *data-frames* separados en R con su columna(s) ID compartida(s).

Ocasionalmente, uno puede querer adjuntar filas de un *data-frame* y luego filas de otro. En este caso, los nombres de la columna y el número de columnas de los dos diferentes set de datos tienen que ser iguales.

Ejemplos: Por lo general, hay algunas formas de combinar columnas y filas desde set de datos distintos en R:

- merge () ----Esta función fusiona columnas sobre la columna ID compartida por los dos *data-frames* para que las filas asociadas coincidan de manera correcta.

Comando: fusionando en un ID de columna, ejemplo:

```
df_merged <- merge(df1, df2, by = "column_ID_name")
```

Comando: fusionando en dos ID de columna, ejemplo:

```
df_merged <- merge(df1, df2, by = c("column1", "column2"))
```

- cbind () ----esta función simplemente "agrega" de manera conjunta las columnas desde dos *data-frame* (deben tener el mismo número de filas). No hace coincidir las filas por ningún identificador.

Comando: uniendo columnas, ejemplo:

```
df_total <- cbind(df1, df2)
```

- rbind () ----la función une filas de dos *data-frames* verticalmente (las columnas en ambos deben tener el mismo nombre).

Comando: uniendo filas, ejemplo:

```
df_total <- rbind(df1, df2)
```

Usando paquetes en R

Hay múltiples paquetes que nos hacen la vida más sencilla cuando manipulamos datos en R. Antes de poder llamar a las funciones que ellos contienen, necesitan estar instalados en su computadora y ser cargados al inicio de la secuencia de comandos en R. Introduciremos ejemplos de algunos paquetes útiles más adelante en este capítulo.

Por ahora, el comando para la instalación de paquetes es:

```
install.packages("name_of_package_case_sensitive")
```

El comando para cargar el paquete en el ambiente de trabajo de R:

```
library(name_of_package_case_sensitive)
```

Nota----no hay comillas cuando cargamos los paquetes en comparación a la instalación; de la otra manera obtendremos un mensaje de error.

Obteniendo ayuda en R

Hay varios tutoriales online y foros de preguntas y respuestas para obtener ayuda en R. Stackoverflow, Cran y Quick-R son algunos buenos ejemplos. Dentro de la consola R, un signo de pregunta,?, seguido del

nombre de la función de interés abrirá el menú de ayuda para la función, por ejemplo.

```
?head
```

12.3.2 Integración de datos

Objetivo: Esto implica combinar los set de datos separados de los resultados exportados desde distintas consultas de MIMIC en un set de datos de destino de mayor tamaño.

Para asegurarnos que coinciden las observaciones asociadas en filas desde los dos set de datos, debemos usar la columna de ID derecha. En MIMIC, las columnas de ID (identificadores únicos) podría ser subject-id, icustay-id, itemid, etc. Por lo tanto, es importante conocer qué ID de columna se usa para identificar y como están relacionados unos con otros. Por ejemplo, subject-id es usado para identificar cada paciente individual, por lo tanto incluye su fecha de nacimiento (DOB), fecha de defunción (DOD) y varios detalles clínicos y valores de laboratorio en MIMIC. Asimismo, el ID de admisión hospitalaria hadm_id, es usado para identificar específicamente varios eventos y resultados desde una única admisión hospitalaria; y también está asociada con el subject_id del paciente que fue el involucrado en esa admisión hospitalaria en particular. Las tablas extraídas de MIMIC pueden tener una o más columnas de ID. Las diferentes tablas exportadas de MIMIC pueden contener diferentes columnas de ID, lo que nos permite "fusionarlas", haciendo que coincidan las filas correctamente mediante el uso valores de ID únicos que se encuentran en cada una de las tablas de origen.

Ejemplos: Para demostrar esto con datos MIMIC, se construye una consulta SQL simple para extraer algunos datos, guardados como: "population. csv" y "demographics. csv". Vamos a estos archivos extraídos para mostrar como fusionar set de datos en R.

1. Consulta SQL

```
WITH
population AS(
SELECT subject_id, hadm_id, gender, dob, icustay_admit_age,
icustay_intime, icustay_outtime, dod, expire_flg
FROM mimic2v26.icustay_detail
  WHERE subject_icustay_seq = 1
  AND icustay_age_group = 'adult'
  AND hadm_id IS NOT NULL
)
, demo AS(
SELECT subject_id, hadm_id, marital_status_descr, ethnicity_descr
FROM mimic2v26.demographic_detail
WHERE subject_id IN (SELECT subject_id FROM population)
)

--# Extract the the datasets with each one of the following line of
codes in turn:
--SELECT * FROM population
--SELECT * FROM demo
```

Nota: Elimine el – frente al comando SELECT para correr la consulta.

2. Código R: Demostrando integración de datos

Configurar el directorio de trabajo y leer archivos de datos en R:

```
setwd("~/Desktop/MIMIC_data_files")
demo <- read.csv("demographics.csv", sep = ",")
pop <- read.csv("population.csv", sep = ",")
head(demo)

##   subject_id hadm_id marital_status_descr        ethnicity_descr
## 1          4   17296                SINGLE                  WHITE
## 2          6   23467               MARRIED                  WHITE
## 3          3    2075               MARRIED                  WHITE
## …
head(pop)

##   subject_id hadm_id gender                 dob icustay_admit_age
## 1          4   17296      F 3351-05-30 00:00:00          47.84414
## 2          6   23467      F 3323-07-30 00:00:00          65.94048
## 3          3    2075      M 2606-02-28 00:00:00          76.52892
## …


##        icustay_intime     icustay_outtime                 dod
expire_flg
## 1 3399-04-03 00:29:00 3399-04-04 16:46:00
N
## 2 3389-07-07 20:38:00 3389-07-11 12:47:00
N
## 3 2682-09-07 18:12:00 2682-09-13 19:45:00 2683-05-02 00:00:00
Y
## …
```

Fusionando los set de datos pop y demo: Note que para que las filas coincidan correctamente, necesitamos fusionar en ambos el subject-id y hadm-id en este caso. Esto se debe a que cada sujeto/paciente podría tener múltiples hadm-id de diferentes admisiones hospitalarias durante el curso de la HCE de la base de datos MIMIC.

```
demopop <- merge(pop, demo, by = c("subject_id", "hadm_id"))
head(demopop)

##   subject_id hadm_id gender                 dob icustay_admit_age
## 1        100     445      F 3048-09-22 00:00:00          71.94482
## 2       1000   15170      M 2442-05-11 00:00:00          69.70579
## 3      10000   10444      M 3149-12-07 00:00:00          49.67315
## …


##        icustay_intime     icustay_outtime                dod
expire_flg
## 1 3120-09-01 11:19:00 3120-09-03 14:06:00
N
## 2 2512-01-25 13:16:00 2512-03-02 06:05:00 2512-03-02 00:00:00
Y
## 3 3199-08-09 09:53:00 3199-08-10 17:43:00
N
## …

##   marital_status_descr        ethnicity_descr
## 1              WIDOWED  UNKNOWN/NOT SPECIFIED
## 2              MARRIED  UNKNOWN/NOT SPECIFIED
## 3                          HISPANIC OR LATINO
## 4              MARRIED BLACK/AFRICAN AMERICAN
## 5              MARRIED                  WHITE
## 6            SEPARATED BLACK/AFRICAN AMERICAN
```

Como puede ver, todavía hay múltiples problemas con esta base de datos fusionada, por ejemplo, los valores faltantes para la columna “marital-status-descr”. El tratamiento de los datos faltantes es examinado en el capítulo 13.

12.3.3 Transformación de datos

Objetivo: Transformar la presentación de valores de datos de forma tal que el nuevo formato sea más adecuado para el análisis estadístico posterior. Los procesos principales involucrados son normalización, agregación y generalización (ver parte 1 de la explicación).

Ejemplos: Para demostrar esto con un ejemplo de la base de datos MIMIC, veamos una tabla generada a partir de la siguiente consulta simple SQL, la cual exportamos como “comorbidity-scores.csv”.

La consulta SQL selecciona toda la información de comorbilidades de los pacientes de la tabla mimic2v26.comorbidity-scores sobre la condición de (1)

ser un adulto, (2) en su primer/primera admisión en UCI y (3) donde no falte el hadm_id acorde con la tabla mimic2v26. icustay_detail.

1. Consulta SQL

```
SELECT *
FROM mimic2v26.comorbidity_scores
WHERE subject_id IN (SELECT subject_id
        FROM mimic2v26.icustay_detail
        WHERE subject_icustay_seq = 1
                AND icustay_age_group = 'adult'
                AND hadm_id IS NOT null)
```

2. Código R: Demostrando la transformación de datos

```
setwd("~/Desktop/MIMIC_data_files")
c_scores <- read.csv("comorbidity_scores.csv", sep = ",")
```

Note la "clase" o tipo de datos de cada columna/variable y el número total de filas (obs) y columnas (variables) in c_scores:

```
str(c_scores)

## 'data.frame':    27525 obs. of  33 variables:
##  $ subject_id              : int  2848 21370 2026 11890 27223 27520
17928 31252 32083 9545 ...
##  $ hadm_id                 : int  16272 17542 11351 12730 32530
32724 20353 30062 32216 10809 ...
##  $ category                : Factor w/ 1 level "ELIXHAUSER": 1 1 1 1
1 1 1 1 1 1 ...
##  $ congestive_heart_failure: int  0 0 0 0 1 0 0 0 1 1 ...
##  $ cardiac_arrhythmias     : int  0 1 1 0 1 0 0 0 0 1 ...
##  $ valvular_disease        : int  0 0 0 0 1 0 0 0 0 1 ...
##  $ …
```

Aquí agregamos una columna en c_scores para guardar enteramente ELIXHAUSER. La función rep () en este caso se repite 0 veces para nrow (c_scores). La función, colnames (), renombra la nueva o última columna, [ncol (c_scores)], como "ELIXHAUSER_overall".

```
c_scores <- cbind(c_scores, rep(0, nrow(c_scores)))
colnames(c_scores)[ncol(c_scores)] <- "ELIXHAUSER_overall"
```

Observemos al resultado. Note la nueva columna “ELIXHAUSER_overall” agregada al final:

```
str(c_scores)

## 'data.frame':    27525 obs. of  34 variables:
##  $ subject_id              : int  2848 21370 2026 11890 27223 27520
17928 31252 32083 9545 ...
##  $ hadm_id                 : int  16272 17542 11351 12730 32530
32724 20353 30062 32216 10809 ...
##  $ category                : Factor w/ 1 level "ELIXHAUSER": 1 1 1 1
1 1 1 1 1 1 ...
##  $ congestive_heart_failure: int  0 0 0 0 1 0 0 0 1 1 ...
##  $ cardiac_arrhythmias     : int  0 1 1 0 1 0 0 0 0 1 ...
##  $ valvular_disease        : int  0 0 0 0 1 0 0 0 0 1 ...
##  $ …
```

Paso de agregación

Objetivo: Resumir los valores de todas las comorbilidades de ELIXHAUSER a través de cada fila. Usando un bucle “for”, para cada entrada de fila i-th en la columna “ELIXHAUSER-overall”, resumimos todos los scores de comorbilidad de esa fila.

```
for (i in 1:nrow(c_scores)) {
  c_scores[i, "ELIXHAUSER_overall"] <- sum(c_scores[i,4:33])
}
```

Observemos el encabezado del resultado de la primera y última columna:

```
head(c_scores[, c(1, 34)])

##   subject_id ELIXHAUSER_overall
## 1       2848                  1
## 2      21370                  3
## 3       2026                  3
## …
```

Pasos de normalización

Objetivo: Ordenar los valores en la columna ELIXHAUSER-overall asignando un valor entre 0 y 1, por ejemplo en [0,1]. La función, max (), encuentra el máximo valor en la columna ELIXHAUSER-overall. *Entonces*

reasignamos cada entrada en la columna ELIXHAUSER-overall como una proporción de max-score para normalizar/ordenar la columna.

```
max_score <- max(c_scores[,"ELIXHAUSER_overall"])
c_scores[,"ELIXHAUSER_overall"] <- c_scores[ ,
"ELIXHAUSER_overall"]/max_score
```

Agrupamos y eliminamos todas las columnas en c-score, excepto "subject_id", "hadm_id", y "ELIXHAUSER_overall":

```
c_scores <- c_scores[, c("subject_id", "hadm_id",
"ELIXHAUSER_overall")]
head(c_scores)

##   subject_id hadm_id ELIXHAUSER_overall
## 1       2848   16272         0.09090909
## 2      21370   17542         0.27272727
## 3       2026   11351         0.27272727
## …
```

Paso de generalización

Objetivo: Considerar solamente al paciente más enfermo que el promedio del score Elixhauser. La función, which (), devuelve los números (índices) de las filas de todas las entradas TRUE de la condición lógica configurada en c_scores dentro de los paréntesis, donde la condición para la entrada de la columna sea ELIXHAUSER_overall > 0.5. Almacenamos la información de los índices de la fila en el vector "sicker". Entonces podemos usar el vector "sicker" para seleccionar dentro de c_scores solamente las filas/pacientes que están "sicker" (más enfermos) y almacenar esta información en "c_score_sicker".

```
sicker <- which(c_scores[,"ELIXHAUSER_overall"]>=0.5)
c_score_sicker <- c_scores[sicker, ]
head(c_score_sicker)

##    subject_id hadm_id ELIXHAUSER_overall
## 10       9545   10809          0.5454545
## 15      12049   27692          0.5454545
## 59      29801   33844          0.5454545
## …
```

Guardar los resultados para archivar: hay muchas funciones que harán esto, por ejemplo write.table () y write.csv (). Daremos un ejemplo aquí:

```
write.table(c_score_sicker, file = "c_score_sicker.csv", sep = ";",
row.names = F, col.names = F)
```

Si uno chequea en su directorio/carpeta, debería ver el nuevo archivo "c_score_sicker.csv".

12.3.4 Reducción de los datos

Objetivo: Reducir o remodelar la entrada de datos por medio de una representación más efectiva del set de datos sin comprometer la integridad de los datos originales. Una manera de reducir los datos es eliminar los registros redundantes, preservando los datos necesarios, lo cual será demostrado en los ejemplos (parte 1). Otro manera consiste en remodelar el set de datos a un formato ordenado, el cual demostraremos en las secciones siguientes.

Ejemplos Parte 1: eliminando registros redundantes

Para demostrar esto con un ejemplo de la base de datos de MIMIC, observaremos múltiples registros de presión arterial media (PAM) no invasivos para cada paciente. Usaremos los registros de la siguiente consulta SQL, la cual exportamos como "mean_arterial_pressure. csv".

La consulta SQL selecciona todos los pacientes subject_ids y las mediciones no invasivas de PAM de la tabla mimic2v26. chartevents con la condición de (1) ser un adulto, (2) en su primera admisión a UCI, y (3) donde no falta hadm_id de acuerdo a la tabla mimic2v26. icustay-detail.

1. Consulta SQL:

```
SELECT subject_id, value1num
FROM mimic2v26.chartevents
WHERE subject_id IN (
SELECT subject_id
    FROM mimic2v26.icustay_detail
              WHERE subject_icustay_seq = 1
              AND icustay_age_group = 'adult'
              AND hadm_id IS NOT null)
AND itemid=456
AND value1num is not null

-- Export and save the query result as "mean_arterial_pressure.csv"
```

2. Código R:

Hay una variedad de métodos que pueden ser elegidos para agregar registros. En este caso prestaremos atención a promediar múltiples registros de PAM en un solo valor promedio de PAM para cada paciente. Otras opciones que pueden ser elegidas incluyen el uso del primer valor registrado, un mínimo o un máximo, etc.

Para un ejemplo básico, el siguiente código demuestra la reducción de datos promediando todos los múltiples registros de PAM en un único registro por paciente. El código usa la función aggregate ():

```
setwd("~/Desktop/MIMIC_data_files")
all_maps <- read.csv("mean_arterial_pressure.csv", sep = ",")

str(all_maps)

## 'data.frame':    790174 obs. of  2 variables:
##  $ subject_id: int  4 4 4 4 4 4 4 4 3 4 ...
##  $ value1num : num  80.7 71.7 74.3 69 75 ...
```

Este paso promedia los valores de PAM para cada subject-id:

```
avg_maps <- aggregate(all_maps, by=list(all_maps[,1]), FUN=mean,
na.rm=TRUE)

head(avg_maps)

##   Group.1 subject_id value1num
## 1       3          3  75.10417
## 2       4          4  88.64102
## 3       6          6  91.37357
## …
```

Ejemplos Parte 2: Remodelar el set de datos

Objetivo: Idealmente, queremos un set de datos "ordenado" reorganizado de forma tal que siga estas 3 reglas [2, 3]:

1. Cada variable forma una columna
2. Cada observación forma una fila
3. Cada valor tiene su propia celda

Los sets de datos exportados de MIMIC generalmente ya están bastante "ordenados". Por lo tanto, construiremos nuestro propio *data-frame* aquí para facilitar la demostración de la regla 3. También demostraremos como usar algunos paquetes para limpieza de datos.

Código R: Para reflejar nuestro propio *data-frame* MIMIC, construimos un set de datos con una columna de subject_id y una columna con una lista de diagnósticos para la admisión.

```
diag <- data.frame(subject_id = 1:6,   diagnosis = c("PNA, CHF", "DKA",
"DKA, UTI", "AF, CHF", "AF", "CHF"))
diag
##   subject_id diagnosis
## 1          1  PNA, CHF
## 2          2       DKA
## 3          3  DKA, UTI
## …
```

Note que el set de datos arriba no está "ordenado". Hay múltiples variables categóricas en la columna "diagnosis"----rompe la regla de datos "ordenados" número 1. Hay múltiples valores en la columna "diagnosis"----rompe la regla de datos "ordenados" número 3.

Hay muchas formas de limpiar y remodelar este set de datos. Mostraremos un modo de hacerlo usando los paquetes R "splitstackshape" [5] y "tidyr" [4] para hacer el remodelado del set de datos más fácil.

Ejemplo 1 de paquete R----"splitstackshape":

Instalar y cargar el paquete en la consola R.

```
install.packages("splitstackshape")
library(splitstackshape)
```

La función, cSplit (), puede separar los múltiples valores categóricos en cada celda de la columna “diagnosis” en diferentes columnas, “diagnosis-1” y “diagnosis-2”. Si el argumento, dirección, para cSplit () no está especificado, entonces la función separa el set de datos original a formato “ancho”.

```
diag2 <- cSplit(diag, "diagnosis", ",")
diag2

##    subject_id diagnosis_1 diagnosis_2
## 1:          1         PNA         CHF
## 2:          2         DKA          NA
## 3:          3         DKA         UTI
## …
```

Uno podría tomar esto como está si uno estuviera interesado en diagnósticos primarios y secundarios (a pesar de que no está estrictamente “ordenado” todavía).

Otra alternativa, si la dirección del argumento esta especificada como “long”, entonces cSplit separa la función “long” de la siguiente manera:

```
diag3 <- cSplit(diag, "diagnosis", ",", direction = "long")
diag3
##    subject_id diagnosis
## 1:          1       PNA
## 2:          1       CHF
## 3:          2       DKA
## …
```

Note que diag3 todavía no está “ordenado” dado que todavía hay múltiples variables categóricas bajo la columna de “diagnosis”---pero ya no tenemos múltiples valores por celda.

Ejemplo 2 de paquete R----“tidyr”:

Para “limpiar” más el set de datos, el paquete “tidyr” es bastante útil.

```
install.packages("tidyr")
library(tidyr)
```

El objetivo es separar cada variable categórica bajo la columna, “diagnosis”, en sus propias columnas siendo 1=tiene el diagnóstico y 0=no tiene el diagnóstico. Para hacer esto primero construimos una tercera

columna, “yes”, que contiene todos los valores 1 inicialmente (porque la función que vamos a usar requiere una columna de valores que corresponda con una columna de múltiples categorías que queremos “esparcir”).

```
diag3$yes <- rep(1, nrow(diag3))
diag3

##    subject_id diagnosis yes
## 1:          1       PNA   1
## 2:          1       CHF   1
## 3:          2       DKA   1
## …
```

Entonces podemos usar la función de esparcir para separar cada variable categórica en sus propias columnas. El argumento, fill=0, reemplaza los valores faltantes.

```
diag4 <- spread(diag3, diagnosis, yes, fill = 0)
diag4

##    subject_id AF CHF DKA PNA UTI
## 1:          1  0   1   0   1   0
## 2:          2  0   0   1   0   0
## 3:          3  0   0   1   0   1
## …
```

Uno puede ver que este set de datos ahora está “ordenado”, dado que sigue las tres reglas de datos “ordenados”.

12.4 Conclusión

Una variedad de temas de control de calidad son comunes cuando uno usa datos clínicos crudos recolectados para propósitos que no sean estudios. El preprocesamiento de datos es un paso importante en la preparación de los datos crudos para el análisis estadístico. Muchos pasos diferentes están involucrados en el preprocesamiento de datos crudos como fue descrito en este capítulo: limpieza, integración, transformación y reducción. A través del proceso es importante comprender las decisiones tomadas en los pasos de preprocesamiento y cómo diferentes métodos pueden impactar en la validez y aplicabilidad de resultados de estudios. En el caso de los datos de HCEs, como los de base de datos MIMIC, el preprocesamiento habitualmente requiere algo de comprensión del contexto clínico bajo el cual los datos

fueron ingresados, para guiar las elecciones realizadas. El objetivo de todos estos pasos es llegar a un set de datos “limpio” y “ordenado” adecuado para análisis estadístico efectivo evitando introducción inadvertida de sesgos en los datos.

Puntos clave

- Los datos crudos para análisis secundario son frecuentemente “desordenados”, es decir no se encuentran en una forma adecuada para el análisis estadístico; los datos deben ser “limpiados” para obtener una base de datos “ordenada” válida, completa, y efectivamente organizada para ser analizada.
- Hay una variedad de técnicas que pueden ser utilizadas para preparar los datos para su análisis, y dependiendo de qué métodos se usen, este preprocesamiento puede introducir sesgo en el estudio.
- El objetivo del preprocesamiento de datos es preparar los datos crudos disponibles para análisis, sin introducir sesgo por cambiar la información contenida en los datos o por influenciar los resultados finales de algún otro modo.

Referencias

1. Son NH (2006) Data mining course-data cleaning and data pre processing. Warsaw University. Disponible en http://www.mimuw.edu. pl/*son/datamining/DM/4-preprocess.pdf.
2. Grolemund G (2016) R for data science-data tidying. O’Reilly Media. Disponible en http://garrettgman. github. io/tidying/.

3. Wickham H (2014) J Stat Softw 59 (10). Tidy Data. Disponible en http://vita. had. co. nz/papers/tidy-data.pdf.
4. Wickham H (2016) Package 'tidyr'-easily tidy data with spread () and gather () functions. CRAN. Disponible en https://cran.rproject.org/web/packages/tidyr/tidyr.pdf.
5. Mahto A (2014) Package 'splitstackshape'-stack and reshape datasets after splitting concatenated values. CRAN. Disponible en https://cran.rproject.org/web/packages/splitstackshape/splitstackshape.pdf.

Capítulo 13

DATOS FALTANTES

Catia M. Salgado, Carlos Azevedo,
Hugo Proença y Susana M. Vieira

Objetivos de aprendizaje

- Conocer cuáles son los diferentes tipos de datos faltantes y las fuentes de las mismas.
- Conocer las opciones disponibles para tratar los datos faltantes.
- Conocer qué técnicas existen para ayudar a elegir la técnica más apropiada para una base de datos específica.

13.1 Introducción

Los datos faltantes son un problema que afecta a la mayoría de las bases de datos; las historias clínicas electrónicas (HCE) no son la excepción. Como la mayoría de los modelos estadísticos operan solamente sobre observaciones de exposición y variables de resultados completas, es necesario lidiar con los datos faltantes ya sea eliminando las observaciones incompletas o reemplazando cualquier valor faltante con un valor estimado basado en otra información disponible. Este proceso es conocido como "imputación". Ambos métodos pueden afectar en forma significativa las conclusiones que pueden ser obtenidas a partir de los datos.

Es importante identificar las causas de la falta de los datos, ya que influencian la elección de la técnica de imputación. Esquemáticamente, existen varias situaciones posibles: (i) el valor falta porque fue olvidado o perdido; (ii) el valor falta porque no era aplicable al caso; (iii) el valor falta porque no es de interés para el caso. Si tuviéramos que poner esto en un contexto médico: (i) la variable es medida, pero por alguna razón no identificable los valores no son registrados electrónicamente, por ejemplo por una desconexión de los sensores, errores en la comunicación con el servidor de la base de datos, omisión humana accidental, fallas eléctricas y otros; (ii) la variable no es medida durante un período de tiempo debido a una razón identificable, como por ejemplo que el paciente sea desconectado del ventilador por una decisión médica; (iii) la variable no es medida porque

no está relacionada con la condición del paciente y no provee información clínica útil al médico [1].

Es importante distinguir entre aquellos datos que faltan por razones identificables y aquellos que faltan por razones no identificables. En el primer caso, ingresar valores puede ser inadecuado y agregar sesgos a la base de datos, por lo que se dice que los datos son no recuperables. Por otro lado, cuando los datos faltan por razones no identificables se asume que faltan por razones aleatorias e involuntarias. Este tipo de datos faltantes se clasifican como recuperables.

La primera sección de este capítulo se enfoca en describir en forma teórica algunos de los métodos usados habitualmente para manejar los datos faltantes. Para poder demostrar las ventajas y desventajas de estos métodos, en la segunda parte del capítulo se demuestra su aplicación en bases de datos reales que fueron creadas para estudiar la relación entre mortalidad y la inserción de catéter arterial invasivo (CAI) en la unidad de cuidados intensivos.

13.2 Parte 1 - Conceptos Teóricos

La preparación de los datos es la tarea más crucial y que más tiempo consume, en el descubrimiento de conocimientos de las bases de datos, influenciando fuertemente el éxito de una investigación. La selección de las variables consiste en identificar un conjunto de predictores potenciales útiles de una gran cantidad de candidatos (por favor remitirse al Capítulo 5 – Análisis de datos, para más información sobre selección de características). Rechazar variables con un excesivo número de valores faltantes (por ejemplo > 50%) suele ser una buena regla de oro. Sin embargo, no es un procedimiento libre de riesgo. Rechazar una variable puede llevar a una pérdida de potencia predictiva y de la capacidad de detectar diferencias estadísticamente significativas. Además, puede ser una fuente de sesgos que afecte la representatividad de los resultados. Por estas razones, la selección de variables debe adecuarse al mecanismo de los datos faltantes. La imputación puede ser realizada antes y/o después de la selección de variables.

Los pasos generales que deben seguirse para manejar datos faltantes son:

- Identificar patrones y razones para los datos faltantes;
- Analizar la proporción de los datos faltantes;

- Elegir el mejor método de imputación.

13.2.1 Tipos de Faltantes

Los mecanismos por los cuales faltan los datos afectarán algunos supuestos que respaldan nuestros métodos de imputación. Es posible describir tres grandes mecanismos que generan datos faltantes, dependiendo de la relación entre los datos observados (disponibles) y no observados (faltantes).

Para simplificar, consideremos faltantes en casos univariados. Para definirlo en términos matemáticos, una base de datos puede ser dividida en dos partes:

$$X = \{X_o, X_m\}$$

donde X_o corresponde al conjunto datos observados y X_m al conjunto de datos faltantes en la base de datos.

Para cada observación definimos una respuesta binaria según dicha observación sea faltante o no:

$$R = \begin{cases} 1 & \text{Si } X \text{ se observa} \\ 0 & \text{Si } X \text{ es faltante} \end{cases}$$

El mecanismo del valor faltante puede ser entendido en términos de la probabilidad de que una observación falte Pr (R) dadas las observaciones observadas y/o faltantes en la forma:

$$\Pr(R|x_o, x_m)$$

Los tres mecanismos están sujetos a cómo la probabilidad de respuesta R depende o no de los valores observados y/o faltantes:

- **Datos faltantes en forma completamente aleatoria (MCAR, del inglés Missing Completely at Random) –** Cuando las observaciones faltantes dependen de las medidas observadas y no observadas. En este caso, la probabilidad de que una observación falte depende únicamente de sí misma y reduce Pr (R|x_o, x_m) = Pr (R). Como ejemplo, imagine que un médico olvida registrar el género de uno de cada 6 pacientes que

ingresan a la UCI. No hay ningún mecanismo oculto relacionado con ninguna variable y no depende de ninguna característica de los pacientes.

- **Datos faltantes en forma aleatoria (MAR, del inglés Missing at Random) –** En este caso la probabilidad de que un valor falte está relacionada únicamente con los datos observables, esto es, los datos observados están estadísticamente relacionados con los datos faltantes y es posible estimar los faltantes a partir de los observados. En este caso no es completamente "aleatorio", pero es el caso más general donde podemos ignorar el mecanismo por el cual se producen los faltantes, en la medida que controlamos la información de la que dependen las faltantes: los datos observados. Dicho de otra manera, la probabilidad de que un dato falte para una variable no depende de los valores de esa variable, luego de ajustar para valores observados. Matemáticamente la probabilidad de falta se reduce a $Pr(R|x_o, x_m) = Pr(R|x_o)$. Imagine que si las personas ancianas son menos propensas a informar al médico el haber tenido neumonía, la tasa de respuesta de la variable neumonía dependerá de la variable edad.

- **Datos faltantes en forma no aleatoria (MNAR, del inglés Missing Not at Random) –** Esto refiere al caso en el que no aplican los dos casos anteriores. Los datos faltantes dependen tanto de los valores faltantes como de los observados. Determinar el mecanismo subyacente suele ser imposible ya que depende de datos no observados. De allí deriva la importancia de realizar análisis de sensibilidad y probar cómo las inferencias se sostienen bajo diferentes supuestos. Por ejemplo, podemos imaginar que los pacientes con baja presión más probablemente tengan su presión medida menos veces (el dato faltante para la variable presión arterial depende parcialmente de los valores de la presión arterial).

13.2.2 Proporción de datos faltantes

El porcentaje de datos faltantes para cada variable (entre pacientes) y cada paciente (entre variables) debe ser computado para ayudar a decidir qué variables y/o pacientes deben ser considerados candidatos para la eliminación o imputación de datos. Un ejemplo crudo se muestra en la Tabla

13.1 donde quizás querríamos considerar eliminar al paciente 1 y la variable "AST" del análisis, considerando que la mayoría de sus valores faltan.

Tabla 13.1 Ejemplos de datos faltantes en la HCE

	Género	Glucosa	AST	Edad
Paciente 1	?	120	?	?
Paciente 2	M	105	?	68
Paciente 3	F	203	45	63
Paciente 4	M	145	?	42
Paciente 5	M	89	?	80

13.2.3 Lidiando con los datos faltantes

Revisión de los métodos para manejar datos faltantes

Los métodos deberían ser adecuados a la medida del set de datos seleccionado, a las causas que explican los datos faltantes y a la proporción de datos faltantes. En general, un método se elige por su simplicidad y su capacidad de introducir el menor sesgo posible en la base de datos.

Cuando los datos son MCAR o MAR el investigador puede ignorar las razones de los datos faltantes, lo que simplifica la elección del método a aplicar. En este caso, puede aplicarse cualquier método. Aun así es difícil obtener evidencia empírica acerca de si los datos son o no MCAR o MAR. Una estrategia válida es examinar la sensibilidad de resultados a los supuestos de MCAR y MAR comparando varios análisis, en donde las diferencias en los resultados a través de los diversos análisis pueden proveer alguna información acerca de qué supuesto puede ser el más relevante.

Una gran cantidad de evidencia se ha focalizado en comparar el rendimiento de los métodos de manejo de datos faltantes, tanto en general [2-4] como en el contexto de factores específicos como la proporción de datos faltantes y el tamaño de la muestra [5-7]. En los trabajos de Jones y Little, pueden encontrarse aspectos técnicos más detallados y la aplicación de esos métodos en varios campos [8, 9].

En resumen, los métodos más utilizados pueden clasificarse en tres categorías generales que son descritas en más detalle a continuación.

1. Métodos de eliminación (eliminación por listas, por ejemplo análisis de casos completos; eliminación por pares, por ejemplo análisis de casos disponibles)
2. Métodos de imputación individual (sustitución por la media/moda, interpolación linear, *hot deck* y *cold deck*)
3. Métodos basados en modelos [regresión, imputación múltiple, *k* vecinos más cercanos (kNN, del inglés *k-nearest neighbours*)]

Métodos de eliminación

La manera más simple de lidiar con los datos faltantes es descartar los casos u observaciones que tengan valores faltantes. En general, los métodos de eliminación de casos llevan a inferencias válidas únicamente para MCAR [10]. Hay tres maneras de hacer esto: análisis de casos completos; análisis de casos disponibles; y métodos de ponderación.

Análisis de casos completos (eliminación por listas)

En un análisis de casos completos todas las observaciones con al menos un valor faltante son eliminadas (Fig. 13.1).

El supuesto principal es que la submuestra restante es representativa de la población y por lo tanto no introducirá sesgos en el análisis hacia un cierto subgrupo. Este supuesto es bastante restrictivo y asume un mecanismo MCAR. La eliminación por listas a menudo produce pendientes de regresión estimadas sin sesgos, siempre y cuando los faltantes no sean una función de la variable resultado. La mayor ventaja de este método es su simplicidad y siempre es razonable utilizarlo cuando el número de observaciones descartadas es relativamente pequeño en comparación al total. Sus principales desventajas son la reducción de potencia estadística (ya que reduce el número de muestras *n*, las estimaciones tendrán un mayor error estándar), la pérdida de información y el posible sesgo del análisis especialmente si los datos no son MCAR.

Fig 13.1 Ejemplo de eliminación de casos completos. Los casos resaltados en rojos son eliminados.

Género	Glucosa	Edad
M	?	65
F	120	71
F	99	?
F	140	52
M	88	?
F	85	63
M	170	68
?	153	80
M	115	59
F	103	?

Análisis de casos disponibles

El método de casos disponibles descarta únicamente aquellos datos en las variables que son necesarias para un determinado análisis. Por ejemplo, si solo 4 de 20 variables son necesarias para un estudio, este método descartaría únicamente las observaciones faltantes de las 4 variables de interés. En la Fig. 13.2, imagine que cada una de las tres variables representadas son utilizadas para análisis distintos. El análisis es realizado utilizando todos los casos en los que la variable de interés esté presente. Más allá de que este método tiene la capacidad de preservar más información, las poblaciones de cada análisis serían diferentes y posiblemente no comparables.

Fig. 13.2 Ejemplo de la eliminación de caso disponible. Si cada variable es utilizada para análisis separados, solo se descartan los casos en los falta la variable de interés.

Estudio de Caso		
S1	S2	S3
Género	Glucosa	Edad
M	?	65
F	120	71
F	99	?
F	140	52
M	88	?
F	85	63
M	170	68
?	153	80
M	115	59
F	103	?

Análisis de ponderación de casos

La ponderación es una forma de ponderar los casos completos modelando los faltantes para reducir el sesgo introducido en el caso disponible.

Métodos de imputación de valor único

En la imputación simple, los valores faltantes son completados por algún tipo de variable "predicha" [9, 11]. La imputación simple ignora la incerteza y casi siempre subestima la varianza. La imputación múltiple supera este problema teniendo en cuenta tanto la incerteza "entre" como "dentro" de los valores imputados.

Media y mediana

El método de imputación más simple es sustituir los valores faltantes por la media o la mediana de la variable. Utilizar la mediana es más sólido ante la presencia de valores atípicos en los datos observados. Las principales desventajas son (1) que reduce la variabilidad, disminuyendo los errores estimados en comparación a los métodos de eliminación y (2) que desatiende la relación entre variables, disminuyendo por lo tanto su correlación. Mientras que este método disminuye el sesgo de utilizar una muestra no representativa, introduce otros sesgos.

Interpolación lineal

Este método es particularmente adecuado para series de tiempo. En la interpolación lineal, un valor faltante se computa interpolando los valores de la medición previa y posterior disponibles para el paciente. Por ejemplo, si la natremia cambia de 132 a 136 mEq/L en 8 hs, uno podría razonablemente asumir que su valor estuvo cerca de 134 mEq/L en el punto medio

Hot deck y cold deck

En el método *Hot Deck,* un valor faltante de un atributo es reemplazado con un valor que proviene de una distribución estimada de los datos actuales. Se utiliza especialmente en la investigación con encuestas [9]. *Hot deck* suele implementarse en dos etapas. Primero los datos son particionados en grupos y luego cada caso con datos faltantes es asociado con un grupo. Los casos completos en un grupo se utilizan para completar los datos faltantes. Esto puede hacerse calculando la media o modo del atributo dentro del grupo. La imputación C*old Deck* es similar a la *Hot Deck*, salvo porque la fuente de datos es diferente al set de datos actual. La imputación *Hot-Deck* reemplaza el valor faltante por valores realistas que conservan la distribución de la variable. Sin embargo, subestima los errores estándares y la variabilidad [12].

Última observación realizada

A veces llamado el método "muestreo-y-retención" [13]. Este método es específico para diseños longitudinales. Esta técnica imputa el valor faltante con la última observación disponible del individuo. Este método asume que la observación del individuo no ha cambiado desde la última observación medida, lo cual es poco realista [14].

Imputación basada en modelos

En la imputación basada en modelos, se crea un modelo predictivo para estimar los valores que sustituirán los datos faltantes. En este caso, el conjunto de datos es dividido en dos subconjuntos: uno que no tiene valores faltantes para la variable bajo evaluación (utilizado para entrenar el modelo) y uno conteniendo los valores faltantes que buscamos estimar. Se pueden utilizar varios métodos de modelado como: regresión, regresión logística, redes neurales y otras técnicas de modelado paramétricos y no

paramétricos. Hay dos grandes desventajas en este abordaje: el modelo estima valores que usualmente se comportan mejor que los valores verdaderos y los modelos tienen un pobre comportamiento si la variable observada y la faltante son independientes.

Regresión lineal

En este modelo, todas las variables disponibles son utilizadas para crear un modelo de regresión lineal que utilice los valores disponibles de la variable de interés como *resultado*. La ventaja de este método es que considera la relación entre variables, a diferencia de la imputación por la media o la mediana. Las desventajas son que sobreestima la capacidad del modelo y la correlación entre variables, ya que no considera la incerteza en los datos faltantes y subestima varianzas y covarianzas. Un método que fue creado para introducir la incerteza es la regresión lineal estocástica (ver abajo).

El caso de la imputación multivariada es más complejo ya que existen valores faltantes para diversas variables que no siguen el mismo patrón de sus faltantes a través de las observaciones. El método utilizado es la extensión multivariada del modelo lineal y depende de un proceso iterativo realizado hasta la convergencia.

Regresión estocástica

La imputación por regresión estocástica apunta a reducir el sesgo a través de un paso adicional que aumenta cada valor predicho con un factor residual. Este factor residual se distribuye normalmente, con un promedio de cero y una varianza igual a la varianza residual de la regresión del predictor sobre el objetivo. Este método permite preservar la variabilidad en los datos y estimar parámetros sin sesgos con datos MAR. Sin embargo, el error estándar tiende a ser subestimado porque la incerteza acerca de los valores imputados no se incluye, lo que incrementa el riesgo de error de tipo I [15].

Imputación de valor múltiple

La imputación múltiple (IM) es una técnica estadística poderosa desarrollada por Rubin en los años ‘70 para analizar set de datos que contienen valores faltantes [7, 16]. Es una técnica Monte Carlo que requiere 3 pasos (Fig 13.3).

- Imputación: donde los valores faltantes son completados utilizando cualquier método de elección, llevando a un conjunto de datos M>=2 (5-10 suele ser suficiente) [10]. En estos set de datos imputados por múltiplos de M, todos los valores observados son los mismos, pero los valores imputados son diferentes reflejando la incerteza de la imputación [10].
- Análisis: cada set de datos M completado es analizado (por ejemplo, se construye una regresión logística clasificadora para predicción de mortalidad), que da M análisis.
- Agregación: los M análisis son integrados en un resultado final, por ejemplo computando la media (y el IC 95%) de los M análisis.

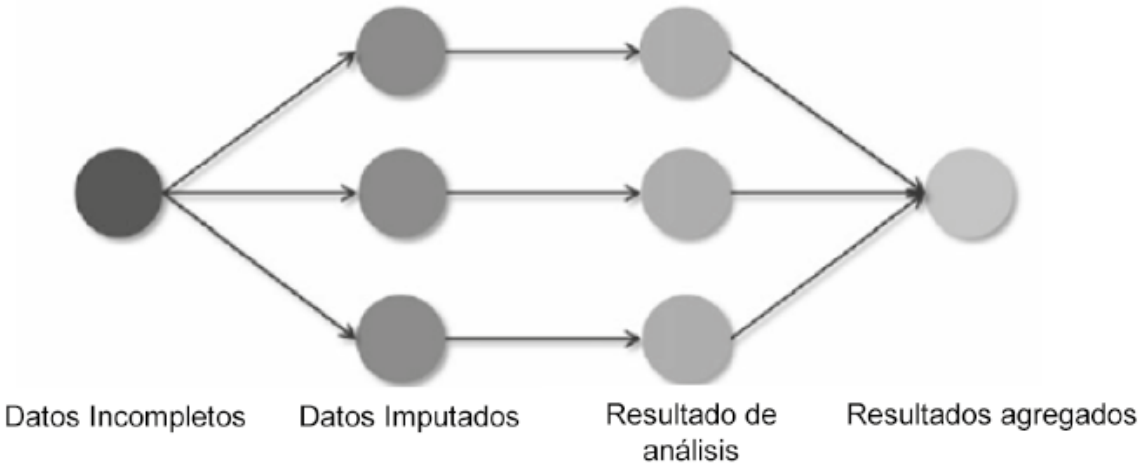

Fig. 13.3 El concepto de imputación múltiple con M=3.

K-Nearest Neighbors

La imputación por vecinos más cercanos (kNN, del inglés *K-Nearest Neighbors*) puede ser utilizada para tratar los valores faltantes. Aquí serán completados con la media de los *k* valores que provengan de las *k* observaciones más parecidas. La similitud de dos observaciones es determinada, luego de la normalización del conjunto de datos, utilizando una función de distancia que puede ser euclideana, Manhattan, Mahalanobis, Pearson, etc. La principal ventaja del algoritmo kNN es que dados suficientes datos puede predecir con una precisión razonable la probabilidad de distribución condicional alrededor de un punto y por lo tanto hacer estimaciones bien informadas. Puede predecir atributos cualitativos y

cuantitativos (discretos y continuos). Otra ventaja de este método es que tiene en consideración la estructura de correlación de los datos. La elección del valor *k* es crítica. Un valor más alto de *k* incluiría atributos que son significativamente diferentes de nuestra observación objetivo, mientras que valores inferiores implican dejar afuera atributos significativos.

13.2.4 La elección del mejor método de imputación

Se espera que los diferentes métodos de imputación funcionen de distinta forma en los diversos set de datos. Describiremos aquí un método simple y genérico que puede ser utilizado para evaluar el desempeño de distintos métodos de imputación en su propio set de datos para ayudar a seleccionar el método más apropiado. Debemos hacer notar que este abordaje simple no evalúa el efecto de los métodos de eliminación. En el caso de estudio presentado a continuación, se describe un abordaje más complejo en el cual se prueba el desempeño de un modelo predictivo en el set de datos completado a través de diferentes métodos de imputación.

Así es como se debe proceder:

1. Utilice una muestra de su set de datos que no contenga datos faltantes (servirá de verdad fundamental).
2. Introduzca proporciones crecientes de datos faltantes aleatoriamente (ejemplo 5-50% en incrementos de 5%).
3. Reconstruya los datos faltantes utilizando los distintos métodos.
4. Compute la suma de los errores al cuadrado (SEC) entre los datos reconstruidos y los datos originales para cada método y cada proporción de datos faltantes.
5. Repita los pasos 1-4 una determinada cantidad de veces (10 veces, por ejemplo) y compute el desempeño promedio de cada método (media SEC).
6. Grafique el promedio de estimación cuadrada de errores en función de la proporción de datos faltantes (un gráfico por cada método de imputación), de manera similar al ejemplo mostrado en la Fig. 13.4.

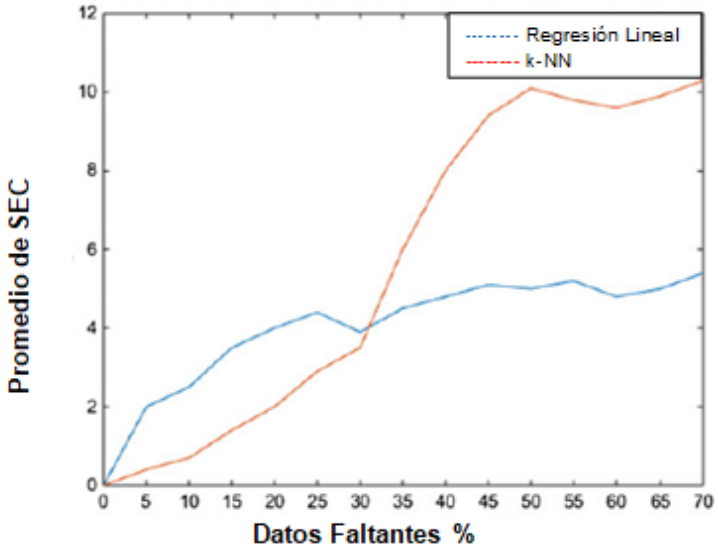

Fig. 13.4 Promedio de la suma de errores al cuadrado (SEC) entre los datos originales y reconstruidos, para varios niveles de datos faltantes y 2 métodos de imputación (los datos tienen únicamente fines ilustrativos).

7. Elija el método que se desempeñe mejor con el nivel de datos faltantes en su set de datos. Por ejemplo, si sus datos tuvieran 10% de datos faltantes, usted querría elegir k-NN; al 40% la regresión lineal actúa mejor (datos inventados únicamente con propósito ilustrativo).

13.3 Parte 2 - Caso de Estudio

En esta sección, aplicaremos varios métodos de imputación a set de datos clínicos del "mundo real" utilizados en un estudio que investigó el efecto que tiene la colocación de un catéter arterial invasivo (CAI) en pacientes con falla respiratoria. Se utilizarán 2 set de datos que incluyen pacientes que recibieron un CAI (grupo CAI) y pacientes que no lo recibieron (grupo no-CAI). Cada set de datos es subdividido en 2 grupos, correspondiendo el grupo 1 a pacientes que fallecieron dentro de los primeros 28 días y el grupo 0 a los sobrevivientes. En primer lugar se discuten la proporción de datos faltantes y las razones potenciales detrás de los mismos. Luego se realizaron los siguientes análisis:

1. Se insertaron distintas proporciones de datos faltantes aleatoriamente en la variable "edad" y luego fueron imputados utilizando los métodos descritos anteriormente. La distribución de las observaciones imputadas se comparó con la distribución original para todos los métodos.

2. Se probó el desempeño de los set de datos imputados con diferentes proporciones de faltantes en un modelo predictivo (regresión logística para predecir mortalidad), primero para datos faltantes univariados (la variable edad), luego para todas las variables (multivariados).

El código utilizado para generar los análisis y las figuras se provee en el documento acompañante de funciones R.

13.3.1 Proporción de datos faltantes y sus posibles causas

La tabla 13.2 muestra la proporción de datos faltantes en algunas de las variables de los set de datos. El subconjunto considerado para probar los diferentes métodos de imputación está representado por 26 variables y fueron seleccionadas sobre la base de que los datos faltantes de estas variables son recuperables.

Debido a que el catéter arterial permanente se utiliza principalmente para monitoreo hemodinámico y para la obtención de muestras de sangre arterial para análisis de gases en sangre podemos esperar un mayor porcentaje de datos faltantes en las variables relacionadas con los gases en sangre en el grupo no-CAI. Además podemos esperar que los diagnósticos de los pacientes brinden explicaciones a la falta de algunos resultados de laboratorio específicos: si una prueba no es pedida porque lo más probable es que no provea información clínica relevante, habrá un valor faltante; tiene sentido estimar que dicho valor se encuentra dentro de un rango normal. En ambos casos, el hecho de que falten datos contiene información sobre la respuesta, por lo que es MNAR. El índice de masa corporal (IMC) tiene un porcentaje de datos faltantes relativamente alto. Asumiendo que esta variable es calculada automáticamente a partir de la altura y peso de los pacientes podemos concluir que este dato es MAR: debido a que faltan el peso y/o la altura, el IMC no puede ser calculado. Si el peso falta porque alguien se olvidó de introducirlo en el sistema entonces es MCAR. Más allá del mecanismo por el cual falta el dato, también es importante considerar la distribución de la muestra en cada variable, ya que algunos métodos de imputación asumen distribuciones específicas de los datos, generalmente la distribución normal.

Tabla 13.2 Datos faltantes en algunas de las variables de los grupos CAI y no-CAI.

	CAI		No-CAI	
	# puntos	%	# puntos	%
Línea arterial durante el día	0	0	792	100
Estadía hospitalaria	0	0	0	0
Edad	0	0	0	0
Género	0	0	0	0
Primer peso	39	3.96	71	8.96
Primer SOFA	2	0.20	4	0.51
Primer Hemoglobina	2	0.20	5	0.63
Primer Bilirrubina	418	42.48	365	46.09

13.3.2 Análisis de datos faltantes univariados

En esta sección, se explorará la influencia específica de cada método de imputación para la variable edad, utilizando todas las otras variables. Se introdujeron artificialmente dos niveles de faltantes (20 y 40%) en los set de datos. El set de datos original representa la verdad de base contra la que se comparó el set de datos imputados utilizando histogramas de frecuencia.

Análisis de casos completos

El método de análisis de casos completos descarta todas las observaciones incompletas con al menos un valor faltante. La distribución del conjunto de datos "imputado" será igual al conjunto de datos original sin las observaciones que tengan valores faltantes en la variable edad.

La figura 13.5 muestra un ejemplo de la distribución de la variable edad en el grupo CAI.

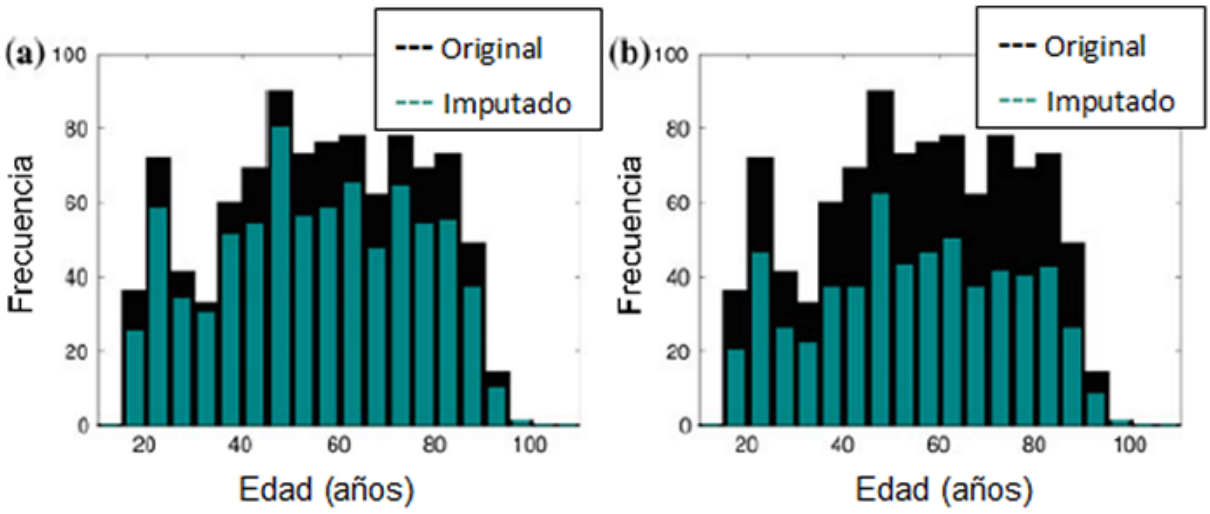

Fig. 13.5 Histograma de la variable edad en el grupo CAI antes y después del método univariado de análisis de casos completos.

Este método solamente es útil cuando hay un porcentaje pequeño de datos faltantes. El mismo no requiere ninguna presunción sobre la distribución de los datos faltantes, más allá de que los casos completos deben ser representativos de la población original, lo que es difícil de probar.

Imputación de valor único

Imputación de media y mediana

Estos métodos constituyen técnicas de imputación muy crudas que ignoran la relación entre la edad y las otras variables y que introducen un sesgo importante hacia los valores medios o medianos. Estos métodos simples nos permiten entender el efecto del sesgo, cosa que se puede ver en los ejemplos de la Fig. 13.6.

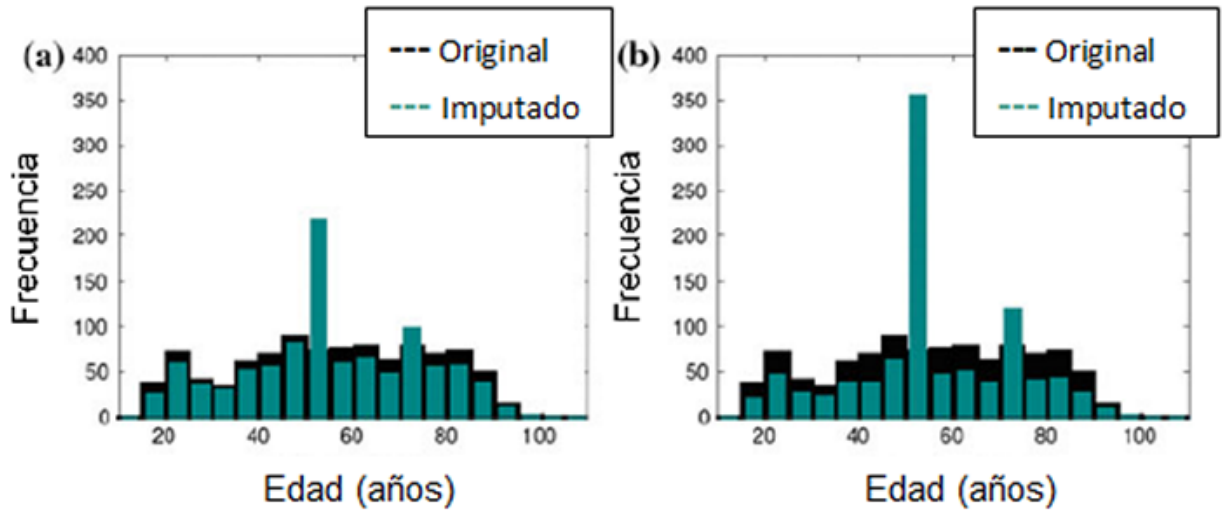

Fig. 13.6 Histograma de la variable edad en el grupo CAI, antes (original) y después (imputado) para la imputación univariada con la media.

Imputación por regresión lineal

El método de regresión lineal imputa la mayoría de los datos al centro de la distribución (ejemplo en la Fig. 13.7). Los extremos de la distribución no están bien modelados y son fácilmente ignorados. Esto se debe a dos características de la técnica: en primer lugar, el supuesto de que los datos faltantes caen dentro de la línea de regresión, transformando la realidad para que se ajuste a la naturaleza determinística del modelo. En comparación con la imputación por la media/mediana, la regresión lineal asume que existe una correlación entre las variables; sin embargo, sobreestima esta relación al asumir que los puntos faltantes están sobre la curva de regresión. El modelo asume que el porcentaje de la varianza explicada es el 100%, en consecuencia subestima la variabilidad.

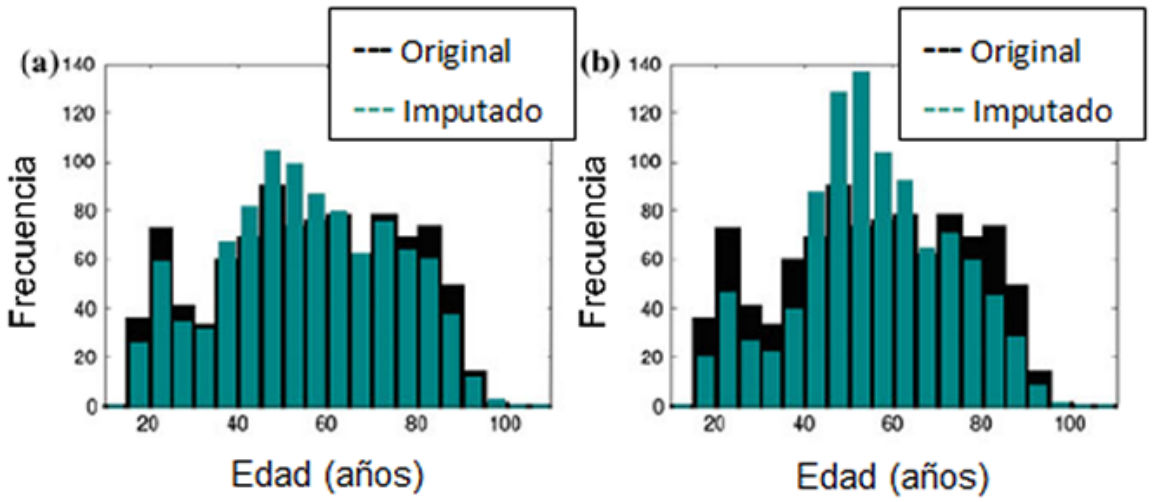

Fig. 13.7 Histograma de la variable edad en el grupo CAI antes (original) y después (imputado) de la imputación univariada por regresión lineal.

Imputación por regresión lineal estocástica

La regresión lineal estocástica es un intento de abandonar las presunciones determinísticas de la regresión lineal. En este caso, la distribución de los datos imputados se asemeja más a los datos originales que en los métodos anteriores (Fig. 13.8). Este método puede introducir valores imposibles, como edades negativas. Es un primer paso para modelar la incertidumbre presente en el conjunto de datos que representa una negociación entre la precisión de los valores y la incerteza introducida por los datos faltantes.

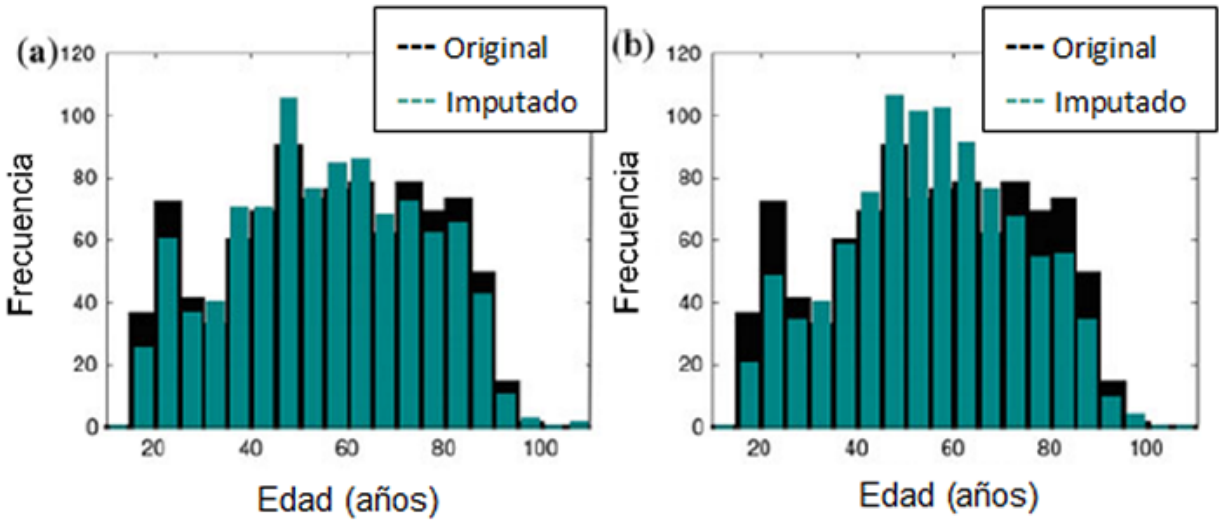

Fig. 13.8 Histograma de la variable edad en el grupo CAI antes (original) y después (imputado) de la imputación univariada por regresión lineal estocástica.

K Nearest Neighbors

Limitaremos la demostración al caso en que k=1. En el caso extremo en que todos los vecinos son usados sin ponderar, este método converge a la imputación media.

La figura 13.9 demuestra que este método introduce en nuestro set de datos particular un gran sesgo hacia el valor central. La razón de esto surge del hecho de que casi la mitad de las variables son binarias, lo que termina teniendo un peso mucho mayor en las distancias que las variables continuas (que siempre son menores a 1, debido a la normalización unitaria realizada en el preprocesamiento de los datos). Los cálculos con kNN incrementan la calidad con el número de observaciones en el set de datos y sin duda se trata de un método muy poderoso en las condiciones adecuadas.

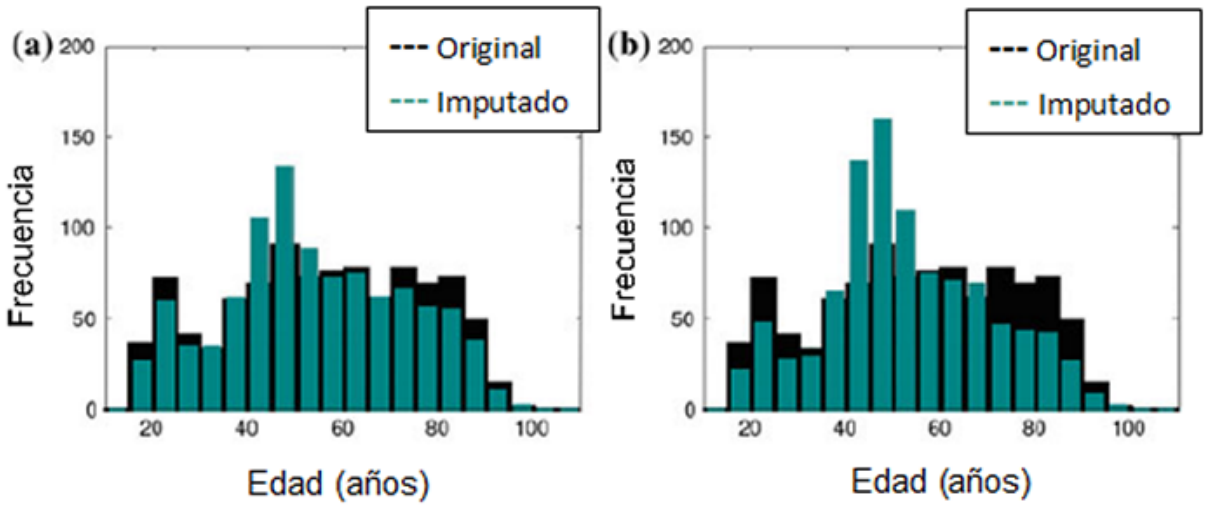

Fig. 13.9 Histograma de la variable edad en el grupo CAI KNN antes (original) y después (imputado) para la imputación univariada.

Imputación múltiple

La imputación múltiple con regresión lineal y la regresión multivariada normal son extensiones de los métodos de imputación única de mismo nombre y utilizan el muestreo para crear distintos set de datos que representen las distintas posibilidades de lo que podría ser el set de datos original. Estos métodos permiten un mejor modelado de la incertidumbre

presente en los valores faltantes y, por lo general, son más sólidos en términos de resultados y propiedades estadísticas. Elegimos trabajar con 10 set de datos, que fueron promediados para que la representación gráfica luciera similar a la de los métodos anteriores.

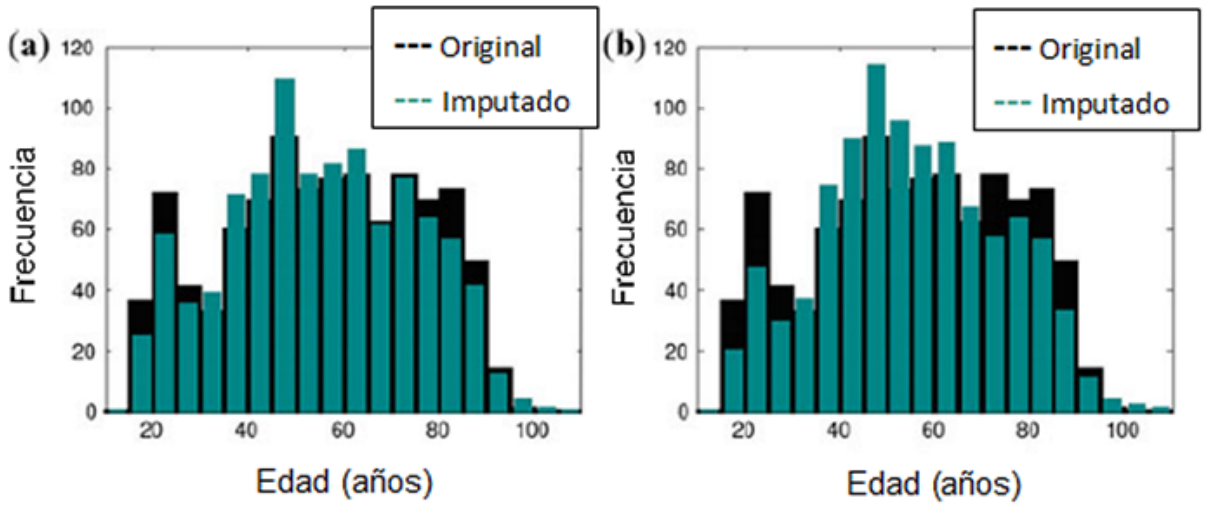

Fig. 13.10 Histograma de la variable edad en el grupo CAI antes (original) y después (imputado) de la imputación múltiple por regresión multivariada normal para la imputación univariada.

Regresión multivariada normal

La regresión multivariada normal con imputación múltiple dio mayor importancia a los valores del centro de la distribución (Fig. 13.10). El supuesto principal de este método es que los datos siguen una distribución multivariada normal, cosa que no es completamente verdad para este set de datos que contiene numerosas variables binarias. Más allá de eso, incluso en presencia de variables categóricas y distribuciones que no son estrictamente normales, debería funcionar razonablemente bien [10,19]. El método de imputación múltiple mejora el modelado de la incertidumbre agregando un muestreo tipo 'bootstrap' al algoritmo de maximización de expectativas, otorgando un incremento a mejores predicciones de los posibles datos faltantes al considerar múltiples posibilidades de los datos originales. Obviamente, cuando se promedian los datos para su representación en el histograma parte de esa riqueza se pierde. Sin embargo, la calidad de la regresión es obvia cuando se compara con los métodos previos.

Regresión lineal

La imputación múltiple por el método de regresión lineal utiliza todas las variables excepto, la variable objetivo (edad) para estimar los datos faltantes de esta última variable. Los datos son modelados utilizando regresión lineal y muestreo de Gibbs. La figura 13.11 demuestra que esto representa de lejos el método de imputación más preciso para este set de datos en particular.

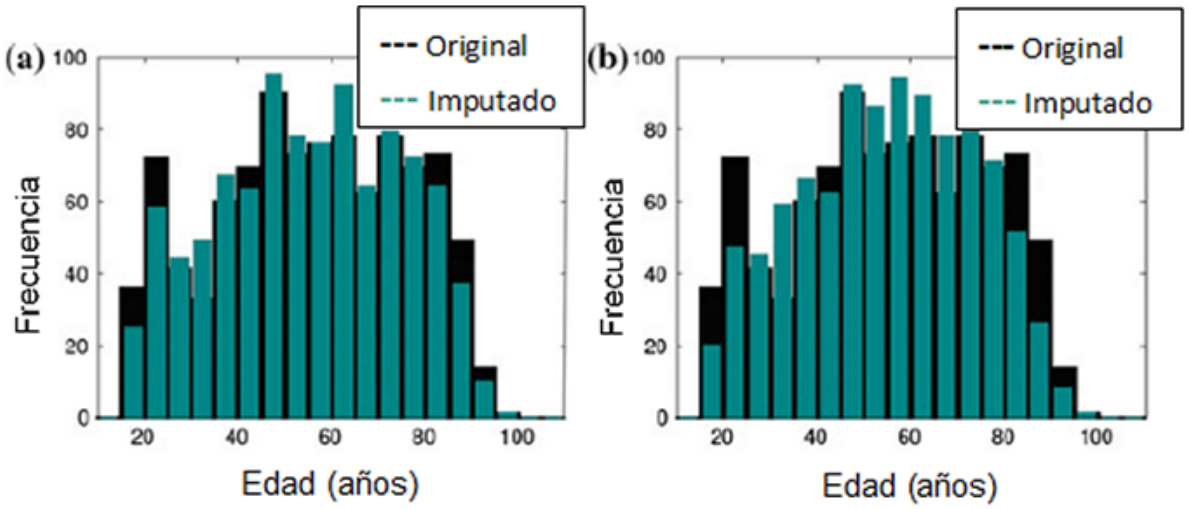

Fig. 13.11 Histograma de la variable edad en el grupo CAI antes (original) y después (imputado) de la imputación múltiple con regresión generalizada para imputación univariada.

13.3.3 Evaluando el desempeño de los métodos de imputación en la predicción de mortalidad

Esta prueba apunta a evaluar la capacidad de generalización de los modelos construidos utilizando datos imputados y verificar su desempeño al compararlos con los datos originales. Todos los métodos descritos previamente fueron utilizados para reconstruir una muestra tanto de set de datos CAI y no-CAI con proporciones crecientes de datos faltantes aleatoriamente, primero solo de la variable edad (univariado) y luego de todas las variables del set de datos (multivariado). Se aplicó un modelo de regresión logística en los datos reconstruidos y se probó en una muestra de los datos originales (que no contiene imputaciones o datos faltantes).

El desempeño del modelo es evaluado en términos de área bajo la curva de la Característica Operativa del Receptor (AUC, del inglés Area Under the Curve), precisión o accuracy (tasa de clasificación correcta), sensibilidad (tasa de verdaderos positivos), especificidad (tasa de verdaderos negativos) y

kappa de Cohen. Todos los métodos se compararon con una regresión logística de referencia que fue construida con los datos originales, sin faltantes. Los resultados se promediaron sobre una validación cruzada de 10 iteraciones y se presentan los resultados del AUC en forma gráfica.

La influencia de una variable tiene un efecto limitado, aún si la edad es la variable que más se correlaciona con mortalidad (Fig. 13.12). Como mucho, el AUC disminuyó de 0,84 a 0,81 para el grupo CAI y de 0,90 a 0,87 para el no-CAI, si excluimos el método de análisis de casos completos que se desempeña pobremente desde un principio. Para valores bajos de faltantes (menos del 50%), todos los otros modelos actúan de manera similar. Entre las técnicas univariadas, los métodos que se desempeñaron mejor en ambos set de datos son los dos métodos de imputación múltiple, a saber, el de regresión lineal y el de regresión multivariada normal, y el algoritmo de kNN. En el caso de faltantes univariadas, el kNN se muestra como un buen estimador si existen varias observaciones completas, como es el caso. Con el incremento de los datos faltantes, los métodos más simples introdujeron mayores sesgos en el modelado del set de datos.

La calidad de los métodos de imputación también se evaluó en presencia de faltantes multivariadas con una probabilidad uniforme en todas las variables (Fig. 13.13).

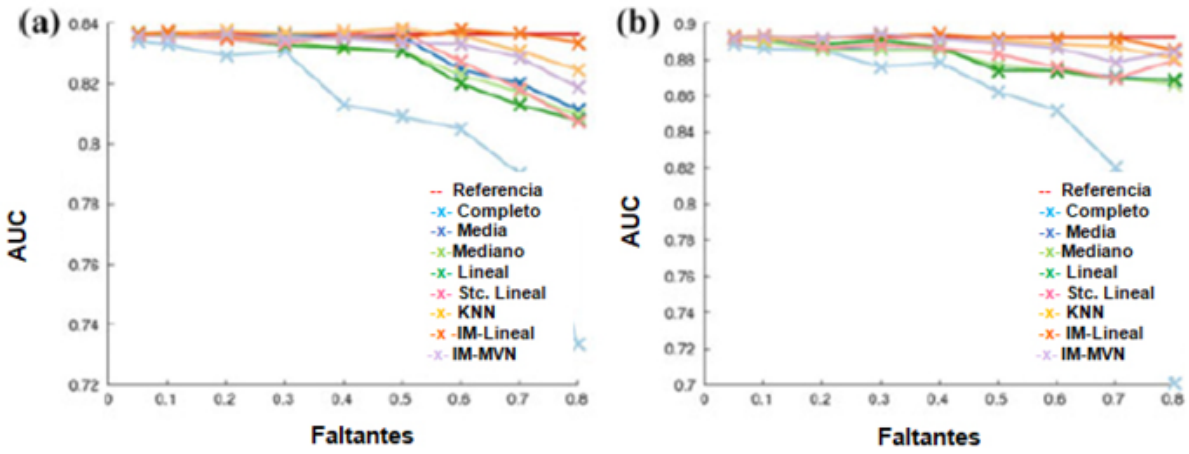

Fig. 13.12 Desempeño promedio del AUC de los modelos de regresión logística modelados con diferentes métodos de imputación para distintos grados de datos faltantes univariados de la variable edad

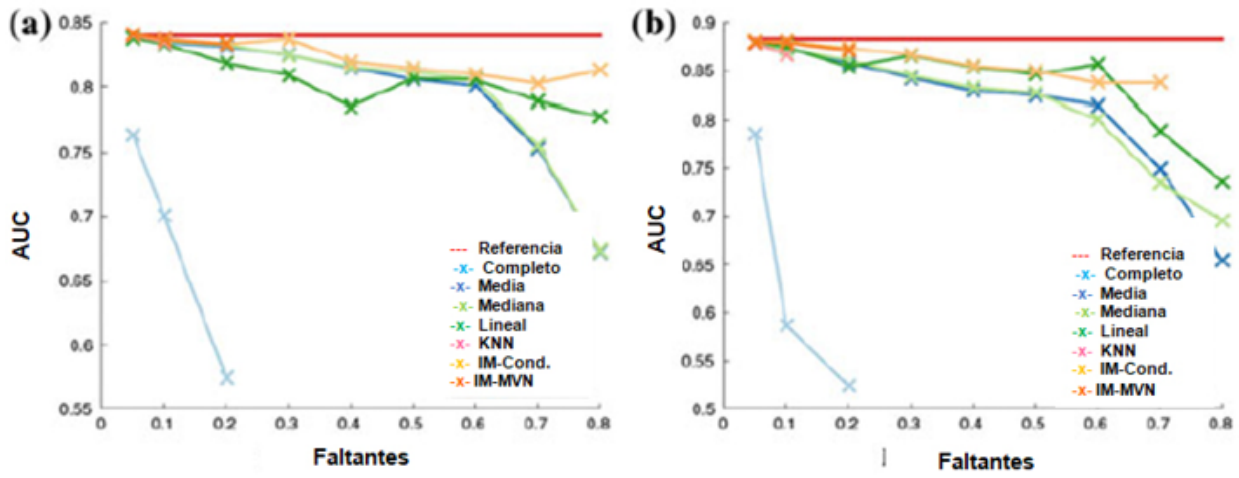

Fig. 13.13 AUC promedio de los modelos de regresión logística para distintos grados de faltantes multivariados

Se ha visto que obtener resultados para más del 40% de faltantes en todas las variables es bastante inviable en la mayoría de los casos y no hay garantías de buen desempeño con ninguno de los métodos. Algunos métodos no fueron capaces de realizar imputaciones completas sobre una gran cantidad de faltantes (por ejemplo, el análisis de casos completos dejó de tener suficientes observaciones luego de un 20% de faltantes).

En suma, y de manera sorprendente, los métodos tuvieron un desempeño aceptable incluso ante faltantes del 80% en cada variable. La razón detrás de esto es que casi la mitad de las variables son binarias y debido a su relación con el resultado; reconstruirlas a partir de los valores frecuentes en cada grupo suele ser el mejor camino. La disminución en el AUC se debió a una disminución de la sensibilidad, a la vez que los valores de especificidad se mantuvieron más o menos inalterados con los incrementos de faltantes. El método que mejor actuó en términos del AUC fue la imputación múltiple por regresión lineal. En el grupo CAI logró un valor mínimo de AUC de 0,81 al 70% de los datos faltantes, correspondiendo a un valor de referencia de AUC de 0,84, y en el grupo no-CAI logró un valor de AUC de 0,85 al 70% de datos faltantes, cercano al valor de referencia de AUC de 0,89.

13.4 Conclusiones

Los datos faltantes son un problema muy frecuente en las HCE debido a la naturaleza propia de la información médica, la gran cantidad de datos

recolectados, la heterogeneidad de los estándares de datos e instrumentos de grabación, la transferencia y conversión de datos y, por último, a las omisiones y errores humanos. Cuando se lidia con el problema de los datos faltantes, de la misma manera que en otros dominios de la minería de datos, no hay un abordaje que sirva para todas las situaciones y los científicos de datos deberían en última instancia contar con herramientas de evaluación robustas a la hora de elegir un método de imputación con el que manejar los valores faltantes de un set de datos particular.

Puntos clave

- Evalúe siempre las razones detrás de los faltantes: ¿es MCAR/MAR/MNAR?
- ¿Cuál es la proporción de datos faltantes por variable y por registro?
- Los abordajes de imputación múltiple suelen actuar mejor que otros métodos.
- Las herramientas de evaluación deben ser utilizadas para ajustar los métodos de imputación en un set de datos particular.

Nota: Las imágenes de este capítulo se encuentran disponibles a color en el ebook; disponible en www.hardineros.com

Referencias

1. Cismondi F, Fialho AS, Vieira SM, Reti SR, Sousa JMC, Finkelstein SN (2013) Missing datain medical databases: impute, delete or classify? Artif Intell Med 58 (1): 63-72.
2. Peng CY, Harwell MR, Liou SM, Ehman LH (2006) Advances in missing data methods and implications for educational research.

3. Peugh JL, Enders CK (2004) Missing data in educational research: a review of reporting practices and suggestions for improvement. Rev Educ Res 74 (4): 525-556.
4. Young W, Weckman G, Holland W (2011) A survey of methodologies for the treatment of missing values within datasets: limitations and benefits. Theor Issues Ergon Sci 12 (1): 15-43.
5. Alosh M (2009) The impact of missing data in a generalized integer-valued autoregressión model for count data. J Biopharm Stat 19 (6): 1039-1054.
6. Knol MJ, Janssen KJM, Donders ART, Egberts ACG, Heerdink ER, Grobbee DE, Moons KGM, Geerlings MI (2010) Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. J Clin Epidemiol 63 (7): 728-736.
7. Little RJA, Rubin DB (2002) Missing data in experiments. In: Statistical analysis with missing data. Wiley, pp 24-40.
8. Jones MP (1996) Indicator and stratification methods for missing explanatory variables in multiple linear regressión. J Am Stat Assoc 91 (433): 222-230.
9. Little RJA (2016) Statistical analysis with missing data. Wiley, New York
10. Schafer JL (1999) Multiple imputation: a primer. Stat Methods Med Res 8 (1): 3-15.
11. de Waal T, Pannekoek J, Scholtus S (2011) Handbook of statistical data editing and imputation. Wiley, New York.
12. Roth PL (1994) Missing data: a conceptual review for applied psychologists. Pers Psychol 47 (3): 537-560.
13. Hug CW (2009) Detecting hazardous intensive care patient episodes using real-time mortality models. Thesis, Massachusetts Institute of Technology.
14. Wood AM, White IR, Thompson SG (2004) Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. Clin Trials 1 (4): 368-376.
15. Enders CK (2010) Applied missing data analysis, 1st edn. The Guilford Press, New York.
16. Rubin DB (1988) An overview of multiple imputation. In: Proceedings of the survey research section, American Statistical Association, pp 79-84.
17. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L-W, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG (2011) Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database. Crit Care Med 39 (5): 952-960.
18. Scott DJ, Lee J, Silva I, Park S, Moody GB, Celi LA, Mark RG (2013) Accessing the public MIMIC-II intensive care relational database for clinical research. BMC Med Inform Decis Mak 13 (1): 9.
19. Schafer JL, Olsen MK (1998) Multiple imputation for multivariate missing-data problems: a data analyst's perspective. Multivar Behav Res 33 (4): 545-571.

Capítulo 14

RUIDO VERSUS VALORES ATÍPICOS

Cátia M. Salgado, Carlos Azevedo,
Hugo Proença y Susana M. Vieira

Objetivos de Aprendizaje

Conocer:

- Qué métodos habituales hay disponibles para la detección de valores atípicos
- Cómo elegir los métodos más apropiados.
- Cómo evaluar el desempeño de un método de detección de valores atípicos y cómo comparar diferentes métodos.

14.1 Introducción

Un valor atípico o "*outlier"* (en inglés) es un dato que es diferente del resto [1]. Los *"outliers"* también son conocidos como anormalidades, discordantes, datos desviados y anomalías [2]. Mientras que el ruido puede ser definido como ejemplos mal catalogados (ruido de clase) o errores en los valores de los atributos (ruido de atributos), un *"outlier"* es un concepto más amplio que incluye no sólo errores sino también datos discordantes que pueden surgir de la variación natural dentro de la población o del proceso. Como tal, a menudo los *"outliers"* contienen información interesante y útil sobre el sistema subyacente. Estas particularidades han sido explotadas en control del fraude, sistemas de detección de intrusión, detección de robots web, pronóstico del tiempo, fuerzas del orden y diagnóstico médico [1], utilizando en general métodos de detección de *"outliers"* supervisados (ver más adelante).

Dentro del dominio médico en general, la principal fuente de *"outliers"* son las fallas en los equipos, errores humanos, anomalías que surgen de comportamientos específicos del paciente y la variación natural entre pacientes. Considere por ejemplo un resultado anormal de un análisis de sangre. Distintas razones pueden explicar la presencia de *"outliers":* estados patológicos graves, ingesta de drogas, comida o alcohol, actividad física reciente, estrés, ciclo menstrual, deficiente recolección y/o manipulación de las muestras de sangre. Mientras que algunas razones pueden señalar la

existencia de características específicas del paciente discordantes con el paciente "promedio", en cuyo caso la observación siendo un *"outlier"* provee información útil, otras razones pueden señalar errores humanos, y por lo tanto, debe considerarse la eliminación o corrección de la observación. Entonces, es crucial considerar las causas que pueden ser responsables de los *"outliers"* en una base de datos dada antes de proceder a cualquier tipo de acción.

Las consecuencias de no investigar los datos en busca de *"outliers"* pueden ser catastróficas. El efecto negativo de los *"outliers"* puede resumirse en: (1) aumento de la variación del error y reducción del poder estadístico; (2) disminución de la normalidad en los casos en que los valores atípicos no se distribuyen al azar; (3) sesgo del modelo al corromper la verdadera relación entre la exposición y el resultado [3].

Se necesita una buena comprensión de los datos en sí mismos antes de elegir un modelo para detectar *"outliers".* Varios factores influyen en la elección de un método de detección de *"outliers",* incluyendo el tipo de datos, su tamaño y distribución, la disponibilidad de los datos verdaderos de base, y la necesidad de interpretación en un modelo [2]. Por ejemplo, los modelos de regresión son más adecuados para hallar *"outliers"* en datos correlacionados linealmente, mientras que los métodos de agrupamiento ("clustering") se recomiendan cuando los datos no se distribuyen linealmente a lo largo de planos de correlación. Si bien este capítulo proporciona una descripción de algunos de los métodos más comunes para la detección de *"outliers",* existen muchos otros.

Evaluar la efectividad de un algoritmo de detección de *"outliers"* y comparar los diferentes enfoques es complejo. Aún más, a menudo los valores verdaderos de base que corresponden a los *"outliers"* no se encuentran disponibles, como en el caso de escenarios no supervisados, lo que dificulta el uso de métodos cuantitativos para evaluar la efectividad de los algoritmos de una manera rigurosa. El analista de datos solo tiene la alternativa de la evaluación cualitativa e intuitiva de los resultados [2]. Para superar esta dificultad, utilizaremos en este capítulo modelos de regresión logística para investigar el desempeño de diferentes técnicas de identificación de *"outliers"* en el estudio de un caso médicamente relevante.

14.2 Parte 1 - Conceptos Teóricos

Los métodos de identificación de *"outliers"* pueden clasificarse en métodos supervisados y no supervisados, dependiendo de si se dispone o no de información previa sobre las anormalidades en los datos. Las técnicas pueden dividirse a su vez en métodos univariables o multivariables, según el número de variables consideradas en la base de datos de interés.

La forma más simple de detección de *"outliers"* es el análisis de valores extremos de datos unidimensionales. En ese caso, el principio central del descubrimiento de los *"outliers"* es determinar las colas estadísticas de la distribución subyacente y asumir que los valores muy grandes o muy pequeños son *"outliers".* Con el fin de aplicar este tipo de técnica a un set de datos multidimensional, el análisis se realiza de a una dimensión a la vez. En dicho análisis multivariable, los *"outliers"* son muestras que tienen combinaciones inusuales con otras muestras en el espacio multidimensional. Es posible tener "outliers" con valores marginales razonables (ej., el valor parece normal cuando se limita a una dimensión), pero debido a las combinaciones lineales y no lineales de múltiples atributos estas observaciones revelan patrones inusuales en relación al resto de la población en estudio.

Para comprender mejor esto, la Fig. 14.1 proporciona un ejemplo gráfico de un escenario donde los *"outliers"* solo son visibles en un espacio bidimensional. Un análisis del diagrama de caja no revelará ningún "outlier" (ningún punto de datos por encima y por debajo de 1,5 RIC (el rango intercuantil, véase Capítulo 15 – Análisis Exploratorio de Datos), un método de detección de "outliers" ampliamente utilizado), mientras que una observación cercana de los conglomerados naturales presentes en los datos revelará patrones irregulares. Los "outliers" pueden identificarse por inspección visual, destacando los puntos de datos que parecen estar relativamente fuera de los grupos inherentes de datos 2D.

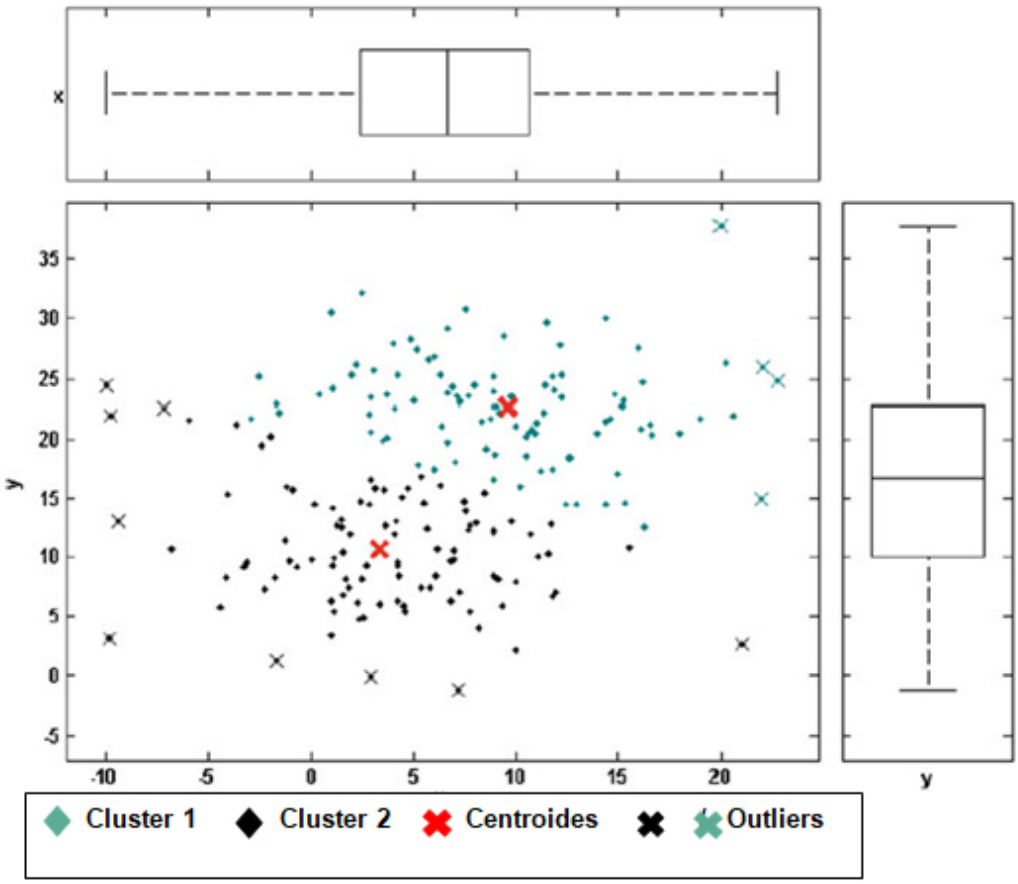

Figura 14.1 Investigación de "outliers" univariable (diagrama de caja) versus multivariable (diagrama de dispersión).

14.3 Métodos Estadísticos

En el campo de la estadística, se supone que los datos siguen un modelo de distribución (ej., distribución normal) y una instancia es considerada un "outlier" si se desvía significativamente del modelo [2, 4]. El uso de distribuciones normales simplifica el análisis, ya que la mayoría de las pruebas estadísticas existentes, como el score-Z, pueden interpretarse directamente en términos de probabilidades de significancia. Sin embargo, en muchos set de datos del mundo real la distribución subyacente de los datos es desconocida o compleja. Las pruebas estadísticas aún proveen una buena aproximación de los *"outliers"*, pero los resultados de las pruebas necesitan ser interpretados con cuidado y no pueden expresarse estadísticamente [2]. Los próximos apartados describen algunas de las

pruebas estadísticas más ampliamente utilizadas para la identificación de "outliers".

14.3.1 Método de Tukey

Los cuartiles son los valores que dividen un conjunto de números en cuartos. El (RIC) es la distancia entre el cuartil inferior (Q1) y el superior (Q3) en el diagrama de caja, que es RIC = Q3 – Q1. Puede utilizarse como una medida de cuan expandidos se encuentran los valores. Los "límites" interiores están situados a una distancia de 1.5 RIC por debajo de Q1 y por encima de Q3, y los límites exteriores a una distancia de 3 RIC por debajo de Q1 y por encima de Q3 [5]. Un valor entre los límites interiores y exteriores es un "outlier" **posible**, mientras que un valor que cae fuera de los límites exteriores es un "outlier" **probable**. La eliminación de todos los "outliers" posibles y probables se conoce como el método Intercuartil (IC), mientras que en el método de Tukey sólo son descartados los "outliers" probables.

14.3.2 Score Z

La prueba del valor Z calcula el número de desviaciones estándar con el que los datos varían de la media. Presenta un criterio razonable para la identificación de "outliers" cuando los datos se distribuyen normalmente. Se define como:

$$z_i = \frac{x_i - \bar{x}}{s} \qquad (14.1)$$

donde $\bar{x}$ y s indican la media de la muestra y la desviación estándar, respectivamente. En casos donde la media y la desviación estándar de la distribución pueden estimarse con precisión (o están disponibles a partir del conocimiento del dominio), una buena "regla de oro" es considerar los valores con $|Z_i| \geq 3$ como "outliers". Cabe destacar que este método es de valor limitado para set de datos pequeños, ya que la máxima puntuación de Z es como mucho $n - 1/\sqrt{n}$ [6].

14.3.3 Score Z Modificado

Los estimadores utilizados en el Score Z, la media y la desviación estándar de la muestra, pueden verse afectados por valores extremos presentes en los datos. Para evitar este problema, el score Z modificado utiliza la mediana $\tilde{x}$ y la desviación mediana absoluta (DMA) en lugar de la media y la desviación estándar de la muestra [7]:

$$M_i = \frac{0.6745(x_i - \tilde{x})}{\text{DMA}} \qquad (14.2)$$

donde

$$DMA = \text{mediana}\ \{|x_i - \tilde{x}|\} \qquad (14.3)$$

Los autores recomiendan utilizar el score Z modificado con $|M_i| \geq 3.5$ como potenciales "outliers". El supuesto de la normalidad de los datos aún se mantiene.

14.3.4 Rango Intercuartilo con Distribución Logarítmica-Normal

Las pruebas estadísticas discutidas anteriormente están basadas específicamente en el supuesto de que los datos se distribuyen con suficiente normalidad. En el ámbito de la atención sanitaria es común encontrar datos asimétricos, por ejemplo los tiempos de procedimientos quirúrgicos o la oximetría de pulso [8]. Véase el Capítulo 15 – Análisis Exploratorio de Datos para una definición formal de asimetría. Si una variable sigue una distribución log-normal entonces los logaritmos de las observaciones siguen una distribución normal. Un enfoque razonable es entonces aplicar el *ln* a los datos originales y aplicar las pruebas previstas a las distribuciones "normalizadas". Nos referimos a este método como log-IC.

14.3.5 Residuos y Residuos Estudentizados

En un modelo de regresión lineal, los residuos se definen como la diferencia entre los valores observados y esperados. Los puntos de datos con grandes residuos difieren de la tendencia general de regresión y pueden representar "outliers". El problema es que sus magnitudes dependen de sus unidades de medida haciendo difícil, por ejemplo, definir un umbral en el cual un punto es considerado un "outlier". Los residuos estudentizados

eliminan las unidades de medida dividiendo los residuos por una estimación de su desviación estándar. Una limitación de este enfoque es que supone que el modelo de regresión está correctamente especificado.

14.3.6 Distancia de Cook

En un modelo de regresión lineal, la distancia de Cook se utiliza para estimar la influencia de un punto de datos en la regresión. El principio de la distancia de Cook es medir el efecto de eliminar una observación dada. Los puntos de datos con una gran distancia pueden representar "outliers". Para el punto *i* en la muestra, la distancia de Cook se define como:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j \hat{y}_{j(i)})^2}{(k+1)s^2} \qquad (14.4)$$

Donde $\hat{y}_{j\ (i)}$ es la predicción de y_j por el modelo de regresión revisado cuando se elimina el punto *i* de la muestra, y *s* es la estimación de la raiz de la media del error al cuadrado. Instintivamente, D_i es una medida normalizada de la influencia del punto *i* en todos los valores medios predichos $\hat{y}_j$ con $j = 1, \ldots, n$. Diferentes valores de corte pueden utilizarse para marcar puntos de gran influencia. Cook ha sugerido que una distancia >1 representa una simple guía operativa. Otros han sugerido un umbral de $4/n$, donde *n* representa el número de observaciones.

14.3.7 Distancia de Mahalanobis

Esta prueba se basa en el método de Wilks diseñado para detectar un solo "outlier" en una muestra normal multivariable. Aproxima la distancia máxima de Mahalanobis al cuadrado (MD, del inglés Mahalanobis Distance) a una formulación de función de distribución-F, que suele ser más apropiada que una distribución X^2 [10]. Para una muestra multivariada p-dimensional X_i ($i=1, \ldots, n$), la distancia de Mahalanobis del caso *i* se define como:

$$MD_i = \sqrt{(x_i - t)^T C^{-1} (x_i - t)} \qquad (14.5)$$

donde *t* es la ubicación multivariable estimada, que suele ser la media aritmética, y *C* es la matriz de covarianza estimada, generalmente la matriz de covarianza de la muestra.

Los “outliers” multivariantes pueden definirse simplemente como observaciones que tienen una gran distancia cuadrada de Mahalanobis. En este trabajo, la distancia cuadrada de Mahalanobis se compara con los cuantiles de la *distribución– F* con *p* y *p-1* grados de libertad. Los valores críticos se calculan usando los límites de Bonferroni.

14.4 Modelos basados en la Proximidad

Las técnicas basadas en la proximidad son sencillas de aplicar y, a diferencia de los modelos estadísticos, no tienen supuestos previos sobre el modelo de distribución de datos. Son adecuados para la detección supervisada y no supervisada de “outliers” multivariantes [4].

El agrupamiento en conglomerados (clustering) es un tipo de técnica basada en la proximidad que comienza partiendo un set de datos N-dimensional en subgrupos *c* de muestras (clusters) basadas en su similitud. Luego, se utiliza alguna medida de ajuste de los puntos de datos a los diferentes clusters con el fin de determinar si los puntos de datos son “outliers” [2]. Una dificultad con este tipo de técnica es que asume formas específicas de los clusters dependiendo de la función de distancia utilizada en el algoritmo de agrupamiento. Por ejemplo, en un espacio tridimensional, la distancia euclidiana consideraría las esferas como equidistantes, mientras que la distancia de Mahalanobis consideraría los elipsoides como equidistantes (donde la longitud del elipsoide en un eje es proporcional a la variación de los datos en esa dirección).

14.4.1 K-medias

El algoritmo K-medias es ampliamente utilizado en minería de datos debido a su simplicidad y escalabilidad [11]. La dificultad asociada con este logaritmo es la necesidad de determinar *k*, el número de clusters, por adelantado. El algoritmo minimiza la suma de cuadrados dentro del cluster, la suma de la distancia entre cada punto de un cluster y el centroide del mismo. En k-medias, el centro de un grupo es la media de las medidas del grupo. Métricas como el Criterio de Información de Akaike o el Criterio de Información Bayesiano, que suman un factor proporcional a *k*, a la función costo utilizada durante la agrupación, pueden ayudar a determinar *k*. Un

valor de *k* demasiado grande aumentará la función costo incluso si reduce la suma de los cuadrados del cluster [12, 13].

14.4.2 K-Medoids

Similar a k-medias, el algoritmo de agrupamiento k-medoids divide el set de datos en grupos de manera que minimice la suma de las distancias entre un punto de datos y su centro. En contraste con el algoritmo k-medias, en k-medoids, el centro de los clusters son miembros del grupo. En consecuencia, si hay una región de "outliers" fuera del área con mayor densidad de puntos, el centro del cluster no será empujado hacia la región de "outliers", como en k-medias. Entonces, el k-medoids es más robusto frente a los "outliers" que k-medias.

14.4.3 Criterios de Detección de "outliers".

Después de determinar la posición del centro del cluster con k-means o bien k-medoids, debe especificarse el criterio para clasificar un ítem como un "outlier", y existen diferentes opciones:

Criterio 1: El primer criterio propuesto para detectar "outliers" se basa en la distancia euclidiana a los centros del cluster *C*, de tal manera que los puntos que están más lejos de su centro que la distancia mínima interclusters son considerados "outliers":

$$x \in C_k \quad \text{es outlier si} \quad d(x, C_k) > \min_{k \neq j}\{\delta(C_k, C_j)\} \times w \qquad (14.6)$$

donde $d(x, C_k)$ es la distancia euclidiana entre el punto *x* y el centro C_k, $\delta(C_k, C_j)$ es la distancia entre los centros C_k y C_j y $w = \{0.5, 0.7, 1, 1.21, 1.5\}$ es un parámetro de ponderación que determina cuán agresivamente el método eliminará los "outliers".

La Figura 14.2 provee un ejemplo gráfico del efecto de la variación de los valores de *w* en la creación de límites para la detección de "outliers". Mientras que valores pequeños de *w* remueve agresivamente los "outliers", a medida que la w aumenta es más difícil identificarlos.

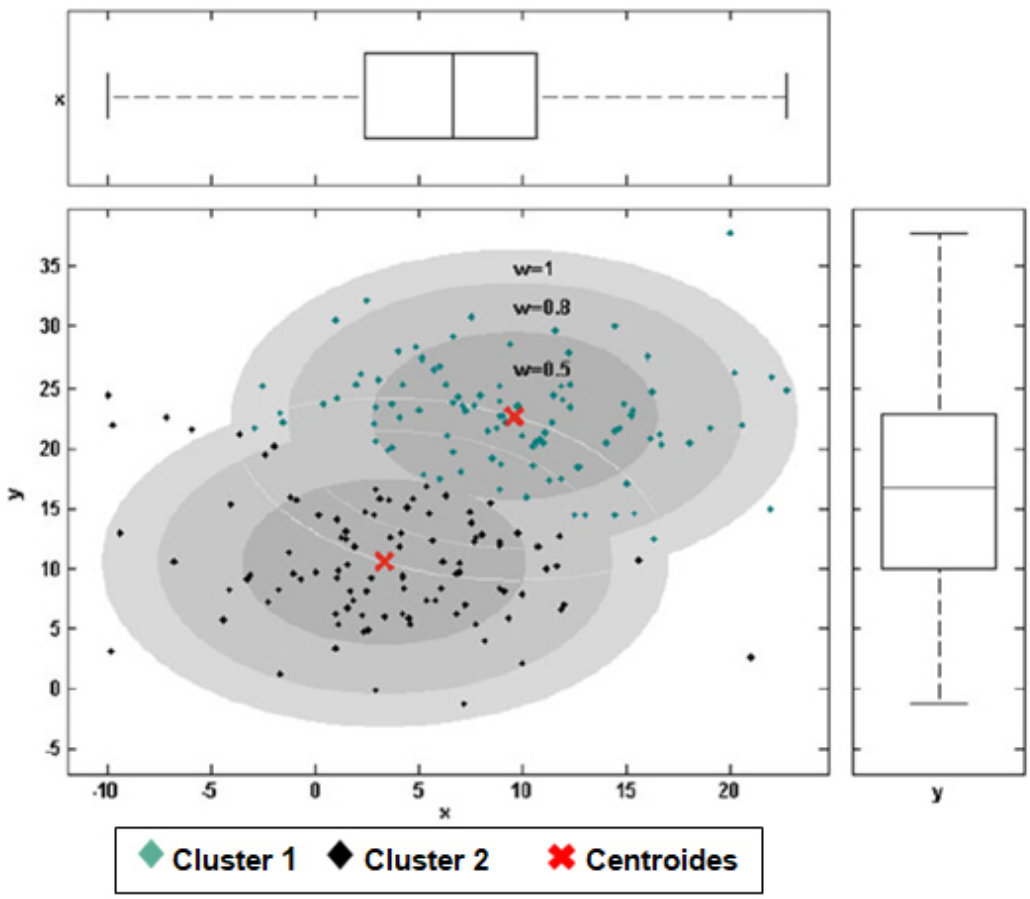

Figura 14.2 Efecto de diferentes pesos *w* en la detección de "outliers" basado en conglomerados, utilizando el criterio 1.

Criterio 2: En este criterio calculamos la distancia de cada punto de datos a su centroide (en el caso de k-means) o su medoid (en el caso de k-medoids) [14]. Si el ratio de la distancia del punto más cercano al centro del cluster y las distancias calculadas son menores que un umbral dado, entonces ese punto es considerado un "outlier". El umbral es definido por el usuario y debe depender del número de clusters seleccionados, ya que cuanto mayor es el número de clusters, más cercanos están los puntos dentro del cluster, es decir, el umbral disminuye con el aumento de *c.*

14.5 Detección Supervisada de valores "outliers"

En muchos escenarios, se puede disponer de conocimientos previos sobre los "outliers" que pueden utilizarse para etiquetar los datos debidamente y para identificar los "outliers" de interés. Los métodos basados en ejemplos

anteriores de “outliers” se denominan métodos supervisados de detección de “outliers” e involucran modelos de entrenamiento de clasificación que posteriormente pueden utilizarse para identificar “outliers” en los datos. Los métodos supervisados generalmente son ideados para detectar anomalías en los dominios de aplicación donde las anomalías son consideradas eventos de interés. Algunos ejemplos incluyen el control del fraude, sistemas de detección de intrusos, detección de robots web o diagnóstico médico [1]. Por lo tanto, las etiquetas representan lo que el analista de datos podría estar buscando específicamente en lugar de uno que podría querer eliminar [2]. La principal diferencia comparado con muchos otros problemas de clasificación es la naturaleza desbalanceada inherente de los datos, ya que los casos etiquetados como “anormales” se presentan con mucha menor frecuencia que los casos etiquetados como “normales”. Los lectores interesados pueden encontrar más información sobre este tema por ejemplo en el libro de Aggarwal [2].

14.6 Análisis de “outliers” utilizando Conocimientos de Expertos

En los análisis univariables, el conocimiento de expertos puede utilizarse para definir umbrales de valores que son normales, críticos (con peligro de muerte) o imposibles porque caen fuera de los rangos permitidos o no tienen sentido físico [15]. Mediciones negativas de frecuencia cardíaca o temperatura corporal son ejemplos de valores imposibles. Es muy importante controlar la presencia de este tipo de “outliers” en el set de datos, ya que se originaron indudablemente por error humano o mal funcionamiento de los equipos, y deberían ser borrados o corregidos.

14.7 Caso de Estudio: Identificación de “outliers” en el Estudio de Uso de Catéter Arterial Invasivo (CAI)

En esta sección, aplicaremos distintos métodos para identificar “outliers” en dos set de datos clínicos del “mundo real” utilizados en un estudio que investigó el efecto de colocar un catéter arterial invasivo (CAI) en pacientes con falla respiratoria. Se utilizan dos set de datos, que incluyen pacientes en los que se usó un catéter arterial invasivo (grupo CAI) y pacientes en que no se usó (grupo no-CAI). El código utilizado para generar el análisis y las figuras están disponible en el repositorio GitHub para este libro.

Tabla 14.1 Rangos normales, críticos e imposibles para las variables seleccionadas, y valores máximos y mínimos presentes en la base de datos.

	Valor de referencia			Datos analizados		
Variable	Rango normal	Crítico	Imposible	Con CAI	Sin CAI	Unidades
Edad	-	->	<17 (adultos)	12,2-99,1	15,2-97,5	Años
SOFA	-	-≥≤<	<0 y >24	1-17	0-14	Sin unidades
(Recuentode glóbulos blancos)	3,9 - 10,7	≥ 100	<0	0,3-86,0	0,2-109,8	$x10^9$ cél/L
Hemoglobina	Varón: 13,5 - 17,5 Mujer: 12-16	≤6 y ≥20	<0	Varón: 3,2-19,0 Mujer: 2,0-18,1	Varón: 4,9-18,1 Mujer: 4,2-18,1	g/dL
Plaquetas	150-400	≤40 y ≥1000	<0	7,0-680,0	9,0-988,0	$x10^9$/L
Sodio	136-145	≤120 y ≥160	<0	105,0-165,0	111,0-154,0	mmol/L
Potasio	3,5-5	≤2,5 y ≥6	<0	1,9-9,8	1,9-8,3	mmol/L
CO2	22-28	≤10 y ≥40	<0	2,0-62,0	5,0-52,0	mmol/L
Cloruro	95-105	≤70 y ≥120	<0 y ≥160	81,0-133,0	78,0-127,0	mmol/L
Nitrógeno ureico en sangre	7-18	≥100	<0	2,0-139,0	2,0-126,0	mg/dL
Creatinina	0,6- 1,2	≥10	<0	0,2-12,5	0,0-18,3	mg/dL
PO2	75 - 105	≤40	<0	25,0-594,0	22,0-634,0	mmHg
PCO2	33-45	≤20 y ≥70	<0	8,0-141,0	14,0-158,0	mmHg

14.8 Análisis de Expertos

La Tabla 14.1 proporciona los valores máximos y mínimos para definir los rangos normales, críticos y permisibles de algunas de las variables analizadas en el estudio, así como los valores máximos y mínimos presentes en la base de datos.

14.9 Análisis Univariado

En esta sección, identificamos “outliers” univariados para cada variable dentro de un grupo predefinido (sobrevivientes y no sobrevivientes), utilizando los métodos estadísticos descritos más arriba.

La tabla 14.2 resume el número y porcentaje de “outliers” identificados con cada método en los grupos con catéter arterial permanente y sin el dispositivo. En general, los métodos de Tukey y log-IC son los más conservadores, es decir que identifican el menor número de puntos como “outliers”, mientras que IC identifica más “outliers” que cualquier otro método. Con unas pocas excepciones, el score Z modificado identifica más “outliers” que el score Z.

Tabla 14.2 Número y porcentaje de "outliers" identificados por cada método.

	CAI Clase 0 (811 pacientes)					Clase 1 (163 pacientes)				
	IC	**Tukey's**	**Log IC**	**Z Score**	**Mod Z score**	**IC**	**Tukey´s**	**Log IC**	**Z Score**	**Mod Z score**
Edad	0 (0,0 %)	0 (0,0 %)	1 (0,1 %)	0 (0,0 %)	0 (0,0 %)	5 (0,6 %)	0 (0,0 %)	8 (1,0 %)	4 (0,5 %)	5 (0,6 %)
SOFA	13 (1,6 %)	0 (0,0 %)	6 (0,7%)	2 (0,2 %)	20 (2,5 %)	16 (2,0 %)	3 (0,4 %)	8 (1,0 %)	1 (0,1 %)	5 (0,6 %)
Recuento GB	20 (2,5 %)	3 (0,4 %)	21 (2,6 %)	5 (0,6 %)	10 (1,2 %)	6 (0,7 %)	1 (0,1 %)	5 (0,6 %)	1 (0,1 %)	3 (0,4 %)
Hemoglobina	8 (1,0 %)	1 (0,1 %)	13 (1,6 %)	5 (0,6 %)	4 (0,5 %)	0 (0,0 %)	0 (0,0 %)	0 (0,0 %)	0 (0,0 %)	0 (0,0 %)
Plaquetas	17 (2,1 %)	1 (0,1 %)	36 (4,4 %)	7 (0,9 %)	7 (0,9 %)	4 (0,5 %)	0 (0,0 %)	2 (0,2 %)	2 (0,2 %)	1 (0,1 %)
Sodio	30 (3,7 %)	8 (1,0 %)	30 (3,7 %)	10 (1,2 %)	26 (3,2 %)	8 (1,0 %)	1 (0,1 %)	8 (1,0 %)	2 (0,2 %)	2 (0,2 %)
Potasio	39 (4,8 %)	30 (1,2 %)	35 (4,3 %)	14 (1,7 %)	26 (3,2 %)	9 (1,1 %)	1 (0,1 %)	7)0,9 %)	2 (0,2 %)	8 (1,0 %)
TCO2	24 (3,0 %)	4 (0,5%)	31 (3,8 %)	13 (1,6 %)	13 (1,6 %)	9 (1,1 %)	2 (0,2 %)	6 (0,7 %)	2 (0,2 %)	2 (0,2 %)
Cloro	21 (2,6%)	3 (0,4 %)	24 (3,0 %)	13 (1,6 %)	18 (2,2 %)	4 (0,5 %)	0 (0,0 %)	3 (0,4 %)	1 (0,1 %)	1 (0,1 %)
Urea	**72 (8,9 %)**	**37 (4,6 %)**	**45 (5,9 %)**	**20 (2,5 %)**	**60 (7,4 %)**	13 (1,6 %)	**9 (1,1 %)**	7 (0,9 %)	**5 (0,6 %)**	**13 (1,6 %)**
Creatinina	50 (6,2 %)	31 (3,8 %)	43 (5,9 %)	18 (2,2 %)	40 (4,9 %)	11 (1,4 %)	2 (0,2 %)	2 (0,2 %)	2 (0,2 %)	8 (1,0 %)
PO2	0 (0,0 %)	0 (0,0 %)	2 (0,2 %)	0 (0,0 %)	0 (0,0 %)	0 (0,0 %)	0 (0,0 %)	0 (0,0 %)	0 (0,0 %)	0 (0,0 %)
PCO2	53 (6,5 %)	22 (2,7 %)	**48 (5,9 %)**	19 (2,3 %)	37 (4,6 %)	11 (1,4 %)	4 (0,5 %)	**13 (1,6 %)**	4 (0,5 %)	9 (1,1 %)
Total Pacientes	220 (27,1 %)	86 (10,6 %)	200 (25,9 %)	91 (11,2 %)	165 (20,3%)	**63 (7,8 %)**	20 (2,5 %)	47 (5,8 %)	23 (2,8 %)	43 (5,3 %)
	Sin CAI **Clase 0 (524 pacientes)**					**Clase 1 (83 pacientes)**				
	IC	**Tukey´s**	**Log IC**	**Z Score**	**Mod Z score**	**IC**	**Tukey´s**	**Log IC**	**Z Score**	**Mod Z score**
Edad	0 (0,0 %)	0 (0,0 %)	0 (0,0 %)	0 (0,0 %)	0 (0,0 %)	1 (0,2 %)	0 (0,0 %)	3 (0,6 %)	1 (0,2 %)	1 (0,2 %)
SOFA	**51 (9,7 %)**	2 (0,4 %)	**48 (9,2 %)**	2 (0,4 %)	7 (1,3 %)	**9 (1,7 %)**	1 (0,2 %)	8 (1,5 %)	1 (0,2 %)	3 (0,6 %)
Recuento GB	21 (4,0 %)	4 (0,8 %)	10 (1,9 %)	4 (0,11 %)	11 (2,1 %)	11 (2,1 %)	1 (0,2 %)	4 (0,8 %)	1 (0,2 %)	3 (0,6 %)
Hemoglobina	1 (0,4 %)	0 (0,0 %)	6 (1,1 %)	2 (0,4 %)	2 (0,4 %)	0 (0,0 %)	0 (0,0 %)	2 (0,4 %)	0 (0,0 %)	0 (0,0 %)

Tabla 14.2 Continuación

	Sin CAI									
	Clase 0 (524 pacientes)					Clase 1 (83 pacientes)				
	IC	Tukey´s	Log IC	Z Score	Mod Z score	IC	Tukey´s	Log IC	Z Score	Mod Z score
Plaquetas	15 (2,9 %)	5 (1,0 %)	21 (4,0 %)	5 (1,0 %)	6 (1,1 %)	4 (0,8 %)	1 (0,2 %)	5 (1,0 %)	**2 (0,4 %)**	2 (0,4 %)
Sodio	25 (4,8 %)	9 (1,7 %)	25 (4,11 %)	9 (1,7 %)	9 (1,7 %)	5 (1,0 %)	1 (0,2 %)	5 (1,0 %)	1 (0,2 %)	1 (0,2 %)
Potasio	22 (4,2 %)	2 (0,4 %)	14 (2,7 %)	6 (1,1 %)	6 (1,1 %)	1 (0,2 %)	0 (0,0 %)	0 (0,0 %)	0 (0,0 %)	0 (0,0 %)
TCO2	27 (5,2 %)	4 (0,8 %)	31 (5,9 %)	8 (1,5 %)	8 (1,5 %)	4 (0,8 %)	1 (0,2 %)	4 (0,8 %)	**2 (0,4 %)**	3 (0,6 %)
Cloro	21 (4,0 %)	4 (0,8 %)	20 (3,11 %)	9 (1,7 %)	9 (1,7 %)	9 (1,7 %)	1 (0,2 %)	**9 (1,7 %)**	1 (0,2 %)	4 (0,8 %)
Urea	35 (6,7 %)	**20 (3,8 %)**	27 (5,2 %)	**13 (2,5 %)**	**34 (6,5 %)**	6 (1,1 %)	2 (0,4 %)	2 (0,4 %)	**2 (0,4 %)**	6 (1,1 %)
Creatinina	29 (5,5 %)	17 (3,2 %)	25 (4,8 %)	8 (1,5 %)	8 (1,5 %)	7 (1,3 %)	2 (0,4 %)	3 (0,6 %)	**2 (0,4 %)**	5 (1,0 %)
PO2	0 (0,0 %)	0 (0,0 %)	0 (0,0 %)	0 (0,0 %)	0 (0,0 %)	1 (0,2 %)	0 (0,0 %)	0 (0,0 %)	0 (0,0 %)	3 (0,6 %)
PCO2	34 (6,5 %)	11 (2,1 %)	33 (6,3 %)	10 (1,9 %)	10 (1,9 %)	8 (1,5 %)	**4 (0,8 %)**	6 (1,1 %)	2 (0,4 %)	8 (1,5 %)
Total Pacientes	**176 (33,6 %)**	59 (11,3 %)	172 (32,8 %)	56 (10,7 %)	111 (21,2 %)	**37 (7,1 %)**	11 (2,1 %)	29 (5,5 %)	11 (2,1 %)	28 (5,3 %)

Total Pacientes representa el número de pacientes identificados al considerar todas las variables juntas. Los resultados en **negrita** resaltan las variables con la mayoría de los outliers en cada método y también el método que remueve el mayor número de pacientes en total en cada clase. Clase 0 representa los sobrevivientes y Clase 1 no sobrevivientes

Una investigación preliminar de los resultados muestra que los valores que caen dentro de los rangos normales de referencia nunca son identificados como “outliers” (ver Tabla 14.1), cualquiera sea el método. Por otro lado, a menudo los valores críticos son identificados como tal. Pueden hacerse observaciones adicionales como (1) más “outliers” son identificados en la variable Nitrogeno ureico sanguineo (BUN, por sus siglas en inglés) que en ninguna otra y (2) la relación entre el número de “outliers” y el número total de pacientes es menor en las cohortes clase 1 (no sobrevivientes). Como se esperaba, para las variables que se acercan más a una distribución log-normal que a una distribución normal, como el potasio, BUN y PCO2, el método IC aplicado a la transformación logarítmica de los datos (método log-IC) identifica menos “outlier” que el IC aplicado a los datos reales. Considere por ejemplo la variable BUN, que sigue aproximadamente una distribución log-normal. La Figura 14.3 muestra un gráfico de dispersión de todos los puntos de datos y los “outliers” identificados en el grupo IAC.

Por otro lado, cuando los valores siguen aproximadamente una distribución normal, como en el caso del cloro (ver Fig. 14.4), el método IC identifica menos “outliers” que log-IC. Cabe destacar que el rango de valores considerados “outliers” difiere entre las clases, es decir, lo que se considera un “outlier” en la clase 0 no es necesariamente un “outlier” en la clase 1. Un ejemplo de esto son los valores menores a 90 mmol/L en el score Z modificado.

Ya que este es un análisis univariado, la investigación de valores extremos mediante el conocimiento de expertos es importante. Para el cloro, los valores normales están en el rango de 95-105 mmol/L, mientras que valores <70 ó >120 mmol/L se consideran críticos, y concentraciones por encima de 160 mmol/L son fisiológicamente imposibles [15].

La Figura 14.4 confirma que los valores normales siempre se mantienen, cualquiera que sea el método usado. Es importante, que algunos valores críticos no son identificados tanto en el score Z como en el score Z modificado (especialmente en la clase 1). Por lo tanto, parece que los métodos identifican “outliers” que no deberían ser eliminados, ya que es probable que representen valores reales en pacientes gravemente enfermos.

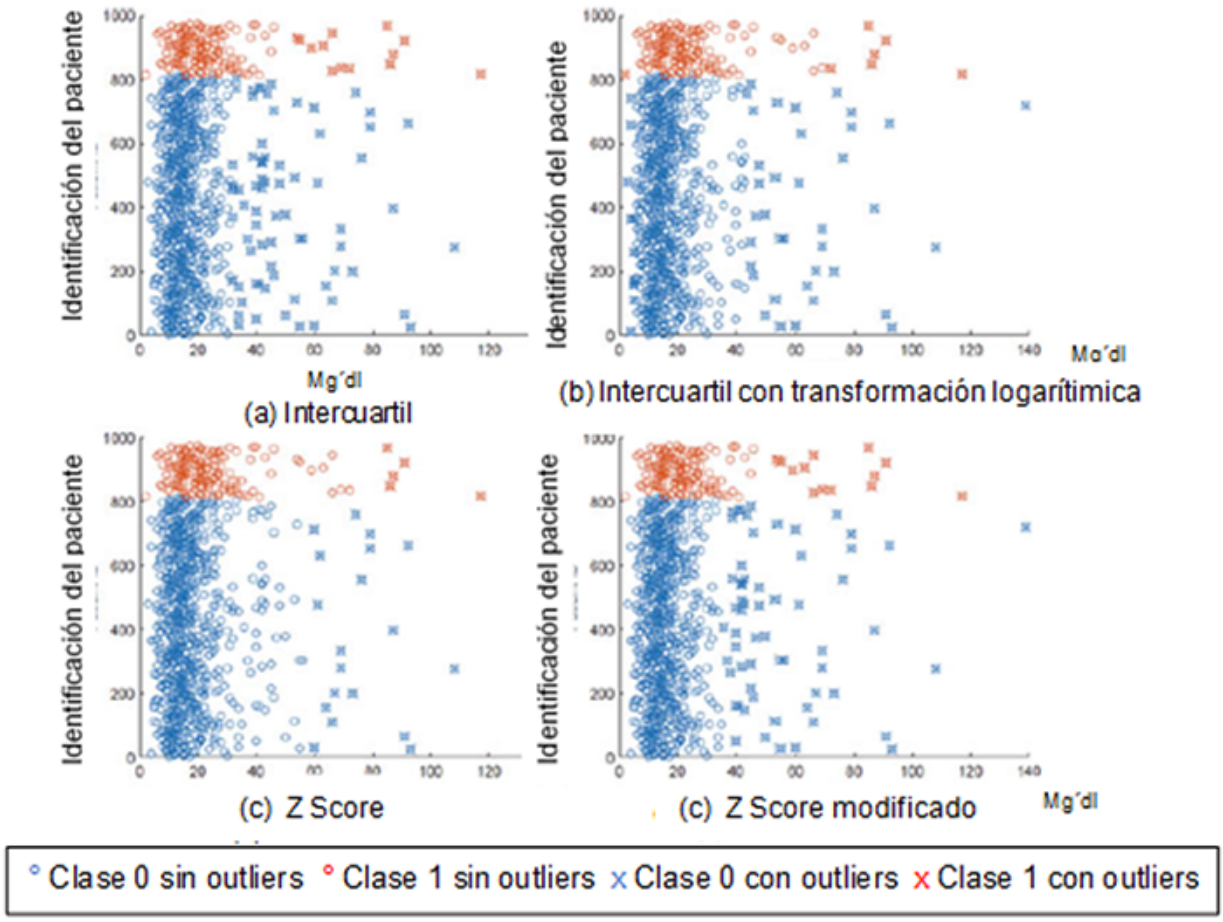

Fig 14.3 "Outliers" identificados por análisis estadístico para la variable BUN, en la cohorte CAI. Clase 0: sobrevivientes, Clase 1: no sobrevivientes.

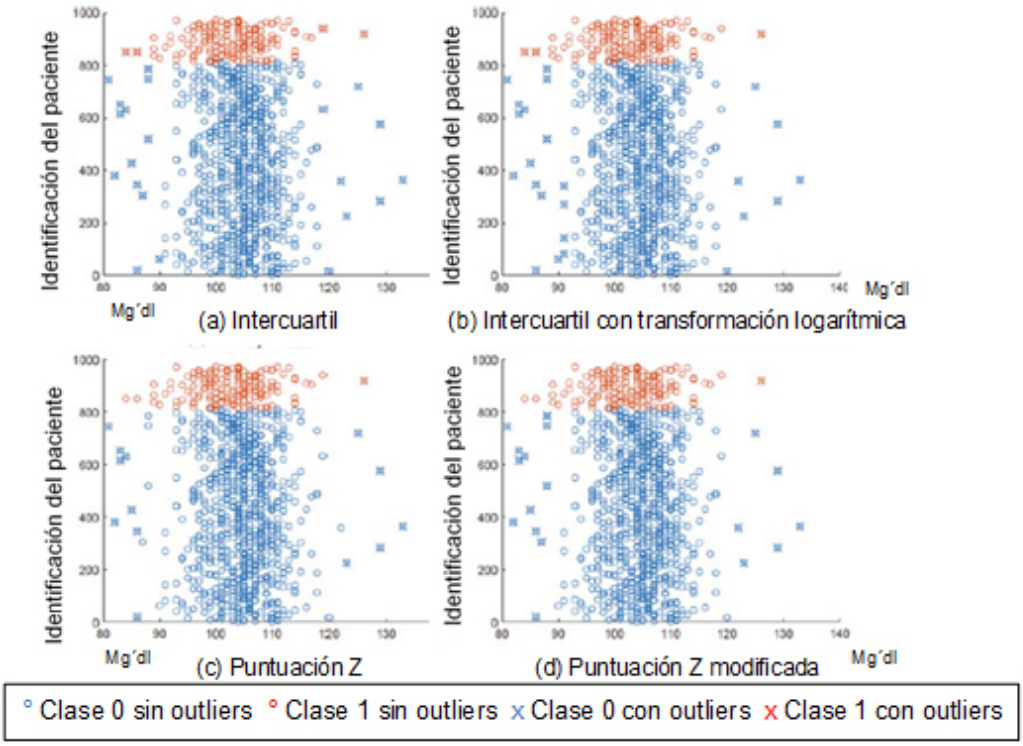

Fig 14.4 "Outliers" identificados por análisis estadístico para la variable cloro, en la cohorte CAI. Clase 0: sobrevivientes, Clase 1: no sobrevivientes.

14.10 Análisis Multivariado

Utilizando enfoques basados en modelos, se puede identificar una combinación inusual de valores para una serie de variables. En este análisis nos preocuparemos por los "outliers" multivariados para el conjunto completo de variables en el set de datos, incluidas las binarias. Con el fin de investigar "outliers" multivariados en los pacientes con y sin CAI, se prueban enfoques basados en la distancia de Mahalanobis y "clusters" (conglomerados), dentro de clases predefinidas.

Tabla 14.3 "Outliers"multivariados identificados por k-medias, k-medoids y distancia de Mahalanobis.

	Criterio	Peso	*c*		% de outliers Clase 0	
			Clase 0	Clase 1	Clase 0	Clase 1
CAI						
K-media, índice de la silueta	1	1.2	4 ± 3.1	2 ± 0.0	25.2±7.4	20.9 ± 11.0
	1	1.5	3 ± 2.9	2 ± 0.0	7.9 ± 4.6	3.3 ± 5.9
	1	1.7	3 ± 2.6	2 ± 0.0	3.6 ± 2.5	0.4 ± 2.2
	1	2	4 ± 3.1	2 ± 0.0	1.0 ± 1.1	0.1 ± 0.3
K-media, c=2	2	0.05	2 ± 0.0	2 ± 0.0	28.5 ± 4.8	2 ± 0.0
	2	0.06	2 ± 0.0	2 ± 0.0	9.3 ± 4.2	2.9 ± 5.2
K-medoids, índice de la silueta	1	1.2	4 ± 3.0	2 ± 0.0	4.1 ± 2.2	0.8 ± 3.1
	1	1.5	3 ± 2.6	2 ± 0.0	1.1 ± 1.0	0.1 ± 0.3
	1	1.7	3 ± 2.9	2 ± 0.0	0.2 ± 0.2	0.0 ± 0.0
	1	2	4 ± 3.0	2 ± 0.0	0.7 ± 0.4	0.0 ± 0.0
K-medoids, c=2	2	0.01	2 ± 0.0	2 ± 0.0	34.6 ± 8.6	2.5 ± 0.0
	2	0.02	2 ± 0.0	2 ± 0.0	20.8 ± 6.1	0.0 ± 0.0
Mahalanobis	-	-	-	-	16.7 ± 5.5	0.0 ± 0.0
Sin CAI						
K-media, índice de la silueta	1	1.2	9 ± 1.8	7 ±2.4	12.82±4.1	20.9 ± 11.0
	1	1.5	9 ± 1.7	7 ±2.5	2.8 ± 1.8	3.3 ± 5.9
	1	1.7	9 ± 1.8	7 ±2.5	0.9 ± 1.2	0.4 ± 2.2
	1	2	9 ± 2.4	7 ±2.5	0.2 ± 0.7	0.1 ± 0.3
K-media, c=2	2	0.05	2 ± 0.0	2 ± 0.0	25.5 ± 4.5	41 ± 11.9
	2	0.06	2 ± 0.0	2 ± 0.0	10.6 ± 2.6	4.8 ± 7.2
K-medoids, índice de la silueta	1	1.2	9 ± 1.5	7 ±2.5	3.8 ± 1.6	1.4 ± 1.6
	1	1.5	9 ± 2.0	7 ±2.4	0.9 ± 1.9	0.0 ± 0.0
	1	1.7	9 ± 2.0	7 ±2.4	0.3 ± 0.6	0.0 ± 0.0
	1	2	9 ± 1.3	7 ±2.5	0.4 ± 0.9	0.0 ± 0.0
K-medoids, c=2	2	0.01	2 ± 0.0	2 ± 0.0	19.7 ± 4.0	2.7 ± 8.8
	2	0.02	2 ± 0.0	2 ± 0.0	11.0 ± 2.8	1.0 ± 5.0
Mahalanobis	-	-	-	-	6.8 ± 2.6	0.8 ±40.0

Los resultados se presentan como media y desviación estandar

La Tabla 14.3 muestra el promedio de los resultados en términos de número de conglomerados c determinados por el coeficiente de silueta, y el porcentaje de pacientes identificados como “outliers”. Con el fin de tener en cuenta la variabilidad, las pruebas se realizaron 100 veces. Los datos se normalizaron sólo para probar los enfoques basados en conglomerados.

Teniendo en cuenta el escenario donde se crean dos conglomerados para el set de datos completo de CAI separados por clases, investigamos los “outliers” mirando las observaciones multivariadas alrededor de los centros de los conglomerados. La Figura 14.5 muestra un ejemplo de la detección de “outlier” utilizando k-medias y k-medoids con el criterio 1 y un peso igual a 1.5. Con fines ilustrativos, presentamos únicamente los resultados gráficos de los pacientes que murieron en el grupo CAI (clase 1). El eje x representa cada una de las características seleccionadas (ver Tabla 14.1) y el eje y representa los valores correspondientes normalizados entre 0 y 1. K-medoids no identifica ningún “outlier”, mientras que K-medias identifica 1 “outlier” en el primer conglomerado y 2 “outliers” en el segundo conglomerado. Esta diferencia puede atribuirse al hecho de que la distancia interconglomerados es menor en k-medoids que en k-medias.

La detección de “outliers” parece ser más influenciada por las características binarias que por las características continuas; las líneas rojas están, con algunas excepciones, bastante cerca de las líneas negras para las variables continuas (1 a 2 y 15 a 25) y distantes en las variables binarias. Una explicación posible es que el agrupamiento fue diseñado esencialmente para datos continuos multivariables; las variables binarias producen una máxima separación, ya que solo existen dos valores, 0 y 1, con nada entre ellos.

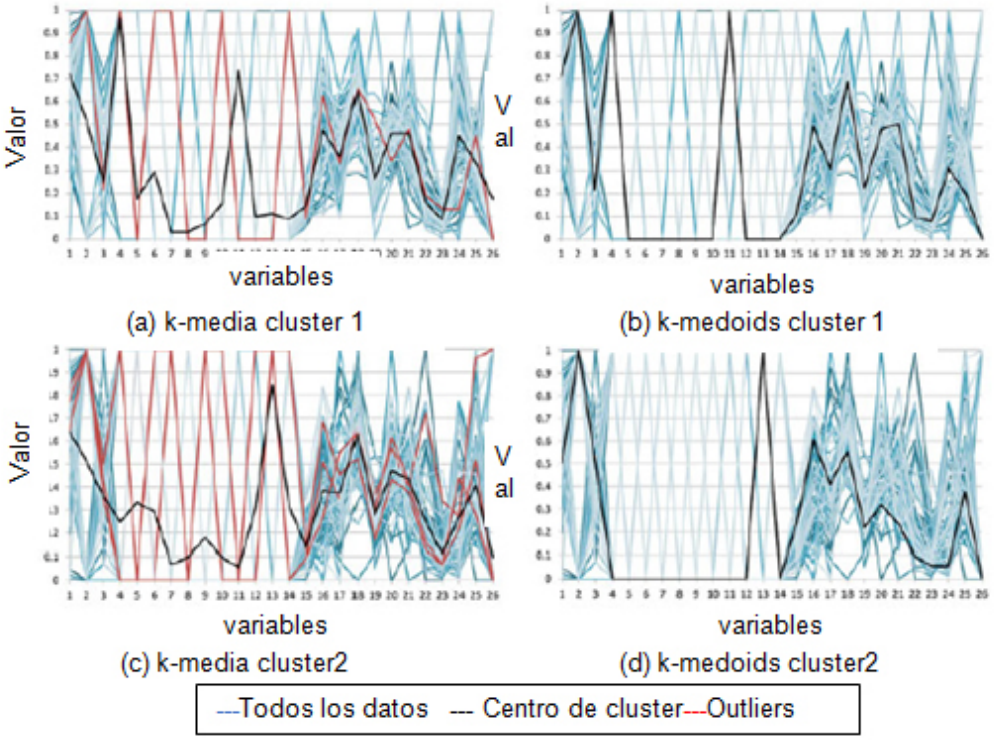

Fig. 14.5 "Outliers" identificados por enfoques basados en conglomerados ("clusters") de pacientes que murieron después del CAI. Se utilizó el criterio 1, basado en distancia interconglomerados, con *c=2* y *w=1.5*. K-medoids no identifica "outliers", mientras que k-medias identifica 1 "outlier" en el conglomerado 1 y 2 "outliers" en el conglomerado 2.

14.11 Clasificación de la Mortalidad en Pacientes con Catéter Arterial Invasivo y sin Catéter Arterial Invasivo

Se crearon modelos de regresión logística para evaluar el efecto de la eliminación de "outliers" utilizando diferentes métodos en la clasificación de la mortalidad en pacientes con y sin CAI, siguiendo el mismo razonamiento que en el Capítulo 13 – Datos faltantes. Se utilizó un enfoque de validación cruzada de 10 iteraciones para evaluar la validez y la robustez de los modelos. En cada iteración, se aplicaron todos los métodos de identificación de "outliers" por separado, para cada clase del set de entrenamiento y los resultados se promediaron a lo largo de las iteraciones. Antes de la validación cruzada, los valores fueron normalizados entre 0 y 1 utilizando el procedimiento de mínimo y máximo. Para el método log-IC, los datos fueron transformados en forma logarítmica antes de la normalización, excepto las variables que contenían valores nulos (variables binarias en la Tabla 14.1, SOFA y creatinina). También investigamos el escenario en el que sólo se

consideró el 10% de los peores ejemplos detectados por cada método estadístico dentro de cada clase, y el caso en el que no se eliminaron los valores atípicos (se utilizaron todos los datos). En los enfoques basados en "clustering", el número de conglomerados *c* fue elegido entre 2 y 10 utilizando el método del coeficiente de silueta. También mostramos el caso donde *c* se fijó en 2. El peso de los enfoques basados en "clustering" se ajustó según las particularidades del método. Dado que el centro del conglomerado en k-medoids es un punto de datos que pertenece al set de datos, la distancia a su vecino más cercano es menor que en el caso de k-medias, especialmente porque se consideran muchas variables binarias. Por esta razón, elegimos valores más altos de *w* para el criterio 2 de k-medias.

El desempeño de los modelos se evaluó en términos de Área Bajo la Curva de la Característica Operativa del Receptor (AUC, por su nombre en inglés, area under the curve), precisión o accuracy (TCC, tasa de clasificación correcta), sensibilidad (tasa de verdaderos positivos), y especificidad (tasa de verdaderos negativos). Una prueba específica sugerida por DeLong y DeLong puede entonces probar si los resultados difieren significativamente [16].

Los resultados del desempeño para el grupo CAI se muestran en la Tabla 14.4, y el porcentaje de pacientes eliminados utilizando cada método en la Tabla 14.5. Para ser más concisos, no se muestran los resultados del grupo no CAI. El mejor rendimiento para CAI es un AUC=0.83 y TCC=0.78 (destacado en negrita). La máxima sensibilidad es de un 87% y la máxima especificidad de un 79%, sin embargo estos dos no se producen simultáneamente. En general, la mejor AUC se obtiene cuando se utilizan todos los datos y cuando se eliminan unos pocos "outliers". El peor desempeño se obtiene utilizando el score z sin recortar los resultados y k-medias y k-medoids utilizando *c=2*, criterio 1 y peso de 1.2. En cuanto al grupo no CAI, el mejor desempeño corresponde a un AUC= 0.88, TCC= 0.84, sensibilidad=0.85 y especificidad=0.85. Nuevamente, el mejor desempeño se alcanza cuando se utilizan todos los datos y en los casos donde se eliminaron menos "outliers". El peor desempeño por lejos se obtiene cuando se eliminan todos los "outliers" identificados por el score Z. De forma similar al grupo CAI, para el criterio 1 de k-medias y k-medoids, valores crecientes de peso proporcionan los mejores resultados.

Tabla 14.4 Resultados de la regresión logística del set de datos CAI utilizando una validación cruzada de 10 iteraciones, luego de la eliminación de los "outliers" y utilizando el set de datos original.

Estadística	Punto de corte	AUC	ACC	Sensibilidad	Especificidad
IQ	-	0.81 ± 0.05	0.76 ± 0.05	0.71 ± 0.14	0.76 ± 0.06
	10	0.82 ± 0.06	0.77± 0.06	0.76 ± 0.11	0.77 ± 0.07
Tukey's	-	0.82 ± 0.05	0.75± 0.06	0.76 ± 0.09	0.75 ± 0.06
	10	**0.83 ± 0.06**	0.78 ± 0.05	0.75 ± 0.10	0.78 ± 0.06
Log-IQ	-	0.81 ± 0.07	0.76 ± 0.07	0.74 ± 0.14	0.76 ± 0.06
	10	**0.83 ± 0.06**	**0.78 ± 0.04**	0.73 ± 0.10	0.79 ± 0.05
Score Z	-	0.78 ± 0.03	0.67 ± 0.06	0.85 ± 0.09	0.64 ± 0.08
	10	0.81 ± 0.07	0.75 ± 0.06	0.74 ± 0.13	0.75 ± 0.07
Score Z modificado	-	0.82± 0.05	0.76 ± 0.05	0.77± 0.14	0.76± 0.05
	10	0.82 ± 0.06	0.77 ± 0.06	0.75 ± 0.10	0.77 ± 0.05
Mahalanobis	-	0.81 ± 0.08	0.75 ± 0.06	0.73 ± 0.10	0.76 ± 0.07
Basado en un cluster	Peso	AUC	ACC	Sensibilidad	Especificidad
K-media, índice de la silueta criterio 1	1.2	0.81 ± 0.08	0.72 ± 0.05	0.80 ± 0.12	0.70 ± 0.06
	1.5	0.82 ± 0.05	0.76 ± 0.06	0.76 ± 0.11	0.76 ± 0.06
	1.7	**0.83 ± 0.06**	**0.78 ± 0.05**	0.77 ± 0.10	0.78 ± 0.06
	2	**0.83 ± 0.06**	**0.78 ± 0.05**	0.74 ± 0.09	0.78 ± 0.06
K-media, c=2 criterio 1	1.2	0.79 ± 0.08	0.66 ± 0.05	0.84 ± 0.10	0.63 ± 0.06
	1.5	0.82 ± 0.06	0.73 ± 0.06	0.79 ± 0.09	0.72 ± 0.07
	1.7	0.82 ± 0.06	0.75 ± 0.06	0.78 ± 0.08	0.75 ± 0.08
	2	**0.83 ± 0.07**	**0.78 ± 0.06**	0.76 ± 0.09	0.78 ± 0.06
K- media criterio 2	0.05	**0.83 ± 0.07**	0.77 ± 0.05	0.74 ± 0.09	0.78 ± 0.06
	0.06	**0.83 ± 0.06**	0.77 ± 0.06	0.75 ± 0.10	0.78 ± 0.06
K-medoids, índice de la silueta criterio 1	1.2	0.81 ± 0.04	0.68 ± 0.04	0.85 ± 0.09	0.64 ± 0.05
	1.5	**0.83 ± 0.05**	0.74 ± 0.04	0.80 ± 0.10	0.73 ± 0.06
	1.7	**0.83 ± 0.05**	0.75 ± 0.06	0.78 ± 0.10	0.74 ± 0.07
	2	**0.83 ± 0.06**	0.77 ± 0.06	0.77 ± 0.09	0.77 ± 0.06
K-medoids, C=2, criterio 1	1.2	0.78 ± 0.06	0.62 ± 0.07	0.87 ± 0.08	0.57 ± 0.07
	1.5	0.81 ± 0.06	0.70 ± 0.06	0.83 ± 0.10	0.68 ± 0.08
	1.7	0.82 ± 0.06	0.72 ± 0.06	0.80 ± 0.10	0.71 ± 0.08
	2	**0.83 ± 0.07**	0.76 ± 0.06	0.77 ± 0.10	0.75 ± 0.07
K- medoids criterio 2	0.01	**0.83 ± 0.07**	0.76 ± 0.06	0.77 ± 0.10	0.75 ± 0.07
	0.02	0.81 ± 0.06	0.67 ± 0.06	0.85 ± 0.09	0.63 ± 0.08
Todos los datos	-	**0.83 ± 0.06**	**0.78 ± 0.05**	0.76 ± 0.11	0.79 ± 0.06

Los resultados se presentan como media ± desviación estándar

Tabla 14.5 Porcentaje de pacientes con CAI eliminados por cada método en el set de entrenamiento, durante la validación cruzada.

Estadística	Corte	Clase 0	Clase 1	Total
IQ	-	23.1 ± 1.4	33.3 ± 1.9	24.8 ± 1.4
	10	3.3 ± 0.2	5.2± 0.3	3.6 ± 0.2
Tukey	-	8.7 ± 0.05	10.1± 1.1	9.0 ± 0.5
	10	1.2 ± 0.1	1.3 ± 0.2	1.3 ± 0.1
Log-IQ	-	22.8 ± 1.1	25.4 ± 2.0	23.2 ± 1.1
	10	**3.1 ± 0.2**	**3.7 ± 0.5**	3.2 ± 0.1
Score Z	-	35.0 ± 1.6	0.67 ± 0.06	32.6 ± 1.4
	10	5.3 ± 0.2	2.9 ± 1.3	4.9 ± 0.3
Score Z modificado	-	18.3± 0.05	24.5 ± 1.3	19.4± 0.5
	10	2.4 ± 0.1	3.5 ± 0.4	2.6 ± 0.1
Mahalanobis	-	19.6 ± 9.6.	17.4 ± 3.0	19.2 ± 8.1

Basado en el cluster	Peso	Clase 0	Clase 1	Total
K-media, índice de la silueta criterio 1	1.2	19.6 ± 9.6	17.4 ± 3.0	19.2 ± 8.1
	1.5	6.1 ± 5.1	1.9 ± 0.5	5.4 ± 4.2
	1.7	2.5 ± 2.6	0.3 ± 0.3	2.2 ± 2.2
	2	0.7 ± 0.9	0.0 ± 0.0	0.6 ± 0.8
K-media, c=2 criterio 1	1.2	29.7 ± 3.5	17.4 ± 3.0	27.6 ± 2.9
	1.5	11.9 ± 3.0	1.9 ± 0.5	10.2 ± 2.5
	1.7	5.5 ± 2.0	0.3 ± 0.3	4.7 ± 1.6
	2	1.7 ± 0.8	0.0 ± 0.0	1.4 ± 0.7
K- media criterio 2	0.05	0.3 ± 0.2	0.0 ± 0.0	0.3 ± 0.2
	0.06	1.1 ± 0.5	0.0 ± 0.0	0.9 ± 0.4
K-medoids, índice de la silueta criterio 1	1.2	25.0 ± 10.7	3.8 ± 2.0	21.5 ± 8.8
	1.5	12.9 ± 7.4	0.0 ± 0.0	10.8 ± 6.2
	1.7	9.5 ± 6.1	0.0 ± 0.0	7.9 ± 5.1
	2	3.1 ± 2.3	0.0 ± 0.0	2.5 ± 1.9
K-medoids, C=2, criterio 1	1.2	34.7 ± 0.7	3.8 ± 2.0	29.5 ± 0.7
	1.5	12.6 ± 0.6	0.0 ± 0.0	16.3 ± 0.5
	1.7	14.9 ± 1.1	0.0 ± 0.0	12.4 ± 0.9
	2	5.1 ± 0.4	0.0 ± 0.0	4.2 ± 0.4
K- medoids criterio 2	0.01	8.3 ± 2.1	0.0 ± 0.0	6.9 ± 1.7
	0.02	28.9 ± 3.9	1.8 ± 3.8	24.4 ± 3.6

Los resultados se presentan como media ± desviación estándar

14.12 Conclusiones y Resumen

El análisis univariado de “outliers” proporcionado en el caso de estudio mostró que se identificaron un gran número de “outliers” para cada variable dentro de clases predefinidas, lo que significa que la eliminación de todos los “outliers” identificados haría que se excluyeran gran parte de los datos. Por esta razón, el clasificar los “outliers” univariados de acuerdo a los valores de los scores y descartar sólo aquellos con mayores scores proporcionó los mejores resultados de la clasificación.

En general, ninguna de las técnicas de eliminación de “outliers” fue capaz de mejorar el desempeño de un modelo de clasificación. Al haber sido limpiada, estos resultados indican que la base de datos no contiene valores imposibles, los valores extremos se deben probablemente a una variación biológica más que a errores experimentales. Por lo tanto, los “outliers” en este estudio parecen contener información útil en sus valores extremos, y la exclusión automática resultó en una pérdida de esta información.

Algunos métodos de modelado ya se acomodan a los “outliers” por lo que tienen un impacto mínimo en el modelo, y pueden ser afinados para ser más o menos sensibles a ellos. Por lo tanto, antes que excluir los “outliers” del set de datos antes del paso de modelado, una estrategia alternativa podría ser utilizar modelos que sean robustos a los “outliers”, como la regresión robusta.

Puntos clave

1. Distinguir los valores atípicos entre los que son útiles o los que no son informativos no está claro.
2. En ciertos contextos, los “outliers” pueden representar información extremadamente valiosa que no se debe descartar.
3. Existen varios métodos que permiten identificar “outliers” posibles o probables, pero debe prevalecer el ojo experto antes de eliminarlos o corregirlos.

Nota: Las imágenes de este capítulo se encuentran disponibles a color en el ebook; disponible en www.hardineros.com

Apéndice: Código

El código usado en este capítulo para este libro se encuentra disponible en: https://github.com/mit–lpc/mitical data-book. Mayor información sobre el código se encuentra disponible en este sitio web.

Referencias

1. Barnett V, Lewis T (1994) Outliers in statistical data, 3rd edn. Wiley, Chichester.
2. Aggarwal CC (2013) Outlier analysis. Springer, New York.
3. Osborne JW, Overbay A (2004) The power of outliers (and why researchers should always check for them). Pract Assess Res Eval 9 (6): 1-12.
4. Hodge VJ, Austin J (2004) A survey of outlier detection methodologies. Artif Intell Rev 22 (2): 85-126.
5. Tukey J (1977) Exploratory data analysis. Pearson.
6. Shiffler RE (1988) Maximum Z scores and outliers. Am Stat 42 (1): 79-80.
7. Iglewicz B, Hoaglin DC (1993) How to detect and handle outliers. ASQC Quality Press.
8. Seo S (2006) A review and comparison of methods for detecting outliers in univariate datasets. 09 Aug 2006 [Internet]. Disponible en: http://d-scholarship. pitt.edu/7948/. [Consultado 07 de Febrero 2016].
9. Cook RD, Weisberg S (1982) Residuals and influence in regressión. Chapman and Hall, New York.
10. Penny KI (1996) Appropriate critical values when testing for a single multivariate outlier by using the Mahalanobis distance. Appl Stat 45 (1): 73-81.
11. Macqueen J (1967) Some methods for classification and analysis of multivariate observations. Presented at the proceedings of 5th Berkeley symposium on mathematical statistics and probability, pp 281-297.
12. Hu X, Xu L (2003) A comparative study of several cluster number selection criteria. En: Liu J, Cheung Y, Yin H (eds) Intelligent data engineering and automated learning. Springer, Berlin, pp 195-

202.
13. Jones RH (2011) Bayesian information criterion for longitudinal and clustered data. Stat Med 30 (25): 3050-3056.
14. Cherednichenko S (2005) Outlier detection in clustering.
15. Provan D (2010) Oxford handbook of clinical and laboratory investigation. OUP Oxford.
16. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44 (3): 837-845.

Capítulo 15

ANÁLISIS EXPLORATORIO DE DATOS

Matthieu Komorowski, Dominic C. Marshall,
Justin D. Salciccioli e Yves Crutain

Objetivos de Aprendizaje

- ¿Por qué el Análisis Exploratorio de Datos es importante durante la exploración inicial de un set de datos?
- ¿Cuáles son las herramientas más importantes del Análisis Exploratorio de Datos gráfico y no gráfico?

15.1 Introducción

El análisis exploratorio de datos (AED) es un paso esencial en el análisis de cualquier investigación. El objetivo principal del análisis exploratorio es examinar la distribución de los datos, los valores atípicos (*outliers)* y anomalías para dirigir la prueba específica de su hipótesis. También provee herramientas para la generación de hipótesis al visualizar y comprender los datos generalmente a través de representaciones gráficas [1]. El AED tiene como objetivo ayudar al analista de datos al reconocimiento de los patrones naturales. Finalmente, las técnicas de selección de características a menudo caen dentro del AED. Desde el trabajo fundamental de Tukey en 1977, el AED ha ganado un gran número de seguidores como la metodología *gold standard* para analizar un set de datos [2, 3]. Según Howard Seltman (Universidad de Carneige Mellon), "en términos generales, cualquier método de análisis de datos que no incluya el modelado estadístico formal y la inferencia cae dentro del término de análisis exploratorio de datos" [4].

El AED es un paso temprano fundamental después de la recopilación de los datos (ver Capítulo 11) y del preprocesamiento (ver Capítulo 12), donde los datos son sencillamente visualizados, graficados, manipulados, sin ningún supuesto, con el fin de ayudar a evaluar la calidad de los datos y construir modelos. La mayoría de las técnicas del AED son de naturaleza gráfica con algunas técnicas cuantitativas. La razón de la gran confianza en los gráficos es que por su propia naturaleza, el papel principal del AED es explorar, y los gráficos dan al analista de datos un poder sin igual para

hacerlo, al mismo tiempo que están preparados para comprender los datos. Existen muchas formas para categorizar las diversas técnicas del AED". [5].

El lector interesado encontrará más información en los libros de texto de Hill y Lewicki [6] o en el e-Handbook NIST/SEMATECH [1]. Los paquetes R relevantes están disponibles en el sitio web CRAN [7].

Los objetivos del AED pueden resumirse como sigue:

1. Maximizar el conocimiento de la base de datos / la comprensión de la estructura de la base de datos;
2. Visualizar relaciones potenciales (dirección y magnitud) entre variables de exposición y resultados;
3. Detectar valores atípicos y anomalías (valores que son significativamente diferentes respecto a otras observaciones);
4. Desarrollar modelos parsimoniosos (un modelo predictivo o explicativo que funcione con la menor cantidad de variables de exposición posibles) o una selección preliminar de modelos apropiados;
5. Extraer y crear variables clínicas relevantes.

Los métodos de AED pueden clasificarse como:

- Métodos gráficos o no gráficos
- Métodos univariados (solo una variable, exposición o resultado) o multivariados (varias variables de exposición solas o con una variable de resultado).

15.2 Parte 1 - Conceptos Teóricos

15.2.1 Técnicas sugeridas para el Análisis Exploratorio de Datos

En las tablas 15.1 y 15.2 se sugieren algunas técnicas de AED según el tipo de datos y el objetivo del análisis.

Tabla 15.1 Técnicas sugeridas de AED según el tipo de datos

Tipo de datos	**Técnicas sugeridas**
Categórico	Estadísticas descriptivas
Univariados	Gráficos de líneas, histogramas

continuos	
Bivariados continuos	Gráficos de dispersión 2D
Matrices 2D	Mapas de calor
Multivariado: trivariado	Gráficos de dispersión 3D o gráficos de dispersión 2D con una 3ra variable representada en otro color, forma o tamaño.
Grupos multivariados	Gráficos de cajas contiguos

Tabla 15.2 Técnicas de AED más útiles de acuerdo al objetivo

Objetivo	**Técnicas de AED sugeridas**
Tener una idea de la distribución de una variable	Histograma
Encontrar valores atípicos “outliers”	Histograma, gráficos de dispersión, gráficos de cajas y bigotes
Cuantificar la relación entre dos variables (una de exposición y una de resultado)	Gráficos de dispersión 2D +/ ajuste de curvas. Covarianza y correlación
Visualizar la relación entre dos variables de exposición y una variable de resultado	Mapa de calor
Visualización de datos de alta dimensión	t-SNE o ACP + gráficos de dispersión 2D/3D

t-SNE Incrustación estocástica de vecinos con corrección t, ACP análisis de componente principal

15.2.2 Análisis Exploratorio de Datos No Gráfico

Estos métodos no-gráficos proporcionarán una visión de las características y la distribución de la(s) variable(s) de interés.

AED no-gráfico univariado

Tabulación de datos categóricos (Tabulación de la frecuencia de cada categoría)

Un método simple de AED no gráfico univariado para variables categóricas es construir una tabla que contenga el conteo y la fracción (o frecuencia) de los datos de cada categoría. Se muestra un ejemplo de tabulación en el caso de estudio (Tabla 15.3).

Tabla 15.3 Ejemplo de Tabla de Tabulación

	Conteo de grupos	**Frecuencia (%)**
Bola verde	15	75
Bola roja	5	25
Total	20	100

Características de los datos cuantitativos: Tendencia Central, Dispersión, Forma de la distribución (Asimetría, Curtosis)

Las estadísticas muestrales expresan las características de una muestra usando un conjunto limitado de parámetros. Son vistos generalmente como estimaciones de los parámetros de la población de la que procede la muestra. Estas características pueden expresar la tendencia central de los datos (media aritmética, mediana, moda), su dispersión (varianza, desviación estándar, rango intercuartílo, valor máximo y mínimo) o algunas características de su distribución (asimetría, Curtosis). Muchas de esas características pueden verse fácilmente cualitativamente en un histograma (ver abajo). Nótese que estas características pueden ser usadas solamente para variables cuantitativas (no categóricas).

Parámetros de Tendencia Central

La media aritmética, o simplemente llamada la media es la suma de todos los datos dividido por el número de valores. La mediana es el valor medio en una lista que contiene todos los valores ordenados. Como la mediana es poco afectada por los valores extremos y atípicos, se dice que es más “robusta” que la media (Fig 15.1).

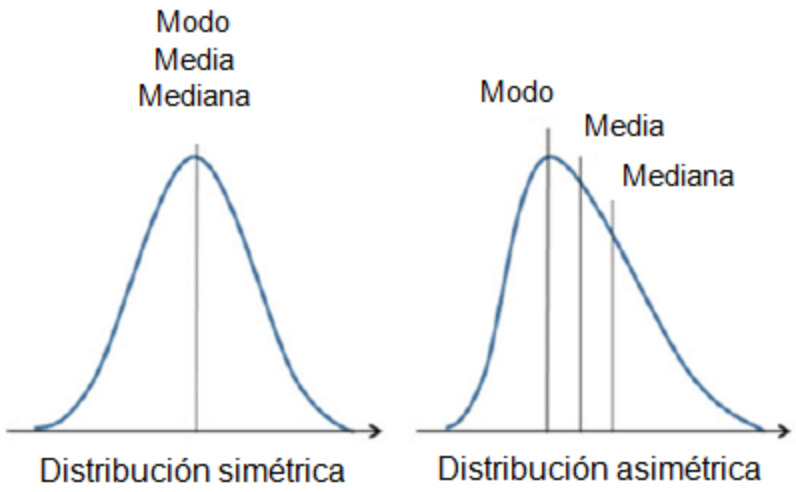

Fig 15.1. Distribución simétrica versus asimétrica, mostrando modo, media y mediana.

Varianza

Cuando se calcula sobre la totalidad de los datos de una población (que raramente ocurre), la varianza σ^2 se obtiene al dividir la suma de las diferencias entre cada valor y la media al cuadrado por n, el tamaño de la población.

La fórmula de la varianza de los datos de una muestra convencionalmente tiene n-1 en el denominador en lugar de n para lograr la propiedad de "insesgado", lo que a grandes rasgos significa que cuando se calcula para muchas muestras aleatorias diferentes de la misma población, la media debe coincidir con la media de la población correspondiente. S^2 es un estimador no sesgado de la varianza de la población σ^2.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \underline{x})^2}{(n-1)} \qquad (15.1)$$

La desviación estándar es simplemente la raíz cuadrada de la varianza. Por lo tanto tiene las mismas unidades que los datos originales, lo que ayuda a hacerla más interpretable.

La desviación estándar de la muestra suele representarse con el símbolo s. Para una distribución teóricamente Gaussiana, la media más o menos 1,2 ó 3 desvíos estándar contiene el 68.3, 95.4 u 99.7% de la densidad de probabilidad, respectivamente.

Rango Intercuartílico (RIC)

El RIC se calcula utilizando los límites de los datos situados entre el 1er y el 3er cuartil. Por favor, consulte el Capítulo 13 “Ruido versus Valores atípicos” para mayor detalle sobre el RIC.

$$RIC=Q_3-Q_1 \quad (15.2)$$

De la misma manera que la mediana es más robusta que la media, el RIC es una medida más robusta de dispersión que la varianza y la desviación estándar y por lo tanto, debe preferirse para las distribuciones pequeñas o asimétricas.

Regla importante:

- **Distribución simétrica** (no necesariamente normal) **y N > 30:** expresar los resultados como la media +/– desviación estándar.
- **Distribución asimétrica o N < 30 ó evidencia de valores atípicos:** usar mediana +/– RIC, lo cual es más robusto.

Asimetría (Skewness) /Curtosis

La asimetría o skewness es una medida de la asimetría de una distribución. Curtosis es una estadística sumaria que brinda información sobre las colas de la distribución (los valores más pequeños y más grandes). Ambas medidas pueden ser utilizadas como una forma de comunicar información sobre la distribución de los datos cuando no pueden usarse métodos gráficos. Más información sobre estas medidas puede encontrarse en [9].

Resumen

Proporcionamos como referencia algunas de las funciones comunes en lenguaje R para la generación de estadísticas sumarias relacionadas con medidas de tendencia central (Tabla 15.4).

Probando la Distribución

Existen varios métodos no gráficos para evaluar la normalidad de un set de datos (si proviene de una muestra con distribución normal), como la prueba de Shapiro-Wilk por ejemplo. Por favor, refiérase a la función

llamada “Distribution” en el repositorio Git-Hub de este libro (ver el apéndice de código al final de este Capítulo).

Tabla 15.4 Principales funciones de R para las medidas básicas de tendencia central y variabilidad.

Función	**Descripción**
summary (x)	Descripción general de un vector
max (x)	Valor máximo
mean (x)	Media
Median (x)	Mediana
min (x)	Valor mínimo
sd (x)	Desviación estándar
var (x)	Variación, medida de dispersión o dispersión de los valores
IQR (x)	Rango Intercuartílico

Encontrando Outliers

Diversos métodos estadísticos para la detección de “*outliers”* caen dentro de las técnicas del AED, como el método de Tukey, score Z, residuos estudentizados, etc [8]. Por favor, referirse al Capítulo 14 “Ruido versus Valores atípicos” para más detalle sobre este tema.

AED no-gráfico multivariado

Tabulación Cruzada

La tabulación cruzada representa la técnica de AED no-gráfica bivariable básica. Es una extensión de la tabulación que se usa para datos categóricos y datos cuantitativos con solo unas pocas variables. Para dos variables, construir una tabla de dos entradas con encabezados de columna que coincidan con los niveles de una variable y los encabezados de fila que coincidan con los niveles de la otra variable, luego llenar los cuadros con el conteo de todos los individuos que coincidan en los niveles de las dos

variables. Las dos variables pueden ser ambas de exposición, ambas variables de resultados, o una de cada una.

Covarianza y correlación

Covarianza y correlación miden el grado de relación entre dos variables al azar y expresan cuánto cambian juntas (Fig15.2)

La covarianza se calcula de la siguiente manera:

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \qquad (15.3)$$

donde *x* e *y* son las variables, *n* es el número de puntos de datos en la muestra, $\overline{x}$ la media de la variable *x* e $\overline{y}$ a media de la variable y.

Una covarianza positiva significa que las variables están positivamente relacionadas (se mueven juntas en la misma dirección), mientras que una covarianza negativa quiere decir que las variables están inversamente relacionadas. Un problema con la covarianza es que su valor depende de la escala de valores de las variables aleatorias. Cuanto mayores sean los valores de x e y, mayor será la covarianza. Esto hace imposible por ejemplo comparar covarianzas de un set de datos con diferentes escalas (ej., libras y pulgadas). Este problema puede resolverse dividiendo la covarianza por el producto del desvío estándar de cada variable aleatoria, lo que da el coeficiente de correlación de Pearson.

La correlación es entonces una versión a escala de la covarianza, utilizada para evaluar la relación lineal entre dos variables y se calcula utilizando la fórmula siguiente.

$$Cor(x,y) = \frac{Cov(x,y)}{s_x s_y} \qquad (15.4)$$

donde *Cov (x, y)* es la covarianza entre *x* e *y*, y s_x, s_y son las desviaciones estándar de las muestras *x* e *y*.

El significado del coeficiente de correlación entre dos variables distribuidas normalmente puede evaluarse utilizando la transformación Z de Fisher (vea la función cor. test en R para más detalles). Existen otras pruebas

para medir la relación no-paramétrica entre dos variables, tales como la rho de Spearman o la tau de Kendall.

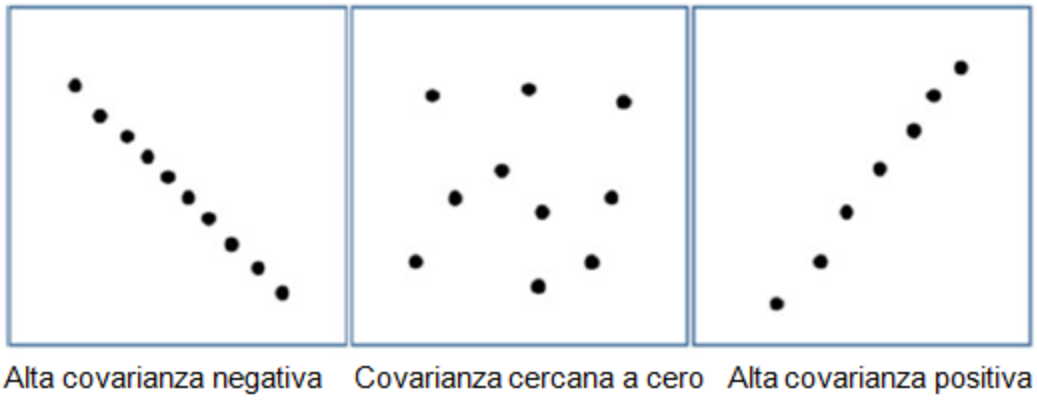

Figura 15.2 Ejemplos de covarianza para 3 set de datos distinto.

15.2.3 AED gráfico

AED gráfico univariado

Histogramas

Los histogramas están entre las técnicas más utilizadas en el AED, y permiten obtener una visión de los datos, incluyendo distribución, tendencia central, dispersión, modalidad y "*outliers".*

Los histogramas son gráficos de barras de conteo versus subgrupos de una variable de exposición. Cada barra representa la frecuencia (recuento) o proporción (el recuento dividido por el recuento total) de casos para un rango de valores. El rango de datos para cada barra se conoce como caja o contenedor. Los histogramas dan una impresión inmediata de la forma de la distribución (simétrica, uni/multimodal, asimétrica, "*outliers"*). El número de contenedores influye en gran medida en el aspecto final del histograma; una buena práctica es probar diferentes valores, generalmente desde 10 hasta 50. Algunos ejemplos de histogramas se muestran debajo así como en los estudios de casos. Por favor, refiérase a la función llamada "Density" en el repositorio Git-Hubde este libro (ver el apéndice de código al final de este Capítulo) (Figs. 15.3 y 15.4).

Los histogramas permiten confirmar que una transformación en los datos fue exitosa. Por ejemplo, si usted necesita realizar una transformación

logarítmica de un conjunto de datos, es interesante graficar el histograma de distribución de los datos antes y después de la transformación (Fig 15.5).

Los histogramas son interesantes para encontrar *"outliers"*. Por ejemplo, la oximetría de pulso puede expresarse en fracciones (rango entre 0 y 1) o porcentaje, en los registros médicos. La Figura 15.6 es un ejemplo de un histograma que muestra la distribución de la oximetría de pulso, mostrando claramente la presencia de *"outliers"* expresados en fracción en lugar de como porcentaje.

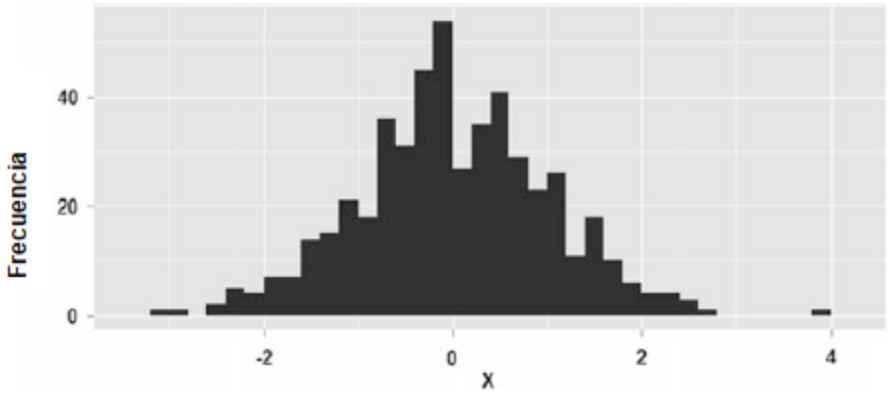

Fig 15.3 Ejemplo de un histograma.

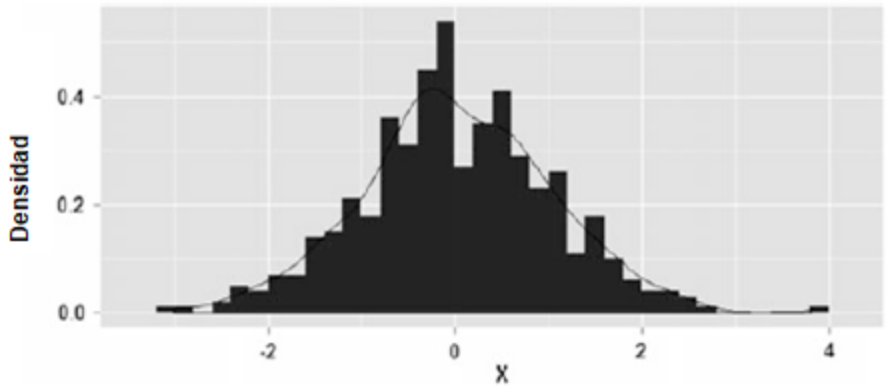

Fig 15.4 Ejemplo de un histograma con estimación de la densidad.

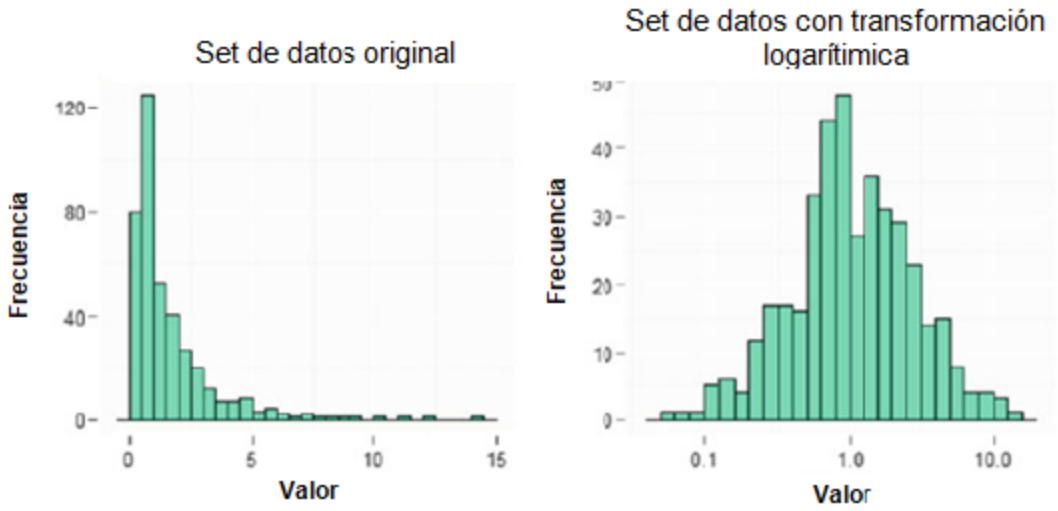

Fig 15.5 Ejemplo del efecto de una transformación logarítmica en la distribución de un set de datos.

Gráficos de Tallo (Stem plots)

Los gráficos de tallo y hojas (también llamados gráficos de tallo) son una sustitución simple de los histogramas. Muestran todos los valores de los datos y la forma de la distribución. Para un ejemplo, por favor, refiérase a la función llamada "StemPlot" en el repositorio Git-Hub de este libro (ver el apéndice de código al final de este Capítulo) (Fig. 15.7).

Fig 15.6 Distribución de la oximetría de pulso.

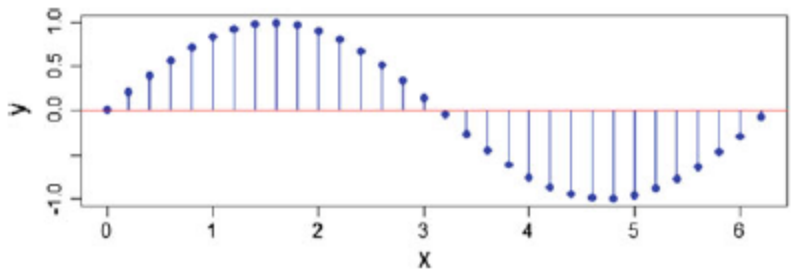

Fig 15.7 Ejemplo de gráfico de tallo.

Gráfico de caja (Boxplots)

Los gráficos de caja son interesantes para representar información acerca de la tendencia central, simetría, asimetría y "outliers", pero pueden esconder algunos aspectos de los datos como la multimodalidad. Los gráficos de caja son una técnica del AED excelente porque se basan en estadísticas sólidas como la mediana y el RIC.

La figura 15.8 muestra un gráfico de caja con anotaciones que explican cómo se construye. El rectángulo central se limita por Q1 y Q3, con la línea media representando la mediana de los datos. Los bigotes se dibujan, en cada dirección, hasta el punto más extremo que está a menos de 1.5 RIC más allá de la bisagra correspondiente (Q1 ó Q3). Los valores más allá de 1.5 RIC son considerados *"outliers".*

Los *"outliers"* identificados por un gráfico de caja, los que pueden ser llamados "valores atípicos del gráfico de cajas" se definen como cualquier punto con más de 1.5 RIC por encima de Q3 ó más de 1.5 RIC por debajo de Q1. Esto no indica por sí mismo un problema con esos puntos de datos. Los gráficos de cajas son una técnica exploratoria, y se debe considerar la designación de un "valor atípico del gráfico de cajas" como una simple sugerencia de que los puntos pueden ser errores o inusuales. Además, los puntos no designados como "valores atípicos del gráfico de cajas" pueden ser también errores. También es importante comprender que el número de "valores atípicos del gráfico de cajas" depende en gran medida del tamaño de la muestra. De hecho, para los datos que están perfectamente distribuidos normalmente, se espera que un 0,70% (cerca de 1 en 140 casos) sean "valores atípicos del gráfico de cajas", con aproximadamente la mitad en cualquier dirección.

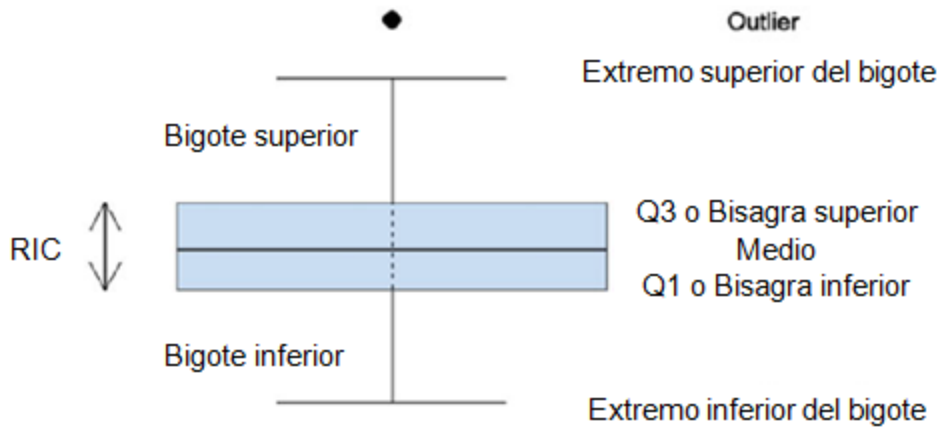

Fig 15.8 Ejemplo de gráfico de caja con anotaciones.

Gráficos de Líneas 2D

Los gráficos de líneas 2D representan gráficamente los valores en una matriz en el eje y, a intervalos regulares en el eje x (Fig 15.9).

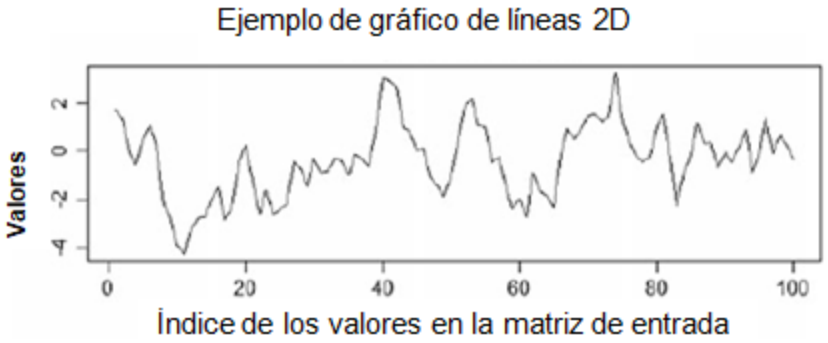

Fig 15.9 Ejemplo de gráfico de líneas 2D.

Gráficos de Probabilidad (Gráfica Cuantil-Normal/Gráfica QN, Gráfica Cuantil-Cuantil/Gráfica QQ)

Los gráficos de probabilidad son pruebas gráficas para evaluar si algunos datos siguen una distribución particular. Se utilizan con mayor frecuencia para probar la normalidad de un set de datos, ya que muchas pruebas estadísticas suponen que las variables de exposición tienen una distribución aproximadamente normal. Estos gráficos también se usan para examinar residuos en modelos que se basan en el supuesto de normalidad de los residuos (ANOVA o análisis de regresión por ejemplo).

La interpretación de un gráfico QN es visual (Fig 15.10): o bien los puntos caen aleatoriamente alrededor de la línea (conjunto de datos distribuidos normalmente) o siguen un patrón de curva en lugar de seguir la línea (no-normalmente). Los gráficos QN también se utilizan para identificar asimetrías, curtosis, colas pesadas, valores atípicos, bimodalidad, etc.

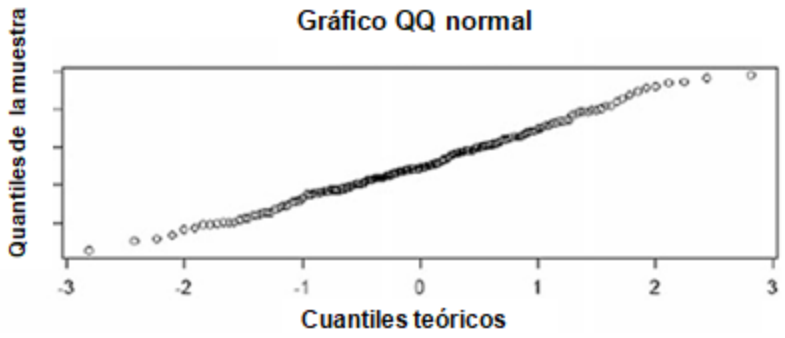

Fig 15.10 Ejemplo de gráfico QQ.

Además de los gráficos de probabilidad, hay muchas pruebas estadísticas cuantitativas (no gráficas) para evaluar la normalidad, tales como Pearson Chi^2, Shapiro-Wilk, y Kolmogowov-Smirnov.

La desviación de la normalidad de la distribución observada hace que muchas herramientas estadísticas poderosas sean inútiles. Nótese que algunos set de datos pueden ser transformados a una distribución más normal, en particular con transformación logarítmica y transformaciones de raíz cuadrada. Si un set de datos es muy asimétrico, otra opción es discretizar sus valores en un conjunto finito.

AED gráfico multivariado

Gráficos de cajas contiguos

Representar varios gráficos de cajas lado a lado permite una comparación sencilla de las características de diversos grupos de datos (ejemplo Fig. 15.11). En el caso de estudio se muestra un ejemplo de este tipo de gráfico.

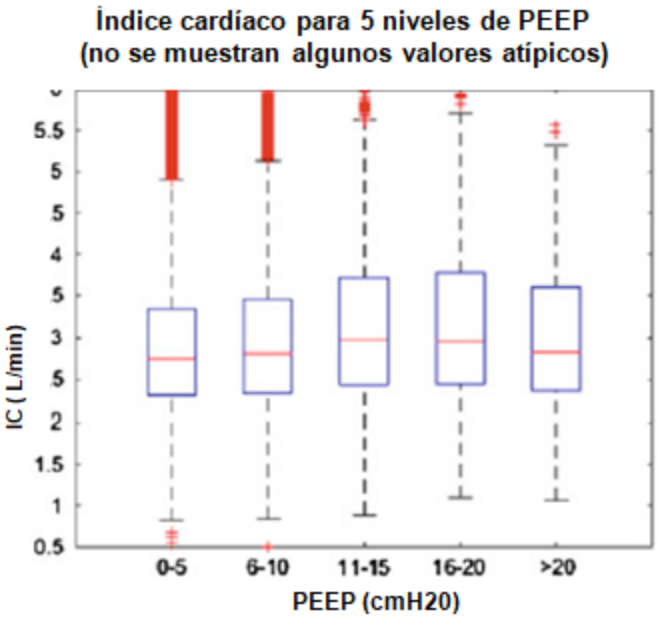

Fig 15.11 Diagrama de cajas contiguos que muestra el índice cardíaco para cinco niveles de Presión Positiva al Final de la Espiración (PEEP).

Gráficos de dispersión (Scatterplots)

Los gráficos de dispersión se construyen utilizando dos variables continuas, cuantitativas ordinales o cuantitativas discretas (Fig. 15.12). La coordenada de cada punto corresponde a una variable. Se puede aumentar la complejidad hasta cinco dimensiones utilizando otras variables diferenciando el tamaño, forma o color de los puntos de datos.

Los gráficos de dispersión también pueden utilizarse para representar datos de alta dimensión en 2 ó 3D (Fig 15.13), usando la incrustación estocática de vecinos con corrección t (t-SNE) o análisis de componente principal (ACP). t-SNE y ACP son características de reducción de dimensiones utilizadas para reducir sets de datos complejos en dos (t-SNE) o más (ACP) dimensiones.

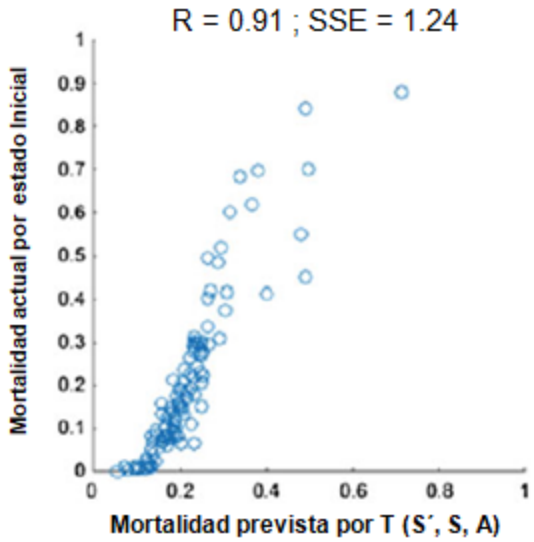

Fig 15.12 Gráfico de dispersión que muestra un ejemplo de la mortalidad real por la tasa de mortalidad predicha.

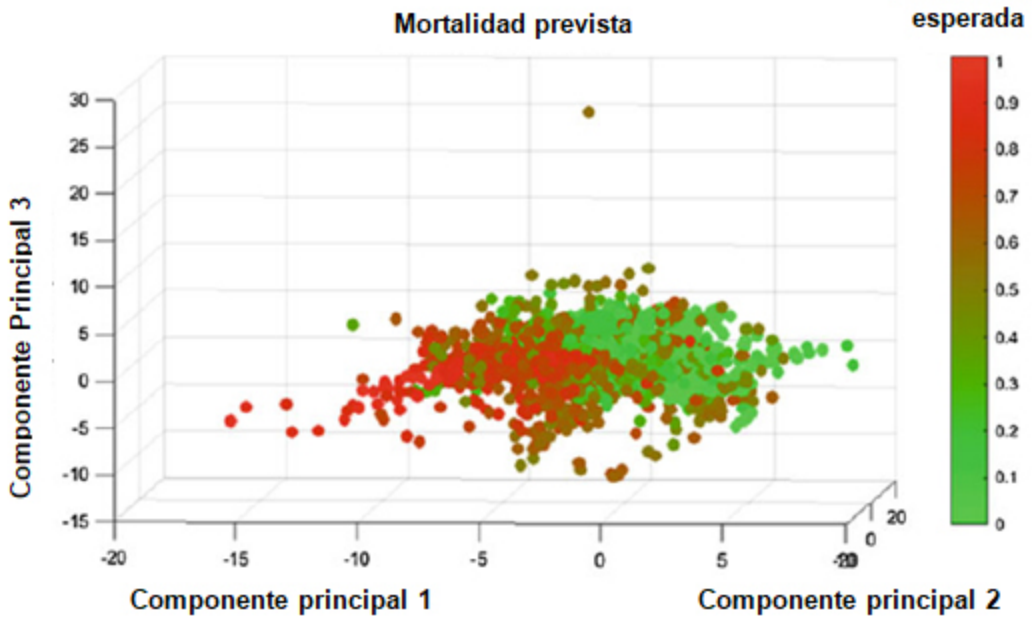

Fig 15.13 Representación 3D de las primeras tres dimensiones de un ACP.

Para las variables binarias (ej., mortalidad a 28 días vs. puntaje SOFA), los gráficos de dispersión 2D no son de gran ayuda (Fig 15.14, izquierda). Dividiendo el conjunto de datos en grupos (en nuestro ejemplo: un grupo por punto SOFA), y trazando el valor promedio del resultado en cada grupo, el gráfico de dispersión se vuelve una herramienta muy poderosa, capaz por

ejemplo de identificar la relación entre una variable y un resultado (Fig 15.14, derecha).

Ajuste de Curvas

El ajuste de curvas es una forma de cuantificar la relación entre dos variables o el cambio en los valores a lo largo del tiempo (Fig 15.15). El método más común para el ajuste de curvas se basa en minimizar la suma de errores al cuadrado (SEC) entre los datos y la función ajustada. Por favor, refiérase a la función "Linear Fit" para crear pendientes de regresión lineal en R.

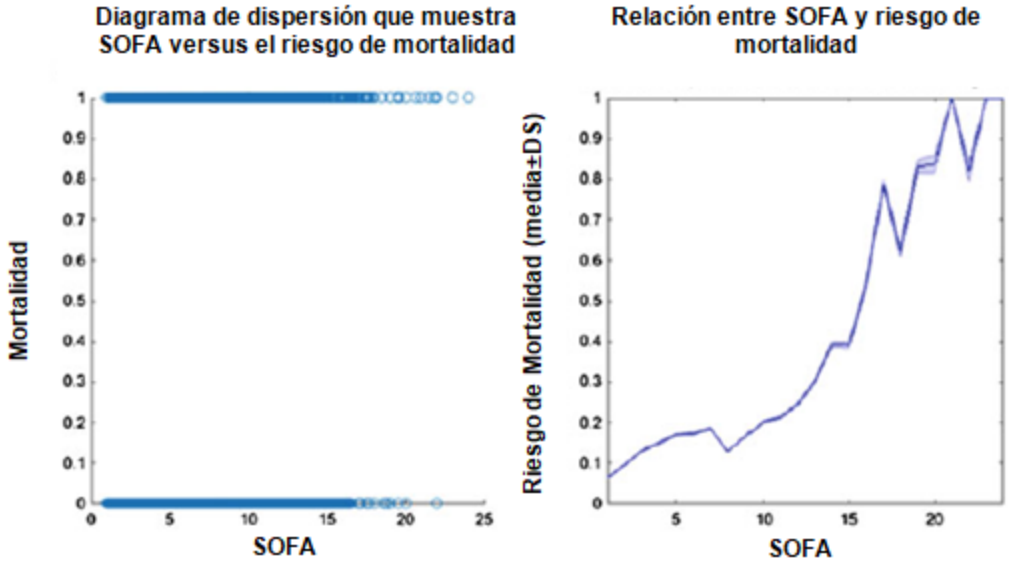

Fig. 15.14 Gráfico de puntaje SOFA versus riesgo de mortalidad.

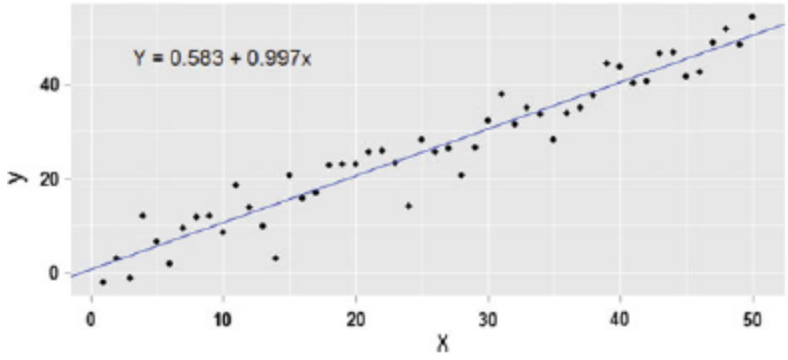

Fig. 15.15 Ejemplo de regresión lineal.

Relaciones más complicadas

Muchos fenómenos de la vida real no se explican adecuadamente mediante una relación directa. Cada vez existen más métodos y algoritmos para lidiar con este problema. Entre los más comunes figuran:

- Añadir variables explicativas transformadas, por ejemplo, añadiendo x^2 ó x^3 al modelo.
- Usar otros algoritmos para manejar relaciones más complejas entre variables (ej., modelos aditivos generalizados, regresión por spline, máquinas de soporte vectorial, etc.).

Mapas de calor y Gráficos de Superficie 3D

Los mapas de calor son simplemente una cuadrícula 2D construida a partir de una matriz 2D, cuyo color depende del valor de cada celda. El conjunto de datos debe corresponder a una matriz 2D cuyas celdas contienen los valores de la variable de resultado. Esta técnica es de utilidad cuando se quiere representar el cambio de una variable de resultado (ej., duración de la estadía) como una función de otras dos variables (ej., edad y score SOFA).

El mapa de colores puede ser personalizado (ej., arcoiris o escala de grises). Interesantemente, la función de Matlab *imagesc* escala los datos a la gama completa de colores. Su equivalente 3D es un gráfico de red o un gráfico de superficie (Fig. 15.16).

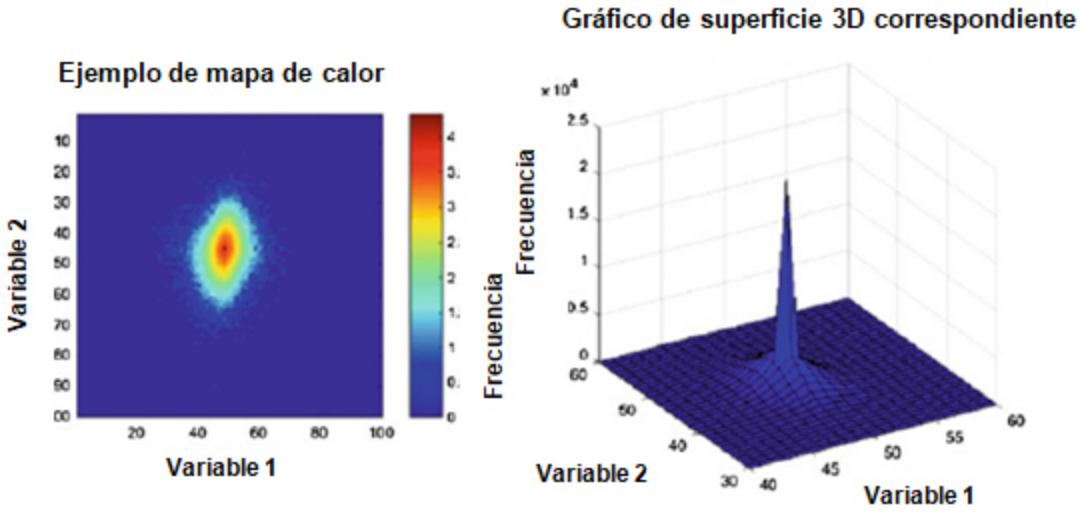

Fig 15.16 Mapa de calor (izquierda) y gráfico de superficie (derecha).

15.3 Parte 2 - Caso de Estudio

Este caso de estudio se refiere a la investigación que evaluaba el efecto de la colocación de catéteres arteriales invasivos (CAIs) en pacientes hemodinámicamente estables con falla respiratoria en cuidados intensivos, de la base de datos MIMIC-II.

Para este caso de estudio, se utilizaron varios aspectos del AED:

- Primero se tabularon los datos categóricos.
- Después se generaron las estadísticas sumarias para describir las variables de interés.
- Se usó AED gráfico para generar histogramas para visualizar los datos de interés.

15.3.1 Análisis Exploratorio de Datos no gráfico

Tabulación

Para analizar, visualizar y probar la asociación o independencia de las variables categóricas, primero deben ser tabuladas. Cuando se generan tablas, cualquier dato que falte se contará en una categoría separada, "NA" (No disponible *"Not Available"*). Por favor, refiérase al Capítulo 13 "Datos faltantes" para enfoques en el manejo de este problema. Hay diversos métodos para crear tablas de frecuencia o contingencia en R, como por

ejemplo, la tabulación de variables de resultados para mortalidad, como se demuestra en el caso de estudio. Refiérase a la función "Tabulate" en el repositorio Git-Hub de este libro (ver el apéndice de código al final de este Capítulo) para detalles sobre cómo computar frecuencias de resultados para diferentes variables.

Pruebas Estadísticas

Hay múltiples pruebas estadísticas disponibles en R. Referimos al lector al Capítulo 16 "Análisis de datos" para información adicional en el uso de pruebas relevantes en R. Para ejemplos de una simple prueba de Chi-cuadrado, por favor consulte la función "Chi-squared" que se encuentra en el repositorio de Git-Hub de este libro (ver el apéndice de código al final de este Capítulo). En nuestro ejemplo, la hipótesis de independencia entre muerte en la UCI y la colocación de CAI no puede rechazarse ($p > 0.05$). Al contrario, la relación de dependencia entre mortalidad a 28 días y el CAI es rechazada.

Estadísticas Sumarias

Las estadísticas sumarias descritas anteriormente incluyen, frecuencia, media, mediana, modo, rango, rango intercuartílo, valores máximos y mínimos. En la tabla 15.5 se muestra un detalle de las estadísticas sumarias de la demografía de los pacientes, los signos vitales, los resultados de laboratorio y las comorbilidades. Por favor consulte la función llamada "EDA Summary" en el repositorioGit-Hub de este libro (ver el apéndice de código al final de este Capítulo) (Tabla 15.5).

Tabla 15.5 Comparación entre las 2 cohortes de estudio (sub muestra de variables exclusivamente)

Variables	**Cohorte Completa (N = 1776)**		
	Sin –CAI	**Con CAI**	**Valor-p**
Tamaño	984 (55.4%)	792 (44.6%)	NA
Edad (años)	51 (35-72)	56 (40-73)	0.009
Género (femenino)	344 (43.5%)	406 (41.3%)	0.4
Peso (kg)	76 (65-90)	78 (67-90)	0.08

Puntaje SOFA	5 (4-6)	6 (5-8)	<0.0001
Co-morbidades			
CHF	97 (12.5%)	116 (11.8%)	0.7
…	…	…	…
Pruebas de laboratorio			
Recuento de Globulos Blancos	10.6 (7.8-14.3)	11.8 (8.5-15.9)	<0.0001
Hemoglobina (g/dL)	13 (11.3-14.4)	12.6 (11-14.1)	0.003
…	…	…	…

Cuando se generan cohortes separadas basadas en una variable común, en este caso la presencia de un catéter arterial invasivo, se presentan las estadísticas sumarias para cada cohorte.

Es importante identificar cualquier diferencia en las características basales del sujeto. Los beneficios de esto son dobles: primero es util para identificar variables potencialmente confundidoras que contribuyan a un resultado además de la variable predictiva (de exposición). Por ejemplo, si la mortalidad es la variable de resultado entonces diferencias en la gravedad de las enfermedades entre las cohortes puede dar cuenta total o parcialmente de cualquier variación de la mortalidad. Identificar estas variables es importante ya que es posible intentar controlarlas mediante métodos de ajuste como regresión logística multivariable. En segundo lugar, puede permitir la identificación de variables que están asociadas con la variable predictora enriqueciendo nuestra comprensión sobre el fenómeno que estamos observando.

La extensión analítica de la identificación de cualquier diferencia utilizando medianas, medias y visualización de datos es probar las diferencias estadísticamente significativas en cualquier característica de un sujeto determinado utilizando por ejemplo la prueba de suma de rangos de Wilcoxon (Wilcoxon-Rank sum test). Consulte el capítulo 16 para mayor detalles sobre pruebas de hipótesis.

15.3.2 Análisis Exploratorio de Datos Gráfico

La representación gráfica de un conjunto de datos de interés es la principal característica del análisis exploratorio.

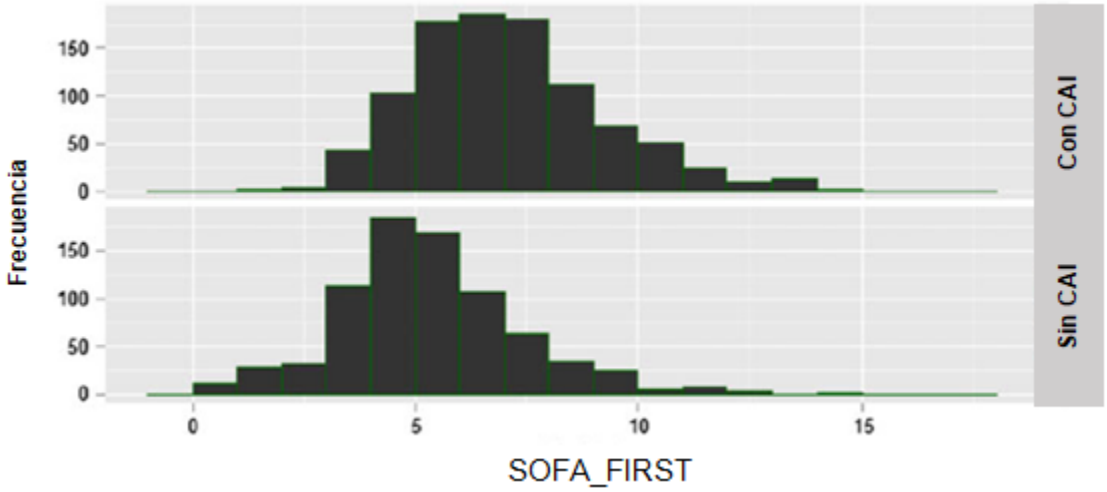

Fig. 15.17 Histogramas de scores SOFA de acuerdo a la presencia de cáteter arterial invasivo.

Histogramas

Los histogramas son considerados como la columna vertebral del AED para los datos continuos. Pueden utilizarse para ayudar al investigador a comprender las variables continuas y proveer información clave como su distribución. Descrito en *ruido y valores atípicos,* el histograma permite al investigador visualizar dónde se ubica el grueso de los puntos de datos entre los valores máximos y mínimos. Los histogramas pueden también permitir una comparación visual de una variable entre cohortes. Por ejemplo, para comparar la gravedad de la enfermedad entre la cohorte de pacientes, los histogramas del score SOFA se pueden trazar uno al lado del otro (Fig 15.17). Se da un ejemplo de esto en el código de este capítulo utilizando el código “side-by-side histogram” (ver el apéndice de código al final de este Capítulo).

Gráfico de caja y ANOVA

Fuera del ámbito de este caso de estudio, el usuario puede estar interesado en el análisis de la varianza. Cuando se realiza el AED una forma efectiva de visualizarla es a través del uso de gráficos de caja. Por ejemplo, para explorar las diferencias en la presión arterial basadas en la severidad

de la enfermedad los sujetos pueden ser categorizados por gravedad de la enfermedad con valores de presión sanguínea en la línea de base trazada (Fig. 15.18). Por favor consulte la función llamada “Box Plot” en el repositorioGit-Hub de este libro (ver el apéndice de código al final de este Capítulo).

El gráfico de caja muestra unos pocos valores atípicos los cuales pueden ser interesantes para explorar individualmente, y que las personas con un score SOFA alto (>10) tienden a tener una presión sanguínea menor que las personas con un score SOFA menor.

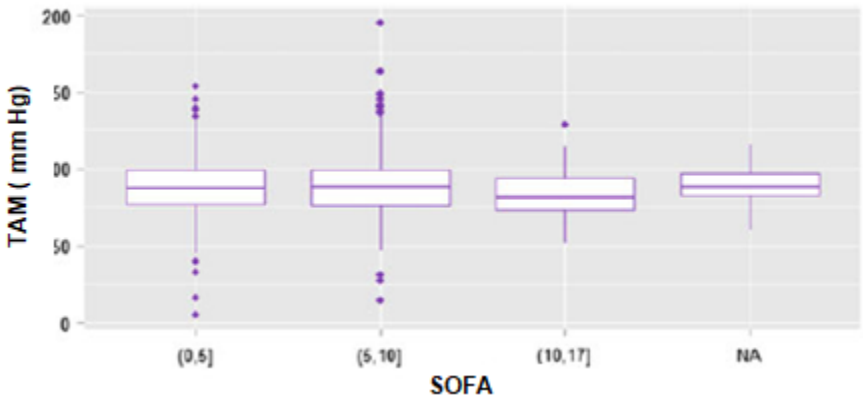

Fig 15.18 Gráfico de cajas lado a lado de PAM
para distintos niveles de gravedad en la admisión.

15.4 Conclusión

En resumen, el AED es un paso esencial en muchos tipos de investigaciones pero es de especial utilidad cuando se analizan historias clínicas electrónicas. Las herramientas descritas en este capítulo pueden permitir al investigador entender mejor las características de un set de datos y también generar nuevas hipótesis.

Puntos clave

1. Siempre empiece a explorar el set de datos con una mente abierta al descubrimiento.
2. El AED permite aprehender mejor las características y posibles problemas de un set de datos.
3. El AED es un paso clave en la generación de hipótesis de investigación.

Nota: Las imágenes de este capítulo se encuentran disponibles a color en el ebook; disponible en www.hardineros.com

Apéndice: Código

El código utilizado en este capítulo se encuentra disponible en el repositorio de GitHub de este libro: https://github.com/MIT-LCP/critical-data-book. En este sitio web se encuentra disponible más información del código.

Referencias

1. Natrella M (2010) NIST/SEMATECH e-Handbook of Statistical Methods. NIST/SEMATECH.
2. Mosteller F, Tukey JW (1977) Data analysis and regressión. Addison-Wesley Pub. Co., Boston.
3. Tukey J (1977) Exploratory data analysis. Pearson, London.
4. Seltman HJ (2012) Experimental design and analysis. [Internet] Disponible en http://www.stat.cmu.edu/*hseltman/309/Book/Book.pdf.
5. Kaski, Samuel (1997) “Data exploration using self-organizing maps”. Acta polytechnica scandinavica: Mathematics, computing and management in engineering series no. 82.1997.
6. Hill T, Lewicki P (2006) Statistics: methods and applications: a comprehensive reference for science, industry, and data mining. StatSoft, Inc., Tulsa.
7. CRAN (2016) The Comprehensive R archive network-packages. Contributed Packages, 10 Jan 2016 [Internet]. Disponible en: https://cran. r-project.org/web/packages/. [Consultado 10 de Enero 2016].
8. Grubbs F (1969) Procedures for detecting outlying observations in samples. Technometrics 11 (1).
9. Joanes DN, Gill CA (1998) Comparing measures of sample skewness and kurtosis. The Statistician 47:183-189.

Capítulo 16

ANÁLISIS DE DATOS

Jesse D. Raffa, Marzyeh Ghassemi, Tristan Naumann,
Mengling Feng y Douglas Hsu

Objetivos de Aprendizaje

- Entender cómo el objetivo del estudio y los tipos de datos determinan el tipo de análisis de datos.
- Entender los conceptos básicos de las tres técnicas más comunes usadas en los estudios que involucran datos en salud.
- Resolver caso de estudio para cumplir con el objetivo de estudio e interpretar los resultados.

16.1 Introducción al Análisis de Datos

16.1.1 Introducción

Este capítulo presenta una visión general del análisis de datos para datos en salud. Brindamos una breve introducción de algunos de los métodos más comunes para el análisis de datos en salud, enfocándonos en elegir una metodología apropiada para los diferentes tipos de objetivos de estudio, y en la presentación y la interpretación del análisis de datos en salud. Brindaremos una visión general de 3 métodos de análisis muy importantes: modelos de regresión lineal, regresión logística y riesgo proporcional (regresión de Cox), los que proveen la base para la mayoría de los análisis de datos conducidos en los estudios clínicos.

Objetivos del capítulo

Al finalizar el capítulo usted debería ser capaz de:

1. Entender cómo los diferentes objetivos de estudio influenciarán el tipo de análisis (Sección 16.1)
2. Realizar los tres tipos de análisis de datos que son comunes para datos en salud (Secciones 16.2-16-4)
3. Presentar e interpretar los resultados de estos tipos de análisis (Secciones 16.2-16.4)

4. Entender las limitaciones y supuestos subyacentes a los diferentes tipos de análisis (Secciones 16.2-16-4)
5. Replicar el análisis de un estudio de caso utilizando alguno de los métodos aprendidos en el capítulo (Sección 16.5)

Esquema del capítulo

Este capítulo está compuesto de cinco secciones. En primer lugar, en esta sección trataremos los aspectos relacionados con la identificación de los tipos de datos y los objetivos del estudio. Estos temas nos permitirán elegir un método de análisis apropiado entre la regresión lineal (Sección 16.2) o logística (Sección 16.3) y el análisis de sobrevida (Sección 16.4), que son las tres secciones siguientes. Luego de este paso, utilizaremos lo aprendido en un caso de estudio usando datos reales de la base Medical Information Martfor Intensive Care II (MIMIC-II), discutiremos brevemente la construcción del modelo y finalmente, resumiremos lo aprendido (Sección 16.5).

16.1.2 Identificando los tipos de datos y objetivos del estudio

En esta sección examinaremos cómo los diferentes objetivos de estudio y tipos de datos afectan los enfoques elegidos para el análisis de los datos. Entender la estructura de los datos y el objetivo del estudio es probablemente el aspecto más importante para elegir una técnica apropiada de análisis.

Objetivos del estudio

Identificar el objetivo del estudio es un aspecto extremadamente importante de la planificación del análisis de datos en salud. Un objetivo descripto vaga o pobremente con frecuencia lleva a un análisis ejecutado pobremente. El objetivo del estudio debería identificar con claridad la población en estudio, el resultado de interés, las covariables de interés, los puntos de tiempo relevantes del estudio, y lo que se desea realizar con estos ítems. Invertir tiempo para construir un objetivo de la investigación muy específico y claro habitualmente ahorrará tiempo en el largo plazo.

Un ejemplo de un objetivo de un estudio claramente definido sería:

Estimar la reducción en la mortalidad a 28 días asociada con uso de vasopresores durante los primeros tres días desde la admisión en la Unidad de Cuidados Intensivos (UCI) Clínica en MIMIC II.

Un ejemplo de un objetivo de un estudio vago y difícil de ejecutar sería:
Predecir la mortalidad en pacientes en UCI.

Mientras ambos pueden estar tratando de alcanzar el mismo objetivo, el primero da un camino mucho más claro para que el científico de datos realice el análisis necesario, dado que identifica la población de estudio (aquellos admitidos a UCI Clínica en MIMIC II), el resultado (mortalidad a los 28 días), covariable de interés (uso de vasopresores en los primeros tres días desde la admisión a UCI Clínica), puntos de tiempo relevantes (28 días para el resultado, dentro de los primeros tres días para la covariable). El objetivo no necesita ser demasiado complicado, habitualmente es conveniente especificar objetivos primarios y secundarios, antes que un único objetivo demasiado complejo.

Tipos de datos

Después de especificar un objetivo claro del estudio, el paso siguiente es determinar los tipos de datos con los que uno está trabajando. La primera distinción que debemos realizar es entre resultados y covariables. Los resultados son lo que el estudio apunta a investigar, mejorar o afectar. En el ejemplo anterior con un objetivo claramente definido, nuestro resultado es la mortalidad a los 28 días. Los resultados también son llamados a veces respuestas o variables dependientes. Las covariables son las variables que uno le gustaría estudiar por su efecto sobre el resultado, o porque cree que pueden tener algún efecto que altere el resultado y uno quisiera controlar. Las covariables también se denominan con diferentes nombres, incluyendo: características, predictores, variables independientes y variables explicativas. En el objetivo de nuestro ejemplo, la covariable primaria de interés es el uso de vasopresores, pero otras covariables también pueden ser importantes en afectar la mortalidad a 28 días, incluyendo edad, género y otras.

Una vez que uno ha identificado los resultados y covariables del estudio, determinar los tipos de datos de los resultados será habitualmente crítico para elegir una técnica de análisis apropiada. Los tipos de datos

generalmente pueden ser identificados como continuos o discretos. Las variables continuas son aquellas plausibles de tomar cualquier valor numérico (número real), a pesar de que este requerimiento frecuentemente no se cumple en forma explícita. Esto contrasta con los datos discretos, que usualmente abarcan solo algunos valores. Por ejemplo, género, puede tomar dos valores: masculino o femenino. Esta es una variable *binaria* dado que abarca dos valores. En el capítulo 11 puede encontrarse mayor análisis sobre tipos de datos.

Hay un tipo especial de datos que pueden ser considerados simultáneamente como continuos y discretos, dado que tienen ambos componentes. Esto sucede frecuentemente con datos como "tiempo al evento" para resultados como mortalidad, donde tanto la ocurrencia de la muerte como el tiempo de sobrevida son de interés. En este caso, el componente discreto es como si el evento (por ejemplo, muerte) ocurriera durante el periodo de observación, y el continuo es el tiempo en el cual la muerte ocurre. El tiempo en el cual la muerte ocurre no está siempre disponible: en este caso, se utiliza el tiempo de la última observación, y los datos son parcialmente *censurados. D*iscutiremos la censura de datos con más detalle en la Sección 16.4.

La figura 16.1 describe el proceso típico por el cual se pueden identificar resultados de covariables, y determinar qué tipo de dato es nuestro resultado. Para cada uno de los tipos de resultados que destacamos – continuo, binario y de sobrevida, hay un set de métodos de análisis más comunes para el uso en datos en salud– que son la regresión lineal, regresión logística, y modelos proporcionales de riesgo de Cox, respectivamente.

Otras consideraciones importantes

Hasta aquí hemos delineado en forma básica cómo elegir un método de análisis de acuerdo al objetivo de estudio. Hay que tener en cuenta que esta discusión ha sido bastante breve y, a pesar de que cubre algunos de los métodos más frecuentemente usados para analizar datos en salud, definitivamente no es exhaustiva. Hay muchas situaciones donde este marco de trabajo y discusión no será adecuado y serán necesarios otros métodos de análisis. Particularmente, resaltamos las siguientes situaciones:

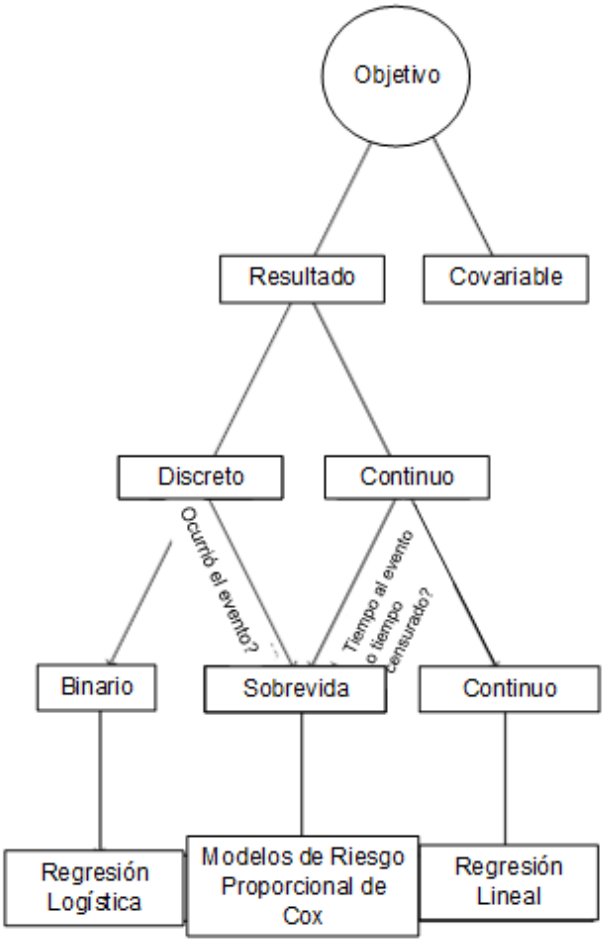

Fig. 16.1 Diagrama de flujo de un proceso simplificado para la elección del método de análisis basado en el objetivo del estudio y los tipos de datos del resultado.

1. Cuando los datos no son a nivel del paciente, como datos agregados (totales) en lugar de datos a nivel individual
2. Cuando los pacientes contribuyen al set de datos con más de una observación (por ejemplo, resultado).

En estos casos, deberían ser usadas otras técnicas.

16.1.3 Datos del caso de estudio

Usaremos un caso de estudio [1] para explorar los enfoques del análisis de datos en salud. Los datos del caso de estudio se originan de un estudio que examina el efecto de los catéteres arteriales invasivos (CAI) en la mortalidad a 28 días en la UCI en pacientes que requirieron ventilación mecánica

durante el primer día de admisión en la UCI. Los datos provienen de MIMIC-II v2.6. En este punto Ud. está listo para hacer el análisis de datos (la extracción de datos y la limpieza ya fueron completadas). Usaremos un archivo separado por comas (.csv) generado después de este proceso, que se puede cargar directamente desde PhysioNet [2, 3]:

```
url <- "http://physionet.org/physiobank/database/mimic2-iaccd/full_cohort_data.csv";
dat <- read.csv(url)
# Or download the csv file from:
# http://physionet.org/physiobank/database/mimic2-iaccd/full_cohort_data.csv
# Type: dat <- read.csv(file.choose())
# And navigate to the file you downloaded (likely in your download directory)
```

Se puede tener acceso al encabezado de este archivo con los nombres de variables usando la función *names* en R.

```
names(dat)
```

```
##  [1] "aline_flg"          "icu_los_day"        "hospital_los_day"
##  [4] "age"                "gender_num"         "weight_first"
##  [7] "bmi"                "sapsi_first"        "sofa_first"
## [10] "service_unit"       "service_num"        "day_icu_intime"
## [13] "day_icu_intime_num" "hour_icu_intime"    "hosp_exp_flg"
## [16] "icu_exp_flg"        "day_28_flg"         "mort_day_censored"
## [19] "censor_flg"         "sepsis_flg"         "chf_flg"
## [22] "afib_flg"           "renal_flg"          "liver_flg"
## [25] "copd_flg"           "cad_flg"            "stroke_flg"
## [28] "mal_flg"            "resp_flg"           "map_1st"
## [31] "hr_1st"             "temp_1st"           "spo2_1st"
## [34] "abg_count"          "wbc_first"          "hgb_first"
## [37] "platelet_first"     "sodium_first"       "potassium_first"
## [40] "tco2_first"         "chloride_first"     "bun_first"
## [43] "creatinine_first"   "po2_first"          "pco2_first"
## [46] "iv_day_1"
```

Hay 46 variables listadas. El foco primario del estudio fue el efecto que tiene la colocación de CAI (aline_flg) en la mortalidad a los 28 días (day_28_flg). Después de haber tratado los conceptos básicos, identificaremos un objetivo de investigación y una técnica de análisis apropiada, y ejecutaremos un análisis abreviado para ilustrar como usar estas técnicas para abordar preguntas científicas reales. Antes de hacer esto necesitamos tratar las técnicas básicas, e introduciremos tres métodos de análisis de datos poderosos frecuentemente usados en el análisis de datos en salud. Usaremos ejemplos del set de datos del caso de estudio para introducir estos conceptos, y volveremos a la pregunta del efecto que tiene el CAI en la mortalidad hacia el final de este capítulo.

16.2 Regresión Lineal

16.2.1 Objetivos de la sección

En esta sección, el lector aprenderá los fundamentos de la regresión lineal, y cómo presentar e interpretar este análisis.

16.2.2 Introducción

La regresión lineal provee las bases de muchos tipos de análisis que hacemos en los datos en salud. En el escenario más simple, tratamos de relacionar un resultado continuo, *y*, a una sola covariable continua, *x*, tratando de encontrar valores para β_0 y β_1 de forma tal que la siguiente ecuación:

$$y = \beta_0 + \beta_1 * x$$

ajuste los datos de manera "optima"[1]. Llamamos estos valores óptimos: $\hat{\beta}_0$y $\hat{\beta}_1$para distinguirlos de los verdaderos valoresβ_0y $\beta_1$los cuales son usualmente desconocidos. En la figura 16.2, vemos un diagrama de dispersión de niveles de TCO2 (y: resultados) versus niveles de PCO2 (x: covariables).

Podemos ver claramente que mientras los niveles de PCO2 aumentan, los niveles de TCO2 también aumentan. Esto sugeriría que podríamos aplicar un modelo de regresión lineal que predice TCO2 a partir de PCO2.

Siempre es una buena idea visualizar los datos cuando es posible, ya que permite evaluar si el análisis subsecuente corresponde a lo que uno puede ver con sus ojos. En este caso, se puede realizar un diagrama de dispersión usando la función *plot:*

```
plot(dat$pco2_first,dat$tco2_first,xlab="PCO2",ylab="TCO2",pch=19,xlim=c(0,175))
```

el cual produce los puntos dispersos en la Figura 16.2.

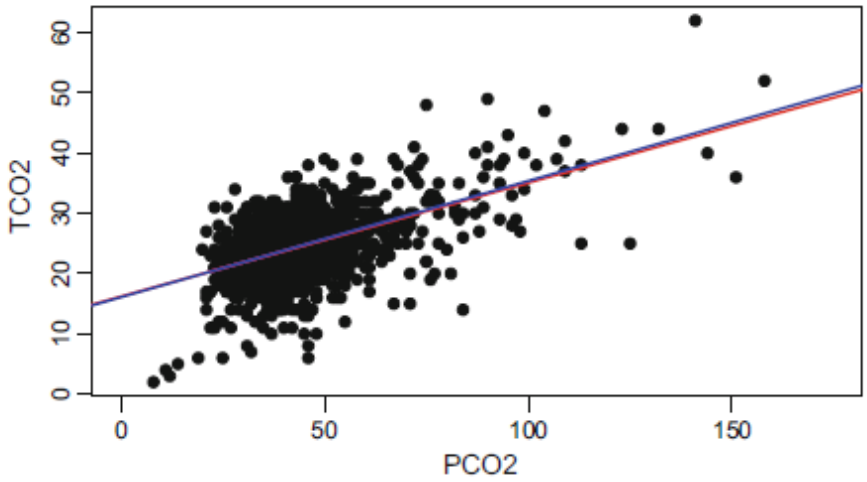

Fig. 16.2 Diagrama de dispersión de PCO2 (eje-x) y TCO2 (eje-y) con la regresión lineal estimada a partir del modelo cuadrático (co2. quad. lm) y el modelo solo lineal (co2. lm)

Es relativamente sencillo encontrar en R la mejor línea de ajuste en el diagrama de dispersión en la Figura 16.2:

```
co2.lm <- lm(tco2_first ~ pco2_first,data=dat)
```

Analizando este comando de izquierda a derecha. La parte *co2. lm<–* asigna la parte derecha del comando a una nueva variable u objeto llamado *co2.lm* la cual contiene información relevante para nuestro modelo de regresión lineal. El lado derecho de este comando ejecuta la función *lm* en *R.*

Lm es una función poderosa en *R* que se ajusta a los modelos lineales. Al igual que en cualquier comando de R, se puede encontrar información de ayuda adicional ejecutando el comando *?lm*. El comando básico lm tiene dos partes. La primera es la fórmula que tiene la sintaxis general *resultadosc̃ovariables.* Aquí, nuestra variable de resultado se llama *tco2_first*. El segundo argumento es separado por una coma y especifica el *dataframe* para usar. En nuestro caso, el *dataframe* es llamado *dat,* así que pasamos *data=dat,* teniendo en cuenta que ambos *tcot2_first* y *pco2_first* son columnas en el *dataframe dat.* El procedimiento general de especificar una fórmula del modelo (*tcot2_firstp̃co2_first*), un *dataframe (data=dat)* y traducirla en una función de R apropiada (*lm*) va a ser usado a lo largo de este capítulo, y es la base para muchos tipos de modelado estadístico en R.

Nos gustaría ver alguna información sobre el modelo que ajustamos, y habitualmente una buena forma de hacerlo es ejecutar el comando *summary* al objeto que creamos:

```
summary(co2.lm)
```

```
##
## Call:
## lm(formula = tco2_first ~ pco2_first, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.8852  -2.5080   0.1891   2.8077  19.2005
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.210859   0.359676   45.07   <2e-16 ***
## pco2_first   0.188572   0.007886   23.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.395 on 1588 degrees of freedom
##   (186 observations deleted due to missingness)
## Multiple R-squared:  0.2647, Adjusted R-squared:  0.2643
## F-statistic: 571.8 on 1 and 1588 DF,  p-value: < 2.2e-16
```

Esto genera información sobre el objeto *lm* que creamos en el paso previo.

La primera parte recuerda el modelo que ajustamos, lo cual es útil cuando ajustamos muchos modelos y estamos intentando compararlos. La segunda parte enumera un resumen de información sobre lo que son llamados *residuos*-un tema importante para validar los supuestos del modelo cubiertos en [8]. Luego enumera los coeficientes estimados– estos son el $\hat{\beta}_0$, (*intercept*), y $\hat{\beta}_1$, *PCO2-first,* los parámetros en la mejor línea de ajuste que estamos tratando de estimar. Este *output* nos dice que la ecuación que mejor se ajusta a los datos es:

$$\texttt{tco2\_first} = 16.21 + 0.189 \times \texttt{pco2\_first}.$$

Estas dos cantidades tienen interpretaciones importantes. El intercepto estimado ($\hat{\beta}_0$) nos dice qué nivel de TCO2 predeciríamos para un individuo con nivel de PCO2 de 0. Esta es la interpretación matemática, y habitualmente esta cantidad tiene un uso práctico limitado. La pendiente estimada ($\hat{\beta}_1$) por otro lado puede ser interpretada como cuán rápido el valor predicho de TCO2 aumenta con cada unidad que aumenta PCO2. En este caso, estimamos que la TCO2 aumenta alrededor de 0.189 mmol/L por cada

mmHg que aumenta en PCO2. Cada coeficiente estimado tiene un *Std. Error* correspondiente (error estándar). Esta es una medida de cuan certera es la estimación. Si el error estándar es relativamente amplio en relación al coeficiente entonces estaremos menos seguros de nuestra estimación. Muchas cosas pueden afectar el error estándar, incluyendo el tamaño de la muestra del estudio. La siguiente columna en esta tabla es *t value*, el cual es simplemente el coeficiente estimado dividido por el error estándar. Esto es seguido por `Pr (>|t|)` que también es conocido como el valor *p.* Las últimas dos cifras son relevantes para un área de la estadística llamada prueba de hipótesis la cual desarrollaremos brevemente a continuación.

Pruebas de hipótesis

La prueba de hipótesis en estadística consiste fundamentalmente en evaluar dos hipótesis competitivas. Una hipótesis llamada *hipótesis nula* es establecida como un "hombre de paja" (un argumento falso que se establece para ser refutado), y es la hipótesis sobre la que uno desearía proveer evidencia en contra. En los métodos de análisis que discutiremos en este capítulo, esto es casi siempre $\beta_k=0$, y usualmente se escribe como H_0: $\beta_k=0$. La hipótesis alternativa (secundaria) comúnmente se asume como $\beta_k\neq0$, y usualmente se escribe H_A: $\beta_k\neq0$. Antes de realizar el análisis debe establecerse un nivel estadísticamente significativo, α. Este valor es conocido como error de Tipo I, y es la probabilidad de rechazar la hipótesis nula cuando la hipótesis nula es verdadera, es decir, de concluir incorrectamente que la hipótesis nula es falsa. En nuestro caso, es la probabilidad de concluir falsamente que el coeficiente es distinto de cero, cuando el coeficiente en realidad es cero. Comúnmente se establece el error de Tipo I en 0.05.

Luego de especificar la hipótesis nula y la alternativa, junto con el nivel de significación, la hipótesis puede ser testeada calculando un valor *p.* El cálculo de un valor *p* se encuentra más allá del alcance de este capítulo, pero explicaremos su interpretación y proporcionaremos algo de información. Los valores *P* son la probabilidad de que los datos sean extremos o más extremos de lo observado, asumiendo que la hipótesis nula es *verdadera*. La hipótesis nula es $\beta_k=0$, entonces, ¿Cuándo podría ser esto improbable? Posiblemente sea poco probable que βk sea igual a 0 cuando estimamos que β_k será bastante mayor. De cualquier manera, ¿qué tan mayor es lo suficiente para rechazar la hipótesis nula? Esto podría depender de cuán seguros

estamos de la estimación de β_k. Si estamos lo bastante seguros, $\hat{\beta}_k$ probablemente no debería ser muy mayor, pero si no lo estamos, no pensaríamos que pudiera ser poco probable incluso para valores muy grandes de $\hat{\beta}_k$. Un valor *p* balancea ambos aspectos, y calcula un solo número. Rechazamos la hipótesis nula cuando el valor *p* es más pequeño que el valor de significación, α.

Volviendo a nuestro modelo de ajuste, vemos que el valor *p* para ambos coeficientes es muy pequeño (<2e-16), y podríamos rechazar ambas hipótesis nulas, concluyendo que probablemente ninguno de los coeficientes es cero. ¿Qué significan estas dos hipótesis en un nivel práctico? Si el intercepto fuese cero, β_0=0 implicaría que la mejor línea de ajuste va desde el origen [(x, y) punto (0,0)], y refutaríamos esta hipótesis. Si la pendiente fuese cero significaría que la mejor línea de ajuste podría ser una línea horizontal plana, y que no aumenta mientras la PCO2 aumenta. Claramente hay una relación entre TCO2 y PCO2, así que también refutaríamos esta hipótesis. En resumen, concluiríamos que necesitamos tanto un intercepto como una pendiente en el modelo. Una siguiente pregunta obvia sería, ¿Podría ser la relación más complicada que una línea recta? Examinaremos esto a continuación.

16.2.3 Selección del modelo

La selección del modelo son técnicas relacionadas con seleccionar el mejor modelo de una lista (quizás una lista bastante grande) de modelos candidatos. Cubriremos algunos conceptos básicos aquí, dado que se tratarán otras técnicas más complejas en un capítulo más adelante. En el caso más simple, tenemos dos modelos, y queremos saber cuál deberíamos usar.

Comenzaremos examinando si la relación entre TCO2 y PCO2 es más compleja que el modelo que ajustamos en la sección previa. Si usted recuerda, ajustamos un modelo donde consideramos un término lineal *pco2_first: tco2-first= β_0 + β_1 x pco2_first.*

Uno podría preguntarse si incluir un término cuadrático ajustaría mejor los datos, por ejemplo si:

$$\texttt{tco2\_first} = \beta_0 + \beta_1 \times \texttt{pco2\_first} + \beta_2 \times \texttt{pco2\_first}^2,$$

es un mejor modelo. Una forma de evaluar esto es testeando la hipótesis nula: β_2=0. Hacemos esto ajustando el modelo mencionado arriba, y mirando el resultado. Sumar un término cuadrático (o cualquier otra función) es bastante fácil usando la función *lm.* Es mejor incluir cualquiera de estas funciones en la función `I` () para estar seguros que son evaluadas como uno espera. La función `I` () fuerza a la fórmula a evaluar lo que se pasó, como el operador ^ tiene un uso diferente en fórmulas en R (ver *`?fórmula`* para más detalles). Ajustando este modelo, y corriendo la función *`summary`* para el modelo:

```
co2.quad.lm <- lm(tco2_first ~ pco2_first + I(pco2_first^2),data=dat)
summary(co2.quad.lm)$coef
```

```
##                     Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)     16.0916260327 0.7713394026 20.8619266 1.309513e-85
## pco2_first       0.1930281243 0.0266927962  7.2314689 7.401248e-13
## I(pco2_first^2) -0.0000356873 0.0002042135 -0.1747548 8.612946e-01
```

Usted notará que hemos abreviado el producto de la función *summary* agregando *$coef* a la función *summary*: esto dice a *R* que queremos información sobre el coeficiente solamente. Mirando primero a los estimadores, vemos que la mejor línea de ajuste es estimada como:

$$\texttt{tco2\_first} = 160.09 + 0.19 \times \texttt{pco2\_first} + 0.00004 \times \texttt{pco2\_first}^2.$$

Podemos agregar ambas mejores líneas de ajuste a la figura 16.2 usando la función *`abline`*:

```
abline(co2.lm,col='red')
abline(co2.quad.lm,col='blue')
```

y uno puede ver que el rojo (término lineal sólo) y el azul (términos lineal y cuadrático) se ajustan de manera casi idéntica. Esto se corresponde con un coeficiente relativamente pequeño estimado para el término `I` *(`pco2_first^2`).* El valor p para este coeficiente es alrededor 0.86, y con un nivel de significación de 0.05 podríamos concluir que no es necesario un término cuadrático en nuestro modelo para ajustar los datos, dado que el modelo con el término lineal ajusta casi tan bien los datos.

Interacciones Estadísticas y Evaluación de Modelos Anidados

Concluimos que un modelo lineal (línea recta) ajusta los datos bastante bien, pero hasta ahora hemos restringido nuestra exploración a solo una

variable por vez. Cuando incluimos otras variables, podemos preguntarnos si la misma línea recta es verdadera para todos los pacientes. Por ejemplo, ¿la relación entre PCO2 y TCO2 podría ser diferente entre hombres y mujeres? Podríamos subagrupar los datos en un *dataframe* para hombres y en un *dataframe* para mujeres, y luego ajustar regresiones separadas para cada género. Una forma más eficiente de cumplir esto es considerar ambos géneros en un solo modelo, e incluir el género como una covariable. Por ejemplo, podemos ajustar:

$$\texttt{tco2\_first} = \beta_0 + \beta_1 \times \texttt{pco2\_first} + \beta_2 \times \texttt{gender\_num}.$$

La variable *gender_num* adquiere un valor de 0 para mujeres y 1 para hombres, y para los hombres el modelo es:

$$\texttt{tco2\_first} = \underbrace{(\beta_0 + \beta_2)}_{\textit{intercepto}} + \beta_1 \times \texttt{pco2\_first},$$

Y en mujeres:

$$\texttt{tco2\_first} = \beta_0 + \beta_1 \times \texttt{pco2\_first}.$$

Como puede verse estos modelos tienen la misma pendiente, pero distintos interceptos (la distancia entre las pendientes es β_2). En otras palabras, las líneas que se ajustan para hombres y mujeres serán paralelas y separadas por una distancia de β_2 para todos los valores de esa función *pco2_first*. Esto no es exactamente lo que querríamos, dado que las pendientes también podrían ser diferentes. Para permitir esto, necesitamos discutir la idea de una interacción entre dos variables. Una interacción es esencialmente el producto de dos covariables. En este caso, lo que llamaremos modelo de interacción, estaríamos ajustando:

$$\texttt{tco2\_first} = \beta_0 + \beta_1 \times \texttt{pco2\_first} + \beta_2 \times \texttt{gender\_num} + \beta_3 \times \underbrace{\texttt{gender\_num} \times \texttt{pco2\_first}}_{\textit{Término de interacción}}.$$

De nuevo, separando los casos para hombres:

$$\texttt{tco2\_first} = \underbrace{(\beta_0 + \beta_2)}_{\textit{Intercepto}} + \underbrace{(\beta_1 + \beta_3)}_{\textit{pendiente}} \times \texttt{pco2\_first},$$

Y mujeres:

$$\texttt{tco2\_first} = \underbrace{(\beta_0)}_{\textit{Intercepto}} + \underbrace{(\beta_1)}_{\textit{pendiente}} \times \texttt{pco2\_first}.$$

Ahora hombres y mujeres tienen diferentes interceptos *y* pendientes.

Ajustar estos modelos en R es relativamente sencillo. Aunque no es absolutamente requerido en esta circunstancia particular, es sabio asegurarse que R maneja los tipos de datos de manera correcta, asegurándonos de que nuestras variables son de la clase correcta. En este caso particular, los hombres están codificados como *1* y las mujeres como *0* (una covariable binaria discreta) pero R piensa que es un dato numérico (continuo):

```
class(dat$gender_num)
```

```
## [1] "integer"
```

Dejar esto inalterado, no afectará el análisis en esta instancia, pero puede ser problemático cuando uno está tratando con otros tipos de datos como datos categóricos con muchas categorías (por ejemplo, etnia). Además, configurando los datos al tipo correcto, el producto que genera R puede ser más informativo. Podemos configurar la variable *`gender_num`* a la clase *`factor`* usando la función *`as. factor`*.

```
dat$gender_num <- as.factor(dat$gender_num)
```

Aquí sobre escribimos la variable antigua en la base de datos *dat* con una nueva copia que es de clase

```
factor:
class(dat$gender_num)
```

```
## [1] "factor"
```

Ahora que tenemos la variable género correctamente codificada, vamos a ajustar los modelos que discutimos previamente. Primero el modelo con género como una covariable, pero sin interacción. Podemos hacer esto simplemente sumando la variable *gender_num* a la fórmula anterior para ajustarla a nuestro modelo *co2. lm.*

```
co2.gender.lm <- lm(tco2_first ~ pco2_first + gender_num,data=dat)
summary(co2.gender.lm)$coef
```

```
##                Estimate  Std. Error    t value      Pr(>|t|)
## (Intercept) 16.3043942 0.377712532 43.1661457 6.337240e-270
## pco2_first   0.1888542 0.007894741 23.9215128 3.015777e-108
## gender_num1 -0.1816540 0.223738366 -0.8119036  4.169687e-01
```

Este resultado es muy similar al que teníamos antes, pero ahora también hay un término *gender_num*. El *1* está presente en la primera columna después de *gender_num*, y nos dice para quién este coeficiente es relevante (sujetos con 1 para los hombres *gender_num*). Esto es siempre relativo para el grupo de base, y en este caso son las mujeres.

La estimación es negativa, entendiendo que la línea para hombres estará por debajo de la línea para las mujeres. Graficando esta curva ajustada en Fig. 16.3:

```
plot(dat$pco2_first, dat$tco2_first, col = dat$gender_num, xlab = "PCO2", ylab = "TCO2",
    xlim = c(0, 40), type = "n", ylim = c(15, 25))
abline(a = c(coef(co2.gender.lm)[1]), b = coef(co2.gender.lm)[2])
abline(a = coef(co2.gender.lm)[1] + coef(co2.gender.lm)[3], b = coef(co2.gender.lm)[2],
    col = "red")
```

Veremos que las líneas son paralelas, pero casi indistinguibles. De hecho, este gráfico fue recortado para observar cualquier diferencia. Desde el estimador del resultado *summary* más arriba, la diferencia entre las dos líneas es -0.182mmol/L, la cual es bastante pequeña, así que quizás este no es un valor tan sorprendente. También podemos ver en el resultado *summary* más arriba que el valor p es alrededor de 0.42, y *no* rechazaríamos la hipótesis nula de que el verdadero valor del coeficiente *gender_num* es cero.

Y ahora avanzaremos en el modelo con una interacción entre *pco2_first* y *gender_num*. *P*ara sumar una interacción entre dos variables use el operador * dentro de la fórmula modelo. Por defecto, R también agregará todos los efectos principales (variables contenidas en la interacción) al modelo, por lo que simplemente agregando

*pco2_first*gender_num*, *R* sumará *pco2_first* y *gender_num* y la interacción entre ellos al modelo de ajuste.

```
co2.gender.interaction.lm <- lm(tco2_first ~ pco2_first*gender_num,data=dat)
summary(co2.gender.interaction.lm)$coef
```

```
##                            Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)            15.85443226 0.48869107 32.442648 1.591490e-177
## pco2_first              0.19939518 0.01072876 18.585105  6.559901e-70
## gender_num1             0.81437833 0.72225677  1.127547  2.596819e-01
## pco2_first:gender_num1 -0.02297002 0.01583758 -1.450348  1.471591e-01
```

Los coeficientes estimados son $\hat{\beta}_0$, $\hat{\beta}_1$,$\hat{\beta}_2$ y $\hat{\beta}_3$ respectivamente, y podemos determinar las mejores líneas de ajuste para hombres:

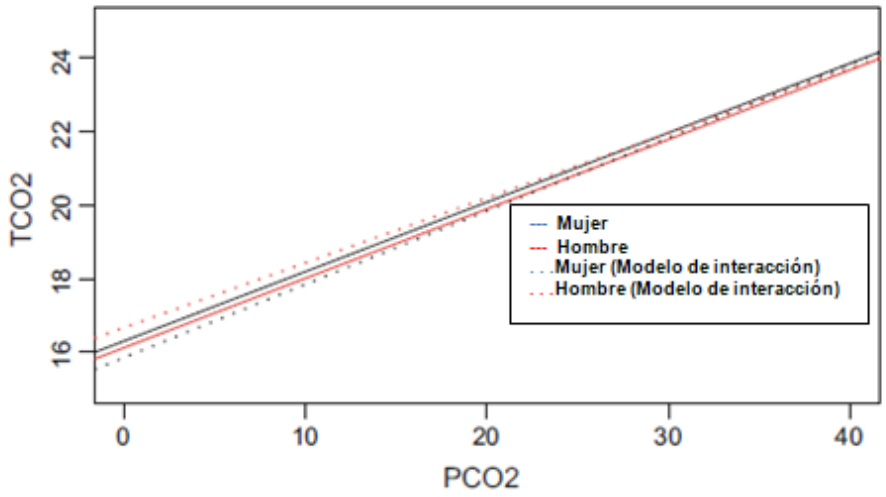

Fig. 16.3 Regresión ajustada de PCO2 en TCO2 con género (mujer negro; hombre rojo; sólido sin interacción; punteada con interacción). *Nota* Ambos ejes están cortados con fines ilustrativos.

$$\begin{aligned}\texttt{tco2\_first} &= (15.85 + 0.81) + (0.20 - 0.023) \times \texttt{pco2\_first} \\ &= 16.67 + 0.18 \times \texttt{pco2\_first},\end{aligned}$$

Y para mujeres:

$$\texttt{tco2\_first} = 15.85 + 0.20 \times \texttt{pco2\_first}.$$

Basado en esto, los interceptos de los hombres deberían ser más altos, pero sus pendientes no deberían ser tan empinadas, en relación a las de las mujeres. Controlemos esto y sumemos los ajustes del nuevo modelo como líneas punteadas y agreguemos una leyenda a la Fig. 16.3.

```
abline(a = coef(co2.gender.interaction.lm)[1], b = coef(co2.gender.interaction.lm)[2],
    lty = 3, lwd = 2)
abline(a = coef(co2.gender.interaction.lm)[1] + coef(co2.gender.interaction.lm)[3],
    b = coef(co2.gender.interaction.lm)[2] + coef(co2.gender.interaction.lm)[4],
    col = "red", lty = 3, lwd = 2)
legend(24, 20, lty = c(1, 1, 3, 3), lwd = c(1, 1, 2, 2), col = c("black", "red",
    "black", "red"), c("Female", "Male", "Female (Interaction Model)", "Male (Interaction Model)"))
```

Podemos ver que los ajustes generados desde este gráfico son un poco diferentes a los generados para un modelo sin interacción. La diferencia más grande es que las líneas punteadas ya no son paralelas. Esto tiene algunas implicancias serias, particularmente cuando se trata de interpretar nuestro resultado. Primero note que el coeficiente estimado para la variable *gender_num* ahora es positivo. Esto significa que en *pco2_first=0,* los hombres (rojo) tienen niveles más altos de *tco2_first* que las mujeres (negro). Si recuerda el ajuste del modelo anterior, las mujeres tenían niveles más altos de *tco2_first* en todos los niveles de *pco2_first*. En algún punto alrededor de *pco2_first=35* esto cambia y las mujeres (negro) tienen niveles de *tco2_first* más altos que los hombres (rojo). Esto significa que el efecto de *gender_num* *puede* variar cuando se cambia el nivel de *pco2_first*, y es el por qué se refieren a las interacciones como modificadores de efecto en la literatura epidemiológica. El efecto no necesita que cambie los signos (por ejemplo, las líneas no necesitan cruzarse) sobre el rango observado de valores para que una interacción esté presente.

La pregunta sigue siendo, ¿Es la variable *gender_num* importante? Observamos esto brevemente cuando examinamos la columna *t value* en el modelo de no interacción que incluía *gender-num*. Qué pasaría si quisiéramos evaluar (simultáneamente) la hipótesis nula: β_2 *y* β_3 *=0*. Hay una prueba útil conocida como F-test que puede ayudarnos en este exacto escenario donde queremos observar si deberíamos usar un modelo más grande (más covariables) o usar un modelo más pequeño (menos covariables). La prueba F-test aplica solo a *modelos anidados* –el modelo más grande *debe* contener cada covariable que es usada en el modelo más pequeño, y el modelo más pequeño *no puede* contener covariables que no están en el modelo más grande. El modelo de interacción y el modelo con género son modelos anidados dado que todas las variables en el modelo con género también están en el modelo más grande de interacción. Un ejemplo de un modelo no anidado sería el modelo cuadrático y el de interacción: el modelo más pequeño (cuadrático) tiene un término (*pco2_first²*) que

no está en el modelo más grande (de interacción). Un F-test no sería apropiado para este último caso.

Para realizar un F-test, primero hay que ajustar los dos modelos que uno desea considerar, y después ejecutar el comando *anova* corriendo los dos modelos.

```
anova(co2.lm,co2.gender.interaction.lm)
```

```
## Analysis of Variance Table
##
## Model 1: tco2_first ~ pco2_first
## Model 2: tco2_first ~ pco2_first * gender_num
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1   1588 30674
## 2   1586 30621  2    53.349 1.3816 0.2515
```

Como se puede ver, el comando *anova* primero genera una lista de los modelos que está considerando. Gran parte del resto de la información va más allá del foco de este capítulo, pero resaltaremos el valor p reportado por el F-test (*Pr (>F)*), que es en este caso 0.2515. En los modelos anidados, la hipótesis nula es que todos los coeficientes en el modelo más grande y no en el modelo más pequeño son cero. En el caso que estamos evaluando, nuestra hipótesis nula es β_2 y β_3=0. Dado que el valor p excede el nivel de significación típicamente usado (α=0.05), no refutaríamos la hipótesis nula, y podríamos decir que el modelo más pequeño explica los datos igual de bien que el más grande. Si estos fueran los únicos modelos que estamos considerando, usaríamos el modelo más pequeño como nuestro modelo final y reportaríamos el modelo final en nuestros resultados. Ahora discutiremos qué es exactamente lo que uno debería reportar y cómo se pueden interpretar los resultados.

16.2.4 Reportando e Interpretando la Regresión Lineal

Discutiremos brevemente como comunicar un análisis de regresión lineal. En general, antes de presentar los resultados, debería hacerse una mención acerca de cómo se obtuvieron los resultados. Es una buena idea informar: si transformamos el resultado o cualquier covariable en cualquier manera (por ejemplo, considerando el logaritmo), qué covariables consideramos y cómo elegimos las covariables que estaban en el modelo que informamos. En nuestro ejemplo anterior, no transformamos el producto (TCO2), consideramos PCO2 como un término lineal y cuadrático, y consideramos el

género por sí mismo y como término de interacción con PCO2. Primero evaluamos si un término cuadrático debería ser incluido en el modelo usando un T-test, luego de lo cual consideramos un modelo con género y una interacción de género-PCO2, e hicimos una selección del modelo con un F-test. Nuestro modelo final involucra solamente un término de PCO2 lineal y un intercepto.

Cuando se reportan los resultados, una buena idea es informar tres aspectos de cada covariable. Primero, deberíamos siempre informar el coeficiente estimado. El coeficiente estimado permite al lector determinar la magnitud del efecto. Hay muchas circunstancias donde el resultado puede ser estadísticamente significativo, pero prácticamente sin sentido. Segundo, junto con el coeficiente estimado deberíamos reportar siempre la medida de incerteza o precisión. Para la regresión lineal, puede ser reportado el error estándar (la columna *Std.Error* en los resultados de R). Más adelante en otra sección, veremos otro método llamado intervalo de confianza. Por último, es también una buena idea informar un valor-p para cada coeficiente. Un ejemplo de presentación apropiada de nuestro modelo final debería ser algo similar a: la TCO2 aumenta 0.18 unidades por unidad de aumento de PCO2 (SE: 0.008, valor p<0.001). Note que informamos un valor p<0.001, cuando de hecho es más chico que este. Es común reportar valores de p muy pequeños como <0.001 ó <0.0001 en vez de usar un número más grande de cifras decimales. Mientras a veces simplemente se reporta si p<0.05 ó no (es decir, si el resultado es estadísticamente significativo o no), práctica que debería evitarse.

Usualmente, también es una buena idea discutir que tan bien se ajusta el modelo general. Hay muchas formas de lograr esto, con frecuencia se reporta una cantidad sin unidad conocida como R^2 (pronunciada R-al cuadrado). Mirando nuevamente los resultados provistos por R para nuestro modelo elegido final, podemos encontrar el valor de R^2 para este modelo bajo *Múltiple R-squared:* 0.2647. Esta cantidad es una proporción (un numero entre 0 y 1) y describe cuánto de la variabilidad total en los datos es explicado por el modelo. Un R^2 de 1 indica un ajuste perfecto, mientras que 0 explica que no hay variabilidad en los datos. Lo que constituye exactamente un “buen” R^2 depende del tema y de cómo será usado. Otra forma de describir el ajuste en su modelo es a través del error estándar residual. Esto también se encuentra en el resultado de *lm* cuando se usa la

función *summary*. Esto estima aproximadamente la raíz cuadrada del promedio de la distancia al cuadrado entre el modelo de ajuste y los datos. Mientras que usa las mismas unidades que el resultado, en general es más difícil de interpretar que R^2. Debería tenerse en cuenta que para evaluar el error de predicción, estos valores son en general demasiado optimistas cuando se aplican los nuevos datos, y debería evaluarse por otros métodos una mejor estimación del error (por ejemplo, validación cruzada y otros que serán tratados en otro capítulo y en otros lados) [4, 5].

Interpretar los resultados

Interpretar los resultados es un componente importante de cualquier análisis de datos. Ya tratamos la interpretación del intercepto, lo cual es la predicción del resultado cuando todas las covariables se establecen en cero. Esta cantidad no es de interés directo en la mayoría de los estudios. Si uno quiere interpretarla, sustraer el promedio de cada una de las covariables del modelo la hará más interpretable– el valor esperado del resultado cuando todas las covariables están configuradas a los promedios del estudio.

Los coeficientes estimados para las covariables son en general las cantidades de mayor interés científico. Cuando la covariable es binaria (por ejemplo, *gender_num*), el coeficiente representa la diferencia entre un nivel de la covariable (1) relativo a otro nivel (0), mientras se mantengan otras covariables constantes en el modelo. A pesar de que no lo trataremos hasta la próxima sección, extender covariables discretas al caso cuando tienen más de dos niveles (por ejemplo, etnia o *service_unit*) es bastante similar, con la excepción que es importante referenciar al grupo de base (es decir, a qué es relativo el efecto). Volveremos a este tema más tarde en este capítulo. Por último, cuando la covariable es continua la interpretación es el cambio esperado en el resultado como resultado del incremento de la covariable en cuestión en una unidad, mientras se mantienen las otras covariables fijas. Esta interpretación es universal para cualquier coeficiente no intercepto, incluyendo datos binarios y otros discretos, pero se apoya más en la comprensión de como R codifica estas covariables con variables ficticias, también conocidas como “variables dummy”.

Examinamos brevemente las interacciones estadísticas aunque este tema puede ser muy difícil de interpretar. En general es recomendable, cuando sea

posible, representar la interacción gráficamente, como hicimos en Fig. 16.3.

Intervalos de confianza y predicción

Como mencionamos antes, un método para cuantificar la incerteza alrededor de los coeficientes estimados es reportando el error estándar. Otro método comúnmente usado es reportar un intervalo de confianza, más comúnmente un intervalo de confianza de 95%. Un intervalo de confianza de 95% para β es un intervalo para el cual, si los datos fueran recolectados repetidamente, alrededor del 95% de los *intervalos* contendrían el *valor real* del parámetro, β, asumiendo que los supuestos del modelo son correctos.

Para obtener los intervalos de confianza de 95% de los coeficientes, R tiene una función *`confint`,* que se aplica un objeto *`lm`*. Entonces producirá los límites del intervalo de confianza 2.5 y 97.5% para cada coeficiente.

```
confint(co2.lm)
```

```
##                 2.5 %     97.5 %
## (Intercept) 15.5053693 16.9163494
## pco2_first   0.1731033  0.2040403
```

El intervalo de confianza 95% para *`pco2_first`* es alrededor de 0.17-0.20, que podría ser un poco más informativo que reportar el error estándar. En general las personas mirarán si el intervalo de confianza incluye cero (sin efecto). Como no lo hace, y de hecho, el intervalo es bastante estrecho y no muy cercano a cero, esto provee evidencia adicional de su importancia. Hay una relación bien conocida entre la evaluación de la hipótesis y los intervalos de confianza en la cual no entraremos en detalle aquí.

Cuando graficamos los datos con el modelo de ajuste, similar a la Fig. 16.2, es una buena idea incluir algún tipo de evaluación de la incerteza también. Para hacer esto en R, primero crearemos un *`dataframe`* con los niveles de PCO2 que queremos predecir. En este caso, nos gustaría predecir el resultado (TCO2) dentro del rango de los valores de la covariable observada (PCO2). Hacemos esto creando un *`dataframe`*, en el que los nombres de las variables coincidan con las covariables usadas en el modelo. En nuestro caso, tenemos solo una covariable (*`pco2_first`*), y predecimos el resultado dentro del rango de los valores de la covariable que observamos, determinados por las funciones *`min`* y *`max`*.

```
grid.pred <- data.frame(pco2_first=seq.int(from=min(dat$pco2_first,na.rm=T),
                                           to=max(dat$pco2_first,na.rm=T)));
```

Luego, usando la función *predict,* podemos predecir los niveles de TCO2 para estos valores de PCO2. La función *predict* tiene tres argumentos: el modelo que construimos (en este caso, usando *lm*), *newdata,* e *interval.* El argumento *newdata* permite pasar cualquier *data frame* con las mismas covariables que el modelo ajustado, que es el motivo por el cual creamos *grid.pred* más arriba. Finalmente, el argumento *interval* es opcional, y permite la inclusión de cualquier intervalo de confianza o predicción. Queremos ilustrar un intervalo de predicción que incorpore tanto la incerteza sobre los coeficientes del modelo, como la incerteza generada por el proceso de generación de datos, entonces introduciremos *interval="prediction".*

```
preds <- predict(co2.lm,newdata=grid.pred,interval = "prediction")
preds[1:2,]
```

```
##        fit      lwr      upr
## 1 17.71943 9.078647 26.36022
## 2 17.90801 9.268186 26.54783
```

Imprimimos las primeras dos filas de nuestras predicciones, *preds*, las cuales son las predicciones del modelo para PCO2 en 8 y 9. Podemos ver que nuestras predicciones (*fit*) están aproximadamente separadas por 0.18, lo cual tiene sentido dado nuestra estimación de la pendiente (0.18). También vemos que nuestros intervalos de predicción de 95% son muy amplios, yendo desde alrededor 9 (*lwr*) a 26 (*upr*). Esto indica que, a pesar de haber generado un modelo que es muy estadísticamente significativo, todavía tenemos mucha incerteza sobre las predicciones que se generan desde dicho modelo. Es una buena idea capturar esta cualidad cuando uno grafica qué tan bien nuestro modelo se ajusta, mostrando las líneas de intervalo como líneas punteadas. Vamos a graficar nuestro modelo final ajustado, *co2. lm,* junto con el gráfico de dispersión y el intervalo de predicción en la Fig. 16.4.

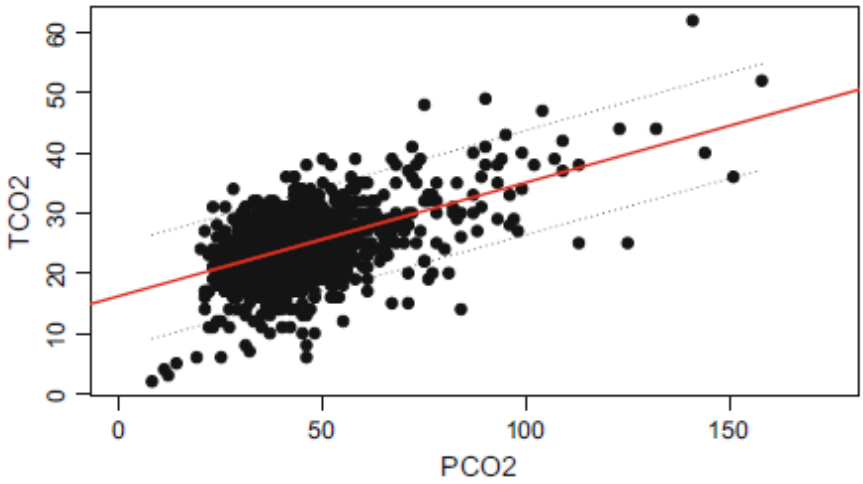

Fig. 16.4 Grafico de dispersión de PCO2 (eje-x) y TCO2 (eje-y) junto con los estimados de regresión lineal del modelo lineal solamente (co2. lm). La línea *punteada* representa el 95% de los intervalos de predicción para el modelo.

```
plot(dat$pco2_first,dat$tco2_first,xlab="PCO2",ylab="TCO2",pch=19,xlim=c(0,175))
co2.lm <- lm(tco2_first ~ pco2_first,data=dat)
abline(co2.lm,col='red',lwd=2)
lines(grid.pred$pco2_first,preds[,2],lty=3)
lines(grid.pred$pco2_first,preds[,3],lty=3)
```

16.2.5 Advertencias y conclusiones

La regresión lineal es una herramienta extremadamente poderosa para hacer análisis de datos sobre resultados continuos. A pesar de esto, hay muchos aspectos a tener en cuenta cuando se realiza este tipo de análisis.

1. La prueba de hipótesis y la generación del intervalo son dependientes de los supuestos del modelo. Hacer gráficos diagnósticos es un componente crítico cuando se lleva a cabo un análisis de datos. Discutimos sobre este tema en otra sección del libro y lo referiremos a [6-8] para más información sobre este importante tema.
2. Los valores *oulier* o atípicos pueden ser problemáticos al ajustar los modelos. Cuando hay valores *"outliers"* en las covariables en general es más fácil transformar una variable numérica en una categórica (dos o más grupos cortan los valores de la covariable). Eliminar los valores atípicos debería evitarse cuando sea posible, dado que en general dan mucha información sobre el proceso de generación de datos. En otros casos, pueden identificar problemas para el proceso de extracción. Por ejemplo, un subconjunto de los datos puede usar diferentes unidades

para la misma covariable (por ejemplo, pulgadas y centímetros para la altura), y entonces los datos necesitan ser convertidos a unidades comunes. R dispone de métodos robustos para tratar valores atípicos y se encuentra disponible una breve introducción de cómo empezar con algunas de las funciones en R [7].

3. Estar atentos a los datos faltantes. R reporta información referente a los datos faltantes en el resultado *summary. P*ara nuestro modelo ajustado *co2.lm,* teníamos 186 observaciones con observaciones faltantes de *pco2-first*. R dejará estas observaciones fuera del análisis, y ajustará las observaciones no faltantes. Siempre chequee el resultado para asegurarse que tiene tantas observaciones como se supone que debería. Cuando muchas observaciones tienen datos faltantes y trata de construir un modelo con un gran número de coeficientes, puede estar ajustando el modelo en solo algunas observaciones disponibles.
4. Valorar potencial multi-colinearidad. La colinearidad puede ocurrir cuando dos o más covariables se encuentran muy correlacionadas. Por ejemplo, si la presión arterial en los brazos izquierdo y derecho son medidas simultáneamente, y ambas son usadas como covariables en el modelo. En este caso, considere tomar la suma, promedio o diferencia (el que sea más útil en el caso particular) para elaborar una única covariable. También puede ocurrir co-linearidad cuando una variable categórica fue generada de una manera no apropiada. Por ejemplo, definir grupos a través de la covariable PCO2 de 0-25, 5-26, 26-50, >50 puede causar que la regresión lineal encuentre algunas dificultades dado que el primer y segundo grupo son casi idénticos (usualmente este tipo de situaciones son errores de programación). Identificar covariables que pueden ser co-lineales es una parte clave del paso de análisis exploratorio, donde se pueden ver usualmente (pero no siempre) graficando los datos.
5. Chequear si los resultados son dependientes. Esto ocurre más comúnmente cuando un paciente contribuye con múltiples observaciones (resultados). Hay métodos alternativos para lidiar con esta situación [9], pero exceden el foco de este capítulo.

Estas preocupaciones no deberían desalentarlo a usar una regresión lineal. Es un método extremadamente poderoso y razonablemente robusto para alguno de los problemas discutidos previamente, dependiendo de la

situación. Frecuentemente un resultado continuo es convertido a uno binario, y en general, no hay una razón convincente para hacerlo. Discretizar el resultado puede generar perdida de información relativa a qué pacientes pueden beneficiarse o pueden ser más dañados por una terapia, dado que un resultado binario puede tratar pacientes con resultados muy diferentes en la escala continua como si fuese el mismo. El marco de referencia general que tomamos en la regresión lineal reflejará el modo en el cual abordamos las otras técnicas de análisis que discutimos más tarde en este capítulo.

16.3 Regresión logística

16.3.1 Objetivos de la sección

En esta sección, el lector aprenderá los fundamentos de la regresión logística y cómo presentar e interpretar dicho análisis.

16.3.2 Introducción

En la Sec. 16.2 cubrimos una metodología muy útil para el modelado cuantitativo o de resultados continuos. Nosotros sabemos, por supuesto, que en salud, los resultados se expresan en todas las diferentes formas de tipos de datos. De hecho, los resultados en salud que en general nos importan-curado/no curado, vivo/muerto, son resultados discretos binarios. Sería ideal si pudiéramos extender el mismo marco general usado en los resultados continuos para los resultados binarios. La regresión logística nos permite incorporar mucho de lo que hemos aprendido en la sección previa y aplicar los mismos principios en los resultados binarios.

Cuando tratamos con datos binarios, quisiéramos poder modelar la probabilidad de un tipo de resultado dada una o más covariables. Uno podría preguntarse, ¿por qué no simplemente usar una regresión lineal? Hay múltiples razones de por qué esto es generalmente una mala idea. Las probabilidades necesitan estar entre cero y uno, y no hay nada en la regresión lineal para limitar las probabilidades estimadas a este intervalo. Esto significaría que uno puede tener una probabilidad estimada 2, o ¡Incluso una probabilidad negativa! Esto es una propiedad poco atractiva de este método (hay otras), y aunque a veces se usa, la disponibilidad de un buen software como R, permite realizar mejores análisis en forma fácil y

eficiente. Antes de introducir este software, deberíamos introducir el análisis de pequeñas tablas de contingencia.

16.3.3 Tablas 2x2

Las tablas de contingencia son la mejor forma de empezar a pensar sobre datos binarios. Una tabla de contingencia tabula en forma cruzada el resultado a través de dos o más niveles de una covariable. Empecemos creando una nueva variable (`age.cat`) que dicotomiza *age* en dos categorías de edad: ≤55 y >55. Tenga en cuenta que dado que tratamos a la edad como variable discreta, también cambiaremos los tipos de datos a factor. Esto es similar a lo que hicimos para la variable gender_num cuando discutimos regresión lineal en la sección previa. Podemos desglosar la nueva variable usando la función `table`.

```
dat$age.cat <- as.factor(ifelse(dat$age<=55, "<=55",">55"))
table(dat$age.cat)
```

```
##
## <=55  >55
##  923  853
```

Queremos ver cómo se distribuye la mortalidad a los 28 días dentro de estas categorías de edad. Podemos hacer esto construyendo una tabla de contingencia, o en este caso, lo que conocemos como tabla 2 x 2.

```
table(dat$age.cat,dat$day_28_flg)
```

```
##
##          0   1
##   <=55 883  40
##   >55  610 243
```

De la tabla de más arriba, se puede ver que 40 pacientes en el grupo joven (≤55) murieron dentro de los 28 días, mientras que 243 en el grupo de mayor edad murieron. Esto corresponde a P (muerte|edad≤55) = 0.043) ó 4.3% y P (muerte|edad>55) =0.284 ó 28.4%, donde "|" puede ser interpretado como "dado" o "para aquellos que tienen". Esta diferencia es bastante marcada, y sabemos que la edad es un factor importante en la mortalidad, así que no nos sorprende.

Las probabilidades de que un evento suceda es un número positivo y puede ser calculado desde la probabilidad de un evento, P, por la siguiente fórmula

$$\text{Odds} = \frac{p}{1-p}.$$

Un evento con Odds de cero no pasa nunca, y un evento con un Odds muy alto probabilidad (>100) es muy probable que pase. Aquí, el odds de morir a los 28 días del grupo joven es 0.043/ (1-0.043) =0.045, y en el grupo más viejo es 0.284/ (1-0.284) =0.40. Es conveniente representar estos dos valores como un ratio o proporción, y la elección de que va en el numerador y en el denominador es algo arbitrario. En este caso, elegiremos poner el odds del grupo más viejo en el numerador y el del más joven en el denominador, y es importante dejar en claro qué grupo está en el numerador y denominador en general. En este caso el odds ratio es 0.40/0.045=8.79, el cual indica una asociación muy fuerte entre la edad y la muerte, y significa que la probabilidad de muerte en el grupo más viejo es casi 9 veces más alta que si lo comparamos con el grupo más joven. Hay un atajo conveniente para hacer cálculos de odds ratio haciendo una X en una tabla 2 x 2 y multiplicando el de arriba a la izquierda por el de abajo a la derecha, y luego dividiendo por el producto del de abajo a la izquierda y el de arriba a la derecha. En este caso, 883*243/610*40=8.79.

Ahora observemos un caso algo diferente– cuando la covariable contiene más de dos valores. Dicha variable es *service_unit*. Observemos cómo las muertes están distribuidas dentro de las diferentes unidades:

```
deathbyservice <- table(dat$service_unit,dat$day_28_flg)
deathbyservice
```

```
##
##          0   1
##   FICU  59   3
##   MICU 605 127
##   SICU 829 153
```

Podemos obtener la frecuencia de estas unidades aplicando la función *prop.table* a nuestra tabla de doble entrada.

```
dbys.proptable <- prop.table(deathbyservice,1)
dbys.proptable
```

```
##
##                0         1
##   FICU 0.9516129 0.0483871
##   MICU 0.8265027 0.1734973
##   SICU 0.8441955 0.1558045
```

Parecería que la FICU podría tener una tasa de muerte más baja que la *MICU o la SICU.* Para calcular un odds ratio, primero hay que calcular los Odds:

```
dbys.proptable[,"1"]/dbys.proptable[,"0"]
```

```
##       FICU       MICU       SICU
## 0.05084746 0.20991736 0.18455971
```

Y después necesitamos elegir cuál de las unidades FICU, MICU o SICU servirá como referencia o grupo de base. Este es el grupo contra el que los otros dos grupos serán comparados. Otra vez la elección es arbitraria pero debería ser dictada por el objetivo del estudio. Si esto fuera un ensayo clínico con dos ramas de tratamiento y una rama placebo, sería tonto usar uno de los tratamientos como grupo de referencia, particularmente si uno quisiera comparar la eficacia de los tratamientos. En este caso particular, no hay un grupo claro de referencia, pero dado que la mortalidad en FICU es tanto más baja que en las otras dos, la usaremos como el grupo de referencia. Calculando el odds ratio para MICU y SICU obtenemos 4.13 y 3.63, respectivamente. Estas son también asociaciones muy fuertes, es decir que la probabilidad de morir en SICU y MICU es relativamente similar pero alrededor de 4 veces más alta que en FICU.

Las tablas de contingencia y las tablas 2 x 2 son en particular las bases del trabajo con datos binarios, y en general es una buena manera de comenzar a mirar los datos.

16.3.4 Introduciendo a la Regresión Logística

Aunque las tablas de contingencia son un modo fundamental de mirar los datos binarios, son algo limitadas. ¿Qué sucede cuando la covariable de interés es continua? Podríamos por supuesto crear categorías desde la covariable estableciendo puntos de corte, pero podríamos todavía perder

algún aspecto importante de la relación entre la covariable y el resultado no eligiendo los puntos de corte correctos. También, qué sucede cuando sabemos que una covariable está relacionada tanto con el resultado como con la covariable de interés. Este tipo de variable es llamado un confundidor y ocurre con frecuencia en datos observacionales, y a pesar de que en las tablas de contingencia hay formas de tener en cuenta los confundidores, esto se vuelve más difícil cuando hay más de uno presente. La regresión logística es una forma de abordar ambos temas y también muchos otros. Si usted recuerda, usar regresión lineal es problemático dado que es propensa a estimar probabilidades fuera del rango [0,1]. La regresión logística no tiene ese problema per se porque usa una función de enlace conocida como la función `logit`, la cual mapea probabilidades en el intervalo [0,1] a un número real (-∞, ∞). Esto es importante por muchas razones prácticas y técnicas. El logit de Px (es decir la probabilidad de un evento para determinados valores de la covariable *x*) está relacionado a las covariables de la siguiente manera:

$$\text{logit}(p_x) = \log(Odds_x) = \log(\frac{p_x}{1-p_x}) = \beta_0 + \beta_1 \times x.$$

Vale la pena señalar aquí que log en este caso, y en la mayor parte de las veces en estadística se refiere al logaritmo natural, algunas veces nombrado como `ln`.

La primer covariable que estábamos considerando, `age.cat` era también una variable binaria, que toma valores 1 cuando `age >55` y 0 cuando `age ≤55`. Así que introduciendo estos valores, primero para el grupo joven (*x=0*):

$$\text{logit}(p_{x=0}) = \log(Odds_{x=0}) = \log(\frac{p_{x=0}}{1-p_{x=0}}) = \beta_0 + \beta_1 \times 0 = \beta_0,$$

Y luego para el grupo más viejo (*x*=1):

$$\text{logit}(p_{x=1}) = \log(Odds_{x=1}) = \log(\frac{p_{x=1}}{1-p_{x=1}}) = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1.$$

Si sustraemos los dos casos

$$\text{logit}\,(p_{x=1}) - \text{logit}\,(p_{x=0}) = \log\,(Odds_{x=1}) - \log\,(Odds_{x=0}),$$

notamos que esta cantidad es igual a β1. Si uno recuerda las propiedades del logaritmo, esto es que la diferencia de dos logaritmos es el logaritmo de su

rango, entonces log ($Odds_{x=1}$) – log ($Odds_{x=0}$) = log ($Odds_{x=1}/Odds_{x=0}$), lo cual puede parecer familiar. Esto es el rango de probabilidades de log o *log odds ratio* en el grupo x=1 relativo al grupo x=0. Por lo tanto, podemos estimar el odds ratio usando regresión logística exponenciando los coeficientes del modelo (a pesar del intercepto, a lo cual llegaremos en un momento).

Ajustemos este modelo, y veamos cómo funciona con un ejemplo real. Ajustamos la regresión logística de una manera muy similar a la que ajustamos los modelos de regresión lineal con algunas pocas excepciones. Primero, usaremos una nueva función llamada `glm`, que es una función muy poderosa en R lo cual permite ajustar una clase de modelos conocidos como modelos lineales generalizados o GLMs (del inglés, Generalized Linear Models) [10]. La función `glm` trabaja de la misma manera que la función *lm*.

Necesitamos especificar una fórmula con el formato:

`resultado_covariables`, especificar que set de datos usar (en nuestro caso el dataframe `dat`), y después especificar la familia. Para la regresión logística nuestra elección será `family="binomial"`. Ud puede ejecutar la función *summary*, de la misma forma que hizo para *lm* y produce un resultado muy similar al que generaba `lm`.

```
age.glm <- glm(day_28_flg ~ age.cat,data=dat,family="binomial")
summary(age.glm)
```

```
##
## Call:
## glm(formula = day_28_flg ~ age.cat, family = "binomial", data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8189  -0.8189  -0.2977  -0.2977   2.5055
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.0944     0.1616  -19.14   <2e-16 ***
## age.cat>55    2.1740     0.1785   12.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1557.9  on 1775  degrees of freedom
## Residual deviance: 1348.7  on 1774  degrees of freedom
## AIC: 1352.7
##
## Number of Fisher Scoring iterations: 5
```

Como Ud puede ver, obtenemos una tabla de coeficientes que es similar a la tabla `lm` que usamos antes. En vez de un `t value`, obtenemos un `z`

`value`, pero esto puede ser interpretado de manera similar. La columna más a la derecha, es un valor-p, para testear la hipótesis nula $\beta=0$. Si usted recuerda, los coeficientes no intercepto son log-odds ratio, así que probar si son 0 es equivalente a probar si el odds ratio es 1. Si un odds ratio es 1, las probabilidades del resultado son iguales en cada grupo. Entonces, determinar si los coeficientes son 0, será un aspecto importante al hacer este tipo de análisis.

Mirando más en detalle a los coeficientes. El intercepto es -3.09 y el coeficiente `age.cat` es 2.17. El coeficiente para `age.cat` es el log odds ratio para la tabla 2 x 2 que hicimos previo al análisis. Cuando exponenciamos 2.17, obtenemos *exp (*2.17*)* =8.79. Esto se corresponde con el estimado usando la tabla 2 x 2. Para completar, miremos los otros coeficientes, el intercepto. Si usted recuerda, log ($Odds_{x=0}$) =β_0, así que β_0 es el log de las probabilidades del resultado del grupo más joven. Exponenciando otra vez, *exp (*-3.09*)* =0.045, y esto se corresponde con el análisis que hicimos previamente. De manera similar, log ($Odds_{x=1}$) = β_0+ β_1, y las probabilidades estimadas de muerte a los 28 días en el grupo más viejo es *exp* (-3.09+2.17) = 0.4, como vimos más arriba. Convertir los Odds estimados en una probabilidad puede ser hecho directamente usando la función *plogis*, pero mostraremos una forma más poderosa y fácil de hacer esto más adelante en esta sección.

Más allá de una Única Covariable Binaria

Mientras el análisis previo es útil para ilustrar, no demuestra nada que no pudiéramos hacer con nuestro ejemplo previo de tabla 2 x 2. La regresión logística nos permite extender la idea básica a por lo menos dos áreas muy relevantes. La primera es el caso donde tenemos más de una covariable de interés. Tal vez, tenemos un confundidor, estamos preocupados por él, y queremos ajustar por él. Alternativamente, quizás hay dos covariables de interés. Segundo, permite usar covariables como cantidades continuas, en lugar de discretizarlas dentro de categorías. Por ejemplo en vez de dividir edad en estratos amplios (como hicimos muy simplemente dividiéndola en dos grupos, ≤55 y >55), podríamos usar la edad como una covariable continua.

Primero, tener más de una covariable es simple. Por ejemplo, si quisiéramos agregar `service_unit` a nuestro modelo previo, solo lo

agregaríamos como hicimos cuando usamos la función *lm* para regresión lineal. Aquí especificamos *~day_28-flg age.cat+service_unit* y ejecutamos la función *summary.*

Se genera una tabla de coeficientes, y ahora tenemos cuatro coeficientes estimados. Los mismos dos, (*Intercept) y age.cat* los cuales fueron estimados en el modelo sin ajustar, pero también tenemos *service_unit_MICU* y *service_unit_SICU* que corresponden a los log odds ratio para MICU y SICU en relación a FICU. Tomando la exponencial de estos resultará en un odds ratio para cada variable, ajustado para las otras variables en el modelo. En este caso en los odds ratios ajustados para Edad>55, MICU y SICU son 8.68,3.25 y 3.08 respectivamente. Concluiríamos que hay al menos un incremento de 9 veces en las probabilidades de mortalidad a los 28 días para aquellos en el grupo >55 años en comparación con el grupo más joven ≤55, mientras mantengamos la unidad de servicio constante. Este ajuste se vuelve importante en varios escenarios en los cuales los grupos de pacientes puedan tener mayor o menor probabilidad de recibir tratamiento, pero también más o menos probabilidad de tener mejores resultados, donde un efecto puede ser confundido posiblemente por muchos otros. Este es casi siempre el caso con los datos observacionales, y es el porque la regresión logística es una herramienta de análisis de datos poderosa en este ámbito.

Otro caso que nos gustaría abordar es cuando tenemos una covariable continua que quisiéramos incluir en el modelo. Uno puede siempre fragmentar la covariable continua en categorías mutuamente excluyentes seleccionando puntos de corte, pero la selección del número y lugar de estos puntos puede ser arbitrario, y en muchos casos innecesario o ineficiente.

Recordemos que en regresión logística estamos ajustando un modelo:

$$\text{logit}(p_x) = \log(Odds_x) = \log(\frac{p_x}{1-p_x}) = \beta_0 + \beta_1 \times x,$$

Pero ahora asumamos que *x* es continua. Imagine un escenario hipotético donde β0 y β1 son conocidas y tiene un grupo de 50 años, y un grupo de 51 años. La diferencia en el log Odds entre los dos grupos es:

$$\begin{aligned}\log(Odds_{51}) - \log(Odds_{50}) &= (\beta_0 + \beta_1 \times 51) - (\beta_0 + \beta_1 \times 50) = \beta_1(51-50) \\ &= \beta_1.\end{aligned}$$

Entonces, el odds ratio para 51 años vs. 50 años es exp (β1). Esto es verdad para cualquier grupo de pacientes que tienen un año de diferencia, y esto es una manera útil de interpretar y usar estos coeficientes estimados para covariables continuas. Trabajemos con un ejemplo. De nuevo, ajustando el resultado de la mortalidad a los 28 días como una función de edad, pero tratando la edad como fue originalmente registrada en el set de datos, una variable continua llamada *age*.

```
agects.glm <- glm(day_28_flg ~ age,data=dat,family="binomial")
summary(agects.glm)$coef
```

```
##               Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept) -5.77800634 0.320774776 -18.01266 1.550034e-72
## age          0.06523274 0.004469569  14.59486 3.028256e-48
```

Vemos que el coeficiente estimado es 0.07 y todavía muy estadísticamente significativo. Exponenciando el log odds ratio para la edad, obtenemos un odds ratio de 1.07, que es por cada año de aumento de edad. ¿Qué pasaría si la diferencia de edad de interés fuera diez años en vez de uno? Hay por lo menos dos formas de hacer esto. Una es reemplazar *age* por *(age/10),* que usa una nueva covariable que es *age* divido por diez. La segunda es usar el log odds ratio estimado *agects.glm,* y multiplicarlo por diez antes de exponenciarlo. Darán equivalentes estimados de 1.92, pero ahora es por cada incremento de 10 años en la edad. Esto es útil cuando el odds ratio estimado (o log odds ratio) son cercanos a uno (o cero). Cuando esto es hecho, una unidad de la covariable es 10 años, así que la interpretación genérica de los coeficientes se mantiene igual, pero las unidades cambian (cada 10 años en vez de cada 1 año).

Esto por supuesto asume que la forma de nuestra ecuación que relaciona el log odds del resultado a la covariable es correcta. En casos donde el Odds del resultado disminuye o aumenta como una función de la covariable, es posible estimar un efecto relativamente pequeño de la covariable lineal, cuando el resultado puede estar fuertemente afectado por la covariable, pero no en la forma en la que está especificado en el modelo. Es posible evaluar en forma gráfica la linealidad del log Odds del resultado y alguna forma discretizada de la covariable. Por ejemplo podemos separar en 5 grupos y estimar el log Odds de la mortalidad a los 28 días en cada grupo.

Graficando estas cantidades en la Fig. 16.5 (izquierda), podemos ver en este caso particular que la edad está efectivamente fuertemente relacionada con el Odds del resultado. Además, expresar la edad en forma lineal pareciera ser una buena aproximación. Si, por otro lado, la mortalidad a los 28 días tiene una curva en forma de "U", podríamos falsamente concluir que no hay relación entre la edad y la mortalidad, cuando la relación puede ser bastante fuerte. Este podría ser el caso cuando en la Fig. 16.5 (derecha) miramos el log odds de mortalidad para la primera temperatura (*temp_1st*).

16.3.5 Test de la Hipótesis y Selección de Modelo

Al igual que en el caso de la regresión lineal, hay una forma de evaluar las hipótesis para la regresión logística. Sigue casi el mismo marco de referencia, con la hipótesis nula siendo β=0. Si usted recuerda, este es el log odds ratio, y probar si es cero es equivalente a probar que el odds ratio sea uno. En este capítulo, nos enfocamos en cómo realizar esta evaluación en R.

Como era el caso cuando utilizábamos *lm*, primero ajustamos los dos modelos que competían, uno más grande (modelo alternativo) y uno más pequeño (modelo nulo). Dado que los modelos son anidados, podemos usar de nuevo la función *anova*, pasando el modelo más pequeño y luego el más grande. Aquí nuestro modelo más grande es el que contenía *service_unit* y *age.cat*, y el modelo más pequeño solo contiene *age.cat*, entonces están anidados. Estamos entonces evaluando si el log odds ratio para los dos coeficientes asociados con *service_unit* son cero. Llamemos estos coeficientes *βMICU y βSICU*. Para evaluar si *βMICU y βSICU*=0, podemos usar la función *anova*, donde esta vez especificaremos el tipo de prueba, en este caso estableceremos el parámetro *test a "Chisq"*.

```
anova(age.glm,ageunit.glm,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: day_28_flg ~ age.cat
## Model 2: day_28_flg ~ age.cat + service_unit
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1774     1348.7
## 2      1772     1343.8  2   4.9315  0.08495 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Aquí el producto de la función *anova* cuando se aplica a objetos *glm* se parece al producto generado cuando se usa en objetos *lm*. Un par de buenas prácticas a incorporar como hábito son primero asegurarse que los dos modelos que compiten sean correctamente especificados. Aquí estamos evaluando *~age.cat* contra *age.cat+service_unit*. A continuación la diferencia entre los grados de libertad residuales *(Resid. Df)* en los dos modelos nos dice cuántos más parámetros tiene el modelo más grande cuando se lo compara con modelo más pequeño.

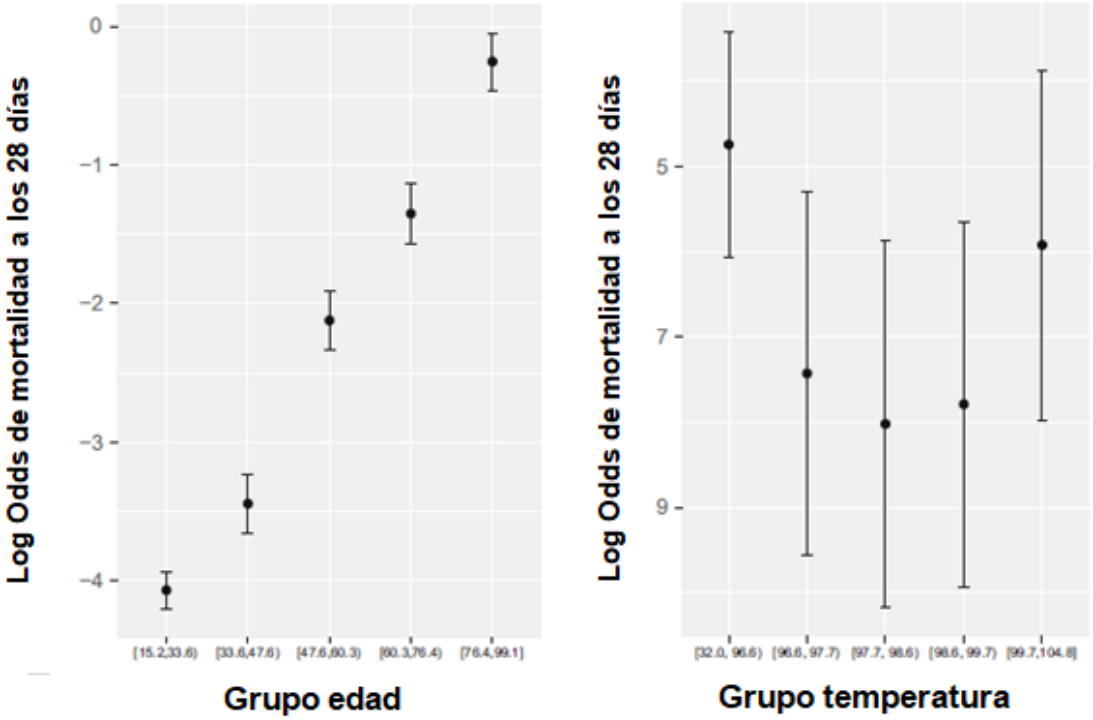

Fig. 16.5 Grafico de log Odds de mortalidad para cada uno de los cinco grupos de edad y temperatura. *Las barras de error* representan el 95% de intervalo de confianza para el log Odds.

Aquí vemos *1774-1772=2* lo cual significa que hay dos coeficientes estimados más en el modelo más grande que en el más pequeño lo cual corresponde con el producto de la tabla *summary* más arriba. Luego mirando el valor p *(Pr (>Chi))*, vemos que una prueba para *βMICU y βSICU=0* tiene un valor p de alrededor de 0.08. Al típico nivel de significación 0.05, no refutaríamos la nulidad, y usaríamos el modelo más simple sin la unidad de servicio. En la regresión logística esta es una forma común de evaluar si una covariable categórica debería retenerse en el modelo, dado que puede ser difícil determinar el uso del valor Z en la tabla *summary*,

particularmente cuando una es muy estadísticamente significativa y la otra no.

16.3.6 Intervalos de confianza

Generar los intervalos de confianza para tanto el log odds ratio como para los odds ratios es bastante sencillo. Vemos como obtener el log odds ratio y los intervalos de confianza respectivos para el modelo *ageunit.glm* el cual incluye tanto edad como unidad de servicio.

```
ageunit.glm$coef
```

```
##      (Intercept)       age.cat>55 service_unitMICU service_unitSICU
##        -4.209013         2.161142         1.178865         1.123442
```

```
confint(ageunit.glm)
```

```
##                         2.5 %    97.5 %
## (Intercept)      -5.66202924 -3.139732
## age.cat>55        1.82211403  2.524682
## service_unitMICU  0.12291680  2.620797
## service_unitSICU  0.07182767  2.563132
```

Aquí los coeficientes estimados y los intervalos de confianza están presentados en casi la misma forma que para la regresión lineal. En la regresión logística, en general es conveniente exponenciar estas cantidades para obtener el resultado en una escala más interpretable.

```
exp(ageunit.glm$coef[-1])
```

```
##       age.cat>55 service_unitMICU service_unitSICU
##         8.681049         3.250684         3.075423
```

```
exp(confint(ageunit.glm)[-1,])
```

```
##                    2.5 %   97.5 %
## age.cat>55       6.18492 12.48693
## service_unitMICU 1.13079 13.74668
## service_unitSICU 1.07447 12.97640
```

En forma similar a la regresión lineal, miraremos si los intervalos de confianza para el log odds ratio incluyen cero. Esto es equivalente a mirar si los intervalos para el odds ratio incluyen 1. En general es más conveniente reportar los odds ratios en vez de los coeficientes en la escala de odds ratio ya que son interpretables en forma más directa.

16.3.7 Predicción

Una vez que haya decidido su modelo final, usted puede querer generar predicciones a partir de su modelo. Dicha tarea puede ocurrir al hacer un análisis por puntaje de propensión (Capítulo 25) o al crear herramientas para apoyo de decisiones clínicas. En el ámbito de la regresión logística esto implica intentar estimar la probabilidad del resultado dada las características (covariables) de un paciente. Esta cantidad, con frecuencia se denomina *P (outcome|X)* y es relativamente fácil de obtener en R usando la función *predict.* Uno debe pasar un set de datos con todas las variables contenidas en el modelo. Asumamos que decidimos incluir *service_unit* en nuestro modelo final, y queremos generar predicciones a partir de esto en un nuevo set de pacientes. Primero creemos un nuevo *dataframe* llamado *newdat* usando la función *expand.grid* la cual calcula todas las combinaciones de los valores de las variables incluidas.

```
newdat <- expand.grid(age.cat=c("<=55",">55"),service_unit=c("FICU","MICU","SICU"))
newdat$pred <- predict(ageunit.glm,newdata=newdat,type="response")
newdat
```

```
##   age.cat service_unit       pred
## 1    <=55         FICU 0.01464341
## 2     >55         FICU 0.11426771
## 3    <=55         MICU 0.04608233
## 4     >55         MICU 0.29546130
## 5    <=55         SICU 0.04370639
## 6     >55         SICU 0.28405645
```

Seguimos esto agregando una columna *pred* a nuestro nuevo *dataframe* usando la función *predict*. La función *predict* para regresión logística funciona de manera similar a cuando se usa para la regresión lineal, pero en este caso también especificamos *type="response"* lo cual asegura que las cantidades calculadas son las que necesitamos, *P (outcome|X)*. El resultado de este nuevo objeto muestra nuestra probabilidad de mortalidad predicha a los 28 días para seis pacientes hipotéticos. Dos en cada unidad de servicio, donde uno se encuentra en el grupo más joven y otro en el más viejo. Vemos que nuestra predicción más baja es para los pacientes más jóvenes en la FICU, mientras que los pacientes con mayor riesgo de mortalidad a los 28 días se encuentran en el grupo más viejo en la MICU, pero la probabilidad predicha no es tanto más alta que los pacientes de la misma edad en SICU.

Para realizar predicciones en un set de datos diferente, solo hay que reemplazar el argumento `newdata` con el otro set de datos. Podríamos, por ejemplo, pasar `newdata=dat` y realizar predicciones para el set de datos sobre el cual construimos el modelo. Como fue el caso con la regresión lineal, evaluar el desempeño predictivo de nuestro modelo en datos usados para construir el modelo generalmente será demasiado optimista en relación a que tan bien se desempeñaría en *el mundo real.* En el capítulo 17 se trata cómo obtener un mejor sentido de la precisión de dichos modelos.

16.3.8 Presentando e Interpretando el Análisis de Regresión Logística

En general, presentar los resultados de un modelo de regresión logística es muy semejante a lo que fue hecho en el ámbito de la regresión lineal. Los resultados deberían siempre ponerse en contexto, incluyendo qué variables fueron consideradas y cuáles variables estuvieron en el modelo final. Reportar los resultados debería siempre incluir alguna forma del coeficiente estimado, una medida de incerteza y un probable valor p. En revistas médicas y epidemiológicas, los coeficientes están usualmente exponenciados de forma tal que no están más en la escala log, y son reportados como odds ratios. Con frecuencia, el análisis multivariado (análisis con más de una covariable) se distingue del univariado (una covariable) denotando los odds ratios estimados como odds ratios ajustados (OR ajustado).

Para el modelo `age.glm`, un ejemplo de lo que podría informarse es:

La mortalidad a los 28 días fue mucho más alta en el grupo de mayor edad (>55 años) que en el más joven (≤55 años), con valores de 28.5 y 4.3% respectivamente (OR=8.79, IC95%: 6.27-12.64, p<0.001)

Cuando se trata la edad como una covariable continua en el modelo `agects.glm` podríamos informar:

La mortalidad a los 28 días se asoció con mayor edad (OR=1.07 por año de aumento, IC95%: 1.06-1.08 p <0.001).

Y para el caso con más de una covariable, `(ageunit.glm)` un ejemplo de lo que podría informarse:

Mayor edad (> 55 vs ≤ 55 años) se asoció en forma independiente con la mortalidad a los 28 días después de ajustar por unidad de servicio (OR ajustado= 8.68, IC95%: 6.18-12.49, p<0.001)

16.3.9 Advertencias y Conclusiones

Como era el caso con la regresión lineal, la regresión logística es una herramienta extremadamente poderosa para el análisis de datos de datos en salud. A pesar de que los resultados del estudio en cada enfoque son diferentes, el marco de referencia y el modo de pensar el problema tienen similitudes. De la misma manera, muchos de los problemas encontrados en la regresión lineal deben tenerse en cuenta en la regresión logística. Los valores *outliers* o atípicos, datos faltantes, colinearidad y resultados dependientes o correlacionados también son problemas para la regresión logística y se puede lidiar con ellos de una manera similar. Del mismo modo, los supuestos del modelo, tema tratado brevemente cuando discutimos si era apropiado usar la edad como una covariable continua en nuestros modelos. A pesar de que las covariables continuas frecuentemente se modelan de esta manera, es importante asegurarse que la relación entre el log Odds del resultado y la covariable es efectivamente lineal. En los casos en donde los datos fueron divididos entre demasiados subgrupos (o el estudio es simplemente demasiado pequeño), uno podría encontrar un nivel de una variable discreta donde ninguno (o muy pocos) o uno de los resultados haya ocurrido. Por ejemplo, si tuviéramos un adicional `service_unit` con 50 pacientes, de los cuales todos vivieron. En un caso semejante, puede no ser apropiado usar el odds ratio estimado y los intervalos de confianza subsecuentes o la hipótesis evaluada. En dicho caso, colapsar la covariable discreta en menos categorías frecuentemente ayudará a llevar el análisis a una forma manejable. Para nuestra hipotética nueva unidad de servicio, crear un nuevo grupo de ella y FICU podría ser una solución posible. A veces una covariable está relacionada muy fuertemente con el resultado, y esto ya no es posible, y la única solución podría ser reportar este hallazgo, y eliminar esos pacientes.

En general, la regresión logística es una herramienta muy *valiosa* para modelar datos binarios y categóricos. Aunque no tratamos este último caso, existe un marco de referencia similar para datos discretos que están ordenados o tienen más de una categoría (ver `?multinom` en el paquete *nnet* en R para detalles acerca de regresión logística multinomial). En [11] se discuten este y otros temas como la evaluación del ajuste del modelo, y el uso de la regresión lineal en diseños de estudio más complicados.

16.4 Análisis de sobrevida

16.4.1 Objetivos de la sección

En esta sección, el lector aprenderá los fundamentos del análisis de sobrevida, y cómo presentar e interpretar este tipo de análisis.

16.4.2 Introducción

Como habrá notado en la sección anterior sobre regresión logística, observamos específicamente el resultado de mortalidad a los 28 días. Esto fue deliberado, e ilustra la limitación del uso de la regresión logística para este tipo de resultados. Por ejemplo, en el análisis previo, alguien que murió en el día 29 fue tratado de manera idéntica que alguien que vivió por más de 80 años. Uno puede preguntarse ¿por qué no tratar el tiempo de sobrevida como una variable continua y hacer el análisis de regresión lineal en este resultado? Hay varias razones, pero la más importante es que uno probablemente no pueda esperar toda una vida por cada participante del estudio. Es probable que en su estudio sólo una fracción de sus sujetos muera antes de que usted esté preparado para publicar sus resultados.

Aunque en general nos concentramos en la mortalidad, esto puede ocurrir para muchos otros resultados, incluyendo tiempo hasta la recaída del paciente, re-hospitalización, reinfección, etc. En cada uno de estos tipos de resultados, se presume que los pacientes están en riesgo del resultado hasta que el evento suceda, o hasta que estén *censurados.* La censura puede ocurrir por una variedad de diferentes razones, pero indica que el evento no fue observado durante el tiempo de observación. En ese sentido, la sobrevida o, más frecuentemente, los datos de tiempo al evento son un resultado bivariable que incorpora el tiempo de observación o estudio en el cual el paciente fue observado y si el evento sucedió durante el período de observación. El caso particular en el que estaremos más interesados es la *censura por la derecha* (los sujetos son observados solamente hasta un punto en el tiempo, y no sabemos lo que pasa después de este punto), pero también existe la *censura por la izquierda* (solamente sabemos que el evento sucedió antes de un punto en el tiempo) y la *censura por intervalo* (los eventos suceden dentro de una ventana de tiempo). La *censura por la derecha* generalmente es la más común, pero es importante entender cómo fueron recolectados los datos para asegurarse que sufrieron efectivamente *censura por la derecha.*

Establecer un origen común en el tiempo (por ejemplo, un lugar para empezar a contar el tiempo) suele ser fácil de identificar (por ejemplo, alta de la unidad de cuidados intensivos, inscripción en un estudio, administración de una droga, etc.), pero en otros escenarios puede que no lo sea (por ejemplo, quizás el interés esté puesto sobre el tiempo de sobrevida desde el comienzo de la enfermedad, pero los pacientes solo son monitoreados desde el momento de diagnóstico). Para ver un buen tratamiento de este tópico y otros asuntos, consulte el Cap. 3 de [12].

Junto con esta complejidad adicional en los datos (en relación a las regresiones logística y lineal) existen aspectos técnicos adicionales y supuestos en los abordajes de análisis de datos. Por lo general, cada abordaje intenta comparar grupos o identificar covariables que modifican las tasas de sobrevida entre los pacientes estudiados.

En general, el análisis de sobrevida global es un área de estudio compleja y fascinante, y aquí únicamente haremos un repaso breve de dos tipos de análisis. Ignoramos en gran medida, los detalles técnicos de estos abordajes, enfocándonos en su lugar en los principios generales y la intuición. Antes de comenzar cualquier análisis de sobrevida, necesitamos cargar el paquete *survival* en R, lo cual puede hacerse ejecutando:

```
library(survival);
```

Normalmente usted podrá saltearse el siguiente paso, pero dado que este set de datos fue utilizado para analizar datos de una manera ligeramente diferente debemos corregir los tiempos de observación para un subconjunto de los pacientes en el set de datos.

```
dat$mort_day_censored[dat$censor_flg==1] <- 731;
```

16.4.3 Curvas de Kaplan-Meier de Sobrevida

Ahora que hemos resuelto las cuestiones técnicas, podemos comenzar visualizando los datos. Así como la tabla de 2 x 2 es un paso fundamental en el análisis de datos binarios, el paso fundamental para los datos de sobrevida suele ser el trazado de lo que se conoce como función de sobrevida de Kaplan-Meier [13]. La función de sobrevida es una función del tiempo y es la probabilidad de sobrevivir al menos esa cantidad de tiempo. Por ejemplo, si hubo una sobrevida del 80% al año, la función de sobrevida a ese momento

es 0,8. Las funciones de sobrevida suelen comenzar a *tiempo=0*, donde la función del sobreviviente es 1 (o 100%, todos están vivos) y solo puede mantenerse igual o decrecer. ¡Si fuera a incrementarse a medida que pasara el tiempo significaría que habría personas volviendo a la vida de la muerte! Las curvas de Kaplan-Meier son una de las curvas más ampliamente utilizadas en la investigación médica.

Antes de graficar la curva de Kaplan-Meier, debemos configurar un objeto *survfit*. Este objeto tiene una forma familiar, pero difiere ligeramente de las metodologías que cubrimos previamente. Especificar una fórmula para un resultado de sobrevida es algo más complicado ya que, como dijimos, el dato de sobrevida tiene dos componentes. Podemos hacer esto creando el objeto *Surv* en R. Este será nuestro resultado de sobrevida para los análisis siguientes.

```
datSurv <- Surv(dat$mort_day_censored,dat$censor_flg==0)
datSurv[101:105]
```

```
## [1] 236.08  731.00+ 731.00+ 731.00+   2.00
```

El primer paso configura un nuevo tipo de objeto en R útil para datos de sobrevida. La función *Surv* normalmente contiene dos argumentos: un vector de tiempo y algún tipo de indicador para aquellos pacientes que tuvieron un evento (muerte en nuestro caso). En nuestro caso, el vector de muerte y los tiempos de censura son *mort_day_censored* y las muertes se codifican con un cero en la variable *censor_flg* (por lo tanto, identificamos el evento allí donde *censor_flg* == 0). El último paso entrega 5 entradas del nuevo objeto (observaciones 101 a 105). Podemos ver que hay tres entradas de 731.00+. El + indica que esta observación es censurada. Las otras entradas no son censuradas indicando muertes en esos momentos.

Construir una curva de Kaplan-Meier es simple luego de haber hecho esto, pero requiere de dos pasos. El primero especifica una fórmula similar a la usada para las regresiones lineal y logística, pero utilizando en este caso la función *survfit*. Queremos que se ajuste por género (*gender_num*), por lo que la fórmula es *datSurv~gender_num*. Luego podemos graficar (comando *plot*) el nuevo objeto creado, pero agregando argumentos adicionales a la función *plot* que incluyen los intervalos de confianza del 95% para las funciones de sobrevida (*conf.int = TRUE*) e incluye una

etiqueta para los ejes x e y (*xlab* e *ylab*). Por último, agregamos una leyenda programando negro para mujeres y rojo para hombres. Este gráfico se encuentra en la Fig. 16.6.

```
gender.surv <- survfit(datSurv~gender_num,data=dat)
plot(gender.surv,col=1:2,conf.int = TRUE,xlab="Days",ylab="Proportion Who Survived")
legend(400,0.4,col=c("black","red"),lty=1,c("Women","Men"))
```

En la Fig. 16.6 aparenta haber una diferencia entre la función de sobrevida entre los 2 grupos por género, con el grupo de hombres (rojo) muriendo a una tasa ligeramente menor que el grupo de mujeres (negro). Hemos incluido líneas para los intervalos de confianza de 95% de la función de sobrevida estimada, que evalúa cuánta certeza tenemos acerca de la sobrevida estimada en cada punto del tiempo. Podemos hacer lo mismo para *service_unit*, pero dado que tiene 3 grupos necesitamos cambiar el argumento de color y la leyenda para asegurarnos que el gráfico sea etiquetado en forma correcta. Observamos dicho gráfico está en la Fig. 16.7.

```
unit.surv <- survfit(datSurv~service_unit,data=dat)
plot(unit.surv,col=1:3,conf.int = FALSE,xlab="Days",ylab="Proportion Who Survived")
legend(400,0.4,col=c("black","red","green"),lty=1,c("FICU","MICU","SICU"))
```

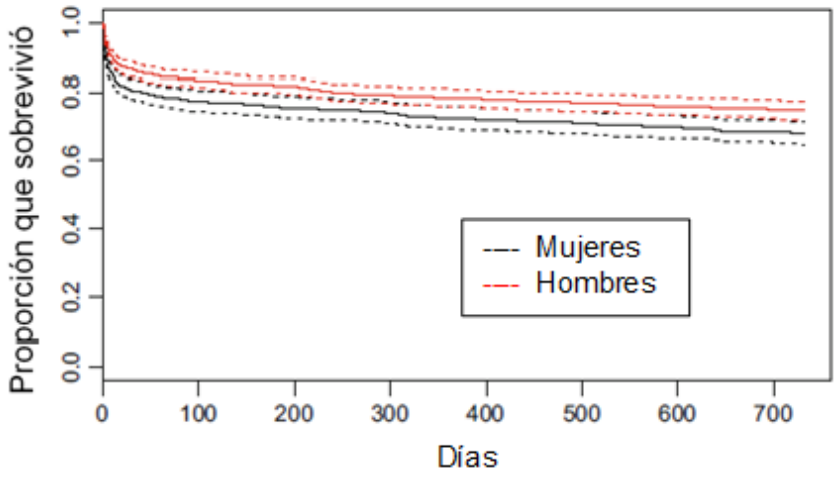

Fig. 16.6 Curva de Kaplan-Meier para la función de sobrevida estimada estratificada por género.

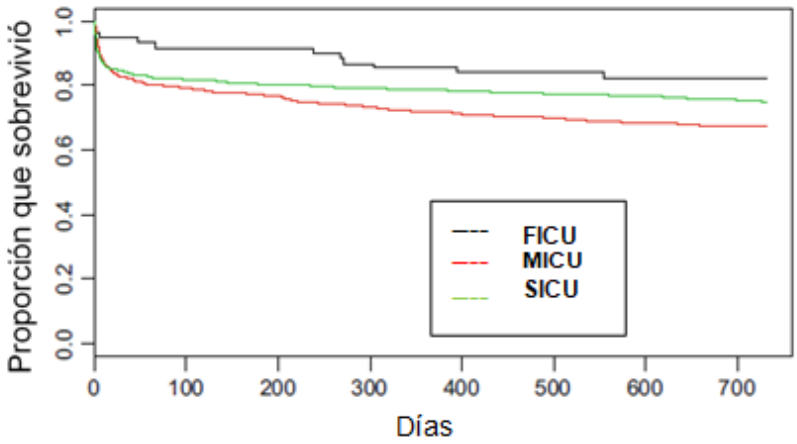

Fig. 16.7 Curva de Kaplan-Meier para la función de sobrevida estimada estratificada por unidad de servicio.

16.4.4 Modelo de riesgos proporcionales de Cox

Las curvas de Kaplan-Meier son un buen primer paso para examinar los datos de tiempo hasta un evento antes de proceder con modelos estadísticos más complejos. Los resultados de tiempo hasta un evento son por lo general más complejos que otro tipo de resultados que hemos examinado hasta aquí. Existen varios modelos de abordaje diferentes, cada uno de los cuales tiene ventajas y limitaciones. El abordaje más popular para datos de salud probablemente sea el Modelo de Riesgos Proporcionales de Cox [14], que también suele llamarse modelo de Cox o regresión de Cox. Como bien implica el nombre, este método modela algo llamado función de riesgo (Hazard Function). No nos detendremos en los detalles técnicos, sino que intentaremos proveer de alguna introducción. La función de riesgo depende del tiempo (horas, días, años) y es aproximadamente la probabilidad instantánea de que ocurra el evento (es decir, la probabilidad de que el evento esté ocurriendo en una ventana de tiempo pequeña) dado que el evento no haya ya ocurrido. Se utiliza frecuentemente para estudiar mortalidad, donde se conoce a veces como *fuerza de mortalidad* o tasa de muerte instantánea y puede ser interpretada de manera simple como el riesgo de muerte en un momento particular, dado que la persona ha sobrevivido hasta ese punto en el tiempo. La parte “proporcional” del modelo de Cox asume que la manera en que las covariables actúan sobre la función de riesgo en los diferentes pacientes es a través de una supuesto de proporcionalidad relativa a la línea de base de la función de riesgo. Para

ilustrarlo, considere un caso simple en el que se administran dos tratamientos. Para el tratamiento 0 (por ejemplo, placebo) determinamos que la función de riesgo es h_0 (t) y para el tratamiento 1 determinamos que la función de riesgo es h_1 (t), donde t es tiempo. El supuesto de riesgos proporcionales es la siguiente:

$$h_1 (t) = HR \times h_0 (t)$$

Es fácil ver que HR = h_1 (t) /h_0 (t). Esta cantidad suele llamarse cociente de riesgo y, si por ejemplo fuera 2, significaría que el riesgo de muerte en el grupo bajo tratamiento 1 es el doble al riesgo en el grupo bajo tratamiento 0. Podemos notar que HR *no* es una función del tiempo, lo que quiere decir que el riesgo de muerte es *siempre* el doble en el primer grupo cuando se lo compara con el segundo. Esta suposición significa que, si el supuesto de riesgos proporcionales es válido, solamente tenemos que conocer la función de riesgo del grupo 0 y el cociente de riesgo (HR) para conocer la función de riesgo del grupo 1. La estimación de la función de riesgo según este modelo suele considerarse una falacia, ya que el foco primario está sobre el cociente de riesgo y esto resulta clave para poder interpretar y ajustar estos modelos. Para un tratamiento técnico más detallado del tópico, lo referimos a [12, 15-17].

De la misma manera que con la regresión logística, modelaremos el logaritmo del cociente de riesgo (Hazard Ratio, HR) en lugar del HR en sí mismo. Esto nos permitirá usar el marco de referencia familiar que hemos utilizado hasta aquí para modelar otros tipos de datos de salud. Como en la regresión logística, cuando el log (HR) es cero, el HR es 1, lo que quiere decir que el riesgo entre los grupos es el mismo. Todavía más, esto se extiende a múltiples covariables del modelo o variables continuas de la misma manera que la regresión logística.

Ajustar los modelos de regresión de Cox en R seguirá el patrón familiar que hemos visto en los casos anteriores de regresiones lineales y logísticas. La función *coxph* (del paquete survival) es la función de ajuste para los modelos de Cox y sigue el patrón general de seguir a una fórmula modelo (*outcome~covariate*) y el set de datos que Ud. quiera utilizar. En nuestro caso, continuemos con nuestro ejemplo de utilizar el género (*gender_num*) para modelar el resultado *dataSurv* que hemos creado y ejecutar la función *summary* para ver qué información resulta.

```
gender.coxph <- coxph(datSurv ~ gender_num,data=dat)
summary(gender.coxph)
```

```
## Call:
## coxph(formula = datSurv ~ gender_num, data = dat)
##
##   n= 1775, number of events= 497
##    (1 observation deleted due to missingness)
##
##                coef exp(coef) se(coef)     z Pr(>|z|)
## gender_num -0.29094   0.74756  0.08978 -3.24  0.00119 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## gender_num    0.7476      1.338    0.6269    0.8914
##
## Concordance= 0.537  (se = 0.011 )
## Rsquare= 0.006   (max possible= 0.983 )
## Likelihood ratio test= 10.43  on 1 df,   p=0.001243
## Wald test            = 10.5  on 1 df,   p=0.001193
## Score (logrank) test = 10.58  on 1 df,   p=0.001146
```

La tabla de coeficientes tiene un formato similar, que ya hemos visto. El *coef* para *gender_num* es aproximadamente –0,29 y este es el estimado de nuestro log-HR. Como fue discutido, tomar el exponente de esto nos da el HR, que el resultado resumen calcula en la siguiente columna *(exp(coef))*. Aquí el HR se estima en 0,75, indicando que los hombres tienen aproximadamente una reducción del 25% en el riesgo de muerte bajo el supuesto de riesgos proporcionales.

La siguiente columna en la tabla de coeficientes tiene el error estándar para el log HR, seguido del score z y el valor p *(Pr(>|z|))*, que es muy similar a lo que vimos en el caso de regresión logística. Aquí vemos que el valor de *p* es bastante pequeño y rechazaríamos la hipótesis nula que sostiene que las funciones de riesgo son las mismas entre hombres y mujeres. Esto es consistente con las figuras exploratorias que produjimos utilizando las curvas de Kaplan-Meier en la sección previa. Para *coxph*, la función *summary* muestra en forma conveniente el intervalo de confianza del HR unas líneas más abajo y aquí nuestro HR estimado es 0,75 (IC95%: 0,63-0,89, p = 0,001). Así es como se reportaría usualmente el HR.

Utilizar más de una covariable funciona de la misma manera que en nuestras otras técnicas de análisis. Agregar una comorbilidad al modelo como la fibrilación auricular (*afib_flg*) puede hacerse de la manera que se haría para la regresión logística.

```
genderafib.coxph <- coxph(datSurv~gender_num + afib_flg,data=dat)
summary(genderafib.coxph)$coef
```

```
##                 coef exp(coef)   se(coef)         z    Pr(>|z|)
## gender_num -0.2591201 0.7717304 0.08987143 -2.883231 0.003936189
## afib_flg    1.3443975 3.8358747 0.10200099 13.180239 0.000000000
```

Aquí el género masculino se asocia con una reducción del tiempo hasta la muerte, mientras que la fibrilación auricular incrementa el riesgo de muerte casi 4 veces. Ambos son estadísticamente significativos en el resultado resumen y sabemos desde antes que podemos probar una gran cantidad de otros tipos de hipótesis estadísticas utilizando la función anova. Nuevamente ejecutamos un anova al modelo anidado más pequeño (*gender_num_only*) y al más grande (*gender_num and afibflg*)

```
anova(gender.coxph,genderafib.coxph)
```

```
## Analysis of Deviance Table
##  Cox model: response is  datSurv
##  Model 1: ~ gender_num
##  Model 2: ~ gender_num + afib_flg
##    loglik  Chisq Df P(>|Chi|)
## 1 -3636.1
## 2 -3567.4 137.37  1 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como era esperado, la fibrilación auricular es muy estadísticamente significativa y, por lo tanto, querríamos conservarla en el modelo.

La regresión de Cox también permite utilizar covariables que cambian en el tiempo. Esto nos permitiría incorporar cambios en tratamiento, en la severidad de la enfermedad, etc. en el mismo paciente sin necesidad de implementar una metodología diferente. El mayor desafío para lograr esto reside en la construcción del set de datos, tema que se discute en alguna de las referencias del final del capítulo. Se necesita algo de cuidado cuando la covariable dependiente de tiempo solo se mide periódicamente, ya que el método requiere que sea conocido cada tiempo de evento para la cohorte completa de pacientes y no solo aquellos relevantes para el paciente en cuestión. Esto es más práctico para cambios en el tratamiento que pueden registrarse con precisión, particularmente en bases de datos como MIMIC II, y menos práctico para resultados de laboratorio que pueden ser medidos cada horas, días o semanas. Se ha visto que interpolar entre valores de

laboratorios o arrastrar la última observación introduce varios tipos de problemas.

16.4.5 Advertencias y conclusiones

Concluiremos esta breve revisión de análisis de sobrevida sabiendo que solo hemos tocado el tema en forma superficial. Existen muchos tópicos que no hemos cubierto o que han sido abordados brevemente.

El análisis de sobrevida se diferencia de otras formas de análisis tratadas en este Capítulo ya que permite que los datos sean censurados. Como fue el caso de otros abordajes que consideramos, existen supuestos del modelo. Por ejemplo, es importante que la censura no sea informativa del tiempo de sobrevida. Por ejemplo, si la censura ocurre cuando el tratamiento se retira porque el paciente está muy enfermo para continuar la terapia constituiría un ejemplo de censura informativa. Entonces, se vuelven inválidos todos los métodos discutidos en esta sección. Se debe tener cuidado de asegurarse que se entienden los mecanismos de censurado para evitar cualquier falsa inferencia.

La evaluación de los supuestos de riesgos proporcionales constituye una parte importante del análisis de regresión de Cox. Lo referimos a las referencias al final del capítulo (particularmente [17] y ver `?cox.zph`) para estrategias y alternativas para cuando se viola el supuesto de riesgo proporcional. En algunas circunstancias, el supuesto de riesgo proporcional no es válido y pueden utilizarse abordajes alternativos. Como siempre, cuando los resultados son dependientes (por ejemplo, un paciente puede contribuir con más de una observación), no deben utilizarse los métodos discutidos en esta sección en forma directa. Generalmente las estimaciones de error estándar serán muy pequeñas y los valores de p serán incorrectos. La preocupación por los valores “outliers”, la colinealidad, datos faltantes y las covariables con escasos resultados de la regresión logística también aplican aquí, al igual que las preocupaciones respecto a especificaciones erróneas del modelo para las covariables continuas.

El análisis de sobrevida es una herramienta técnica poderosa que resulta muy relevante en los estudios en salud. Solamente hemos dado un breve repaso de la materia y lo alentamos a explorar todavía más estos métodos.

16.5 Caso de Estudio y resumen

16.5.1 Objetivos de la sección

En esta sección trabajaremos con un caso de estudio y discutiremos los componentes de análisis de datos que deben incluirse en un artículo de investigación original apto para una revista científica. También discutiremos algunos abordajes para la selección de modelos y atributos.

16.5.2 Introducción

Utilizaremos entonces lo que hemos aprendido en las secciones previas para examinar si los catéteres arteriales invasivos (CAI) tienen algún efecto sobre la mortalidad de los pacientes. Como ya fue reiterado, la identificación clara del objetivo de estudio es importante para un buen análisis de datos. En nuestro caso, nos gustaría estimar el efecto de los CAI sobre la mortalidad, pero reconocemos algunas áreas de potenciales problemas. Primero, los grupos que reciben CAI y aquellos que no lo reciben suelen ser muy diferentes en muchos aspectos y muchas de estas diferencias probablemente tengan a su vez algún efecto sobre la mortalidad. Segundo, nos gustaría ser capaces de limitarnos a eventos de mortalidad que ocurren en la proximidad del alta de la unidad de cuidados intensivos. El set de datos incluye la mortalidad a 28 días, lo que parecería ser próximo al alta de la UCI. Para el primer problema también tenemos muchas covariables que capturan muchos de los atributos que pueden preocuparnos, incluyendo la severidad de enfermedad (`sapsi_first and sofa_first`), edad (`age`), género del paciente (`gender_num`) y comorbilidades (`chf_flg, afib_flg, renal_flg, etc.`).

Con todo esto en mente, deberíamos comenzar por determinar nuestro objetivo de estudio. En nuestro caso podría ser:

Estimar el efecto del uso de CAI durante una internación en UCI sobre la mortalidad a 28 días en pacientes dentro del estudio MIMIC II, que recibieron asistencia respiratoria mecánica, ajustado por la edad, el género, la severidad de la enfermedad y comorbilidades.

Por ahora esto describe bastante bien nuestro resultado y a las covariables. Una de las primeras cosas que suele hacerse es describir a la población calculando las estadísticas resumen de todo el subconjunto de

variables recolectadas en el estudio. Esta descripción le permite al lector entender cuán bien el estudio podría generalizarse a otras poblaciones. Hemos puesto disponible un paquete en R en GitHub que le permitirá construir formas preliminares de dicha tabla de manera rápida. Para instalar el paquete R, primero instale y cargue el paquete "`devtools`":

```
install.package("devtools")
library(devtools)
```

y luego instale y cargue nuestro paquete utilizando la función *`install_github`*.

```
install_github("jraffa/MIMICbook")
library(MIMICbook);
```

Antes de hacer cualquier análisis profundo asegurémonos que estamos utilizando el set de datos original, primero eliminando y luego recargando el dataframe. *`dat`*. Para asegurarnos de que nuestra investigación sea reproducible es una buena idea asegurarnos documentar todo el proceso de análisis. Empezando desde la copia original del set de datos seremos capaces de presentar de manera precisa qué métodos utilizamos en un análisis.

```
rm(dat)
dat <- read.csv(url)
```

Como mencionamos anteriormente, la recodificación de las variables binarias codificadas (aquellas que son 0s y 1s) a datos de clase factor en R a veces puede facilitar la interpretación del resultado de R. El siguiente trozo de código cicla a través de todas las columnas en *`dat`* y convierte cualquier variable binaria en un factor.

```
# Identify which columns are binary coded
bincols <- colMeans((dat == 1 | dat == 0), na.rm = T) == 1
for (i in 1:length(bincols)) {
    # Turn the binary columns into a factor
    if (bincols[i]) {
        dat[[i]] <- as.factor(dat[[i]])
    }
}
```

Así estamos listos para generar un resumen de las características de los pacientes de nuestro estudio. El paquete MIMICbook tiene una función *`produce.table1`*. Esta función genera una tabla resumen del dataframe que se le presenta, utilizando un resumen adecuado para variables continuas (promedio y desviación estándar) y variables categóricas (números y porcentajes) para cada variable. En su forma más simple,

produce.table1 puede ejecutarse en un dataframe como un argumento, cosa que haremos (pasando el dataframe dat). Este resultado no es muy agradable y podemos hacerlo más ameno utilizando un poderoso paquete de R llamado *knitr*, que provee muchas herramientas que colaboran en la realización de investigaciones reproducibles. Puede averiguar más sobre *knitr* (que puede instalarse utilizando *install.packages ('knitr')*) ejecutando *?knitr* en la consola de R luego de cargarlo. Utilizaremos el comando *kable*, que tomará nuestra variable *tab1* –una tabla resumen que generamos utilizando la función *produce.table1*– y la hará lucir un poco mejor.

```
tab1 <- produce.table1(dat);
library(knitr);
kable(tab1,caption = "Overall patient characteristics")
```

Los descriptores de filas no son muy informativos y lo que hemos producido no sería utilizable para una publicación final, pero cumple con nuestro propósito por ahora. *knitr* permite que uno exporte estas tablas en HTML, LaTeX o incluso en un documento de Word, que usted puede editar, haciendo la tabla más informativa. Los resultados están contenidos en la Tabla 16.1.

Podemos notar un par de cosas de las características de base:

1. Algunas variables tienen muchas observaciones faltantes (por ejemplo: *bmi, po2_fist, iv_day_1*).
2. Ninguno de los pacientes tiene sepsis.

Estos dos puntos son importantes e ilustran por qué siempre es una buena idea realizar un análisis básico descriptivo antes de comenzar cualquier modelado.

Tabla 16.1 Características generales de los pacientes

	PROMEDIO (DS), o N (%)
aline_flg==1	984 (55.4%)
icu_los_day	3.3 (3.4)
hospital_los_day	8.1 (8.2)

Age	54.4 (21.1)
gender_num==1	1025 (57.7%) (Faltantes: 1)
weight_first	80.1 (22.5) (Faltantes: 110)
bmi	27.8 (8.2) (Faltantes: 466)
sapsi_first	14.1 (4.1) (Faltantes: 85)
sofa_first	5.8 (2.3) (Faltantes: 6)
service_unit==SICU	982 (55.3%)
service_num==1	982 (55.3%)
day_icu_intime==Saturday	278 (15.7%)
day_icu_intime_num	4.1 (2)
hour_icu_intime	10.6 (7.9)
hosp_exp_flg==0	1532 (86.3%)
icu_exp_flg==0	1606 (90.4%)
day-28-flg==0	1493 (84.1%)
mort_day_censored	614.3 (403.1)
censor_flg==1	1279 (72%)
sepsis_flg==0	1776 (100%)
chf_flg==0	1563 (88%)
afib=flg==0	1569 (88.3%)
renal_flg==0	1716 (96.6%)
liver_flg==0	1677 (94.4%)
copd_flg==0	1619 (91.2%)
cad_flg==0	1653 (93.1%)

stroke_flg==0	1554 (87.5%)
mal_flg==0	1520 (85.6%)
resp_flg==0	1211 (68.2%)
map_1st	88.2 (17.6)
hr_1st	87.9 (18.8)
temp_1st	97.8 (4.5) (Faltantes: 3)
spo2_1st	98.4 (5.5)
abg_count	6 (8.7)
wbc_first	12.3 (6.6) (Faltantes: 8)
hgb_first	12.6 (2.2) (Faltantes: 8)
platelet_first	246.1 (99.9) (Faltantes: 8)
sodium_first	139.6 (4.7) (Faltantes: 5)
potassium_first	4.1 (0.8) (Faltantes: 5)
tco2_first	24.4 (5) (Faltantes: 5)
chloride_first	103.8 (5.7) (Faltantes: 5)
aline_flg==1	984 (55.4%)
bun_first	19.3 (14.4) (Faltantes: 5)
creatinine_first	1.1 (1.1) (Faltantes: 6)
po2_first	227.6 (144.9) (Faltantes: 186)
pco2_first	43.4 (14) (Faltantes: 186)
iv_day_1	1622.9 (1677.1) (Faltantes: 143)

Los datos faltantes están primordialmente relacionados a peso/IMC o valores de laboratorio. Para el propósito de este capítulo, vamos a ignorar ambas clases de variables. Aunque podríamos desear ajustar por alguna de

estas covariables en una versión final del articulo científico a publicar y el Cap. 11 nos brinda técnicas útiles para lidiar con esta situación, nos vamos a focalizar en el conjunto de covariables que hemos identificado en el objetivo del estudio, que no incluyen estas variables. El tópico relacionado con la sepsis también es destacable. Con certeza la sepsis contribuiría a tasas de mortalidad más elevadas en comparación a pacientes sin sepsis, pero como no tenemos pacientes con sepsis, no podemos ni necesitamos ajustar esta covariable por sí misma. Lo que necesitamos es asentar este aspecto revisando el objetivo del estudio. Originalmente identificamos nuestra población como pacientes dentro de MIMIC, pero debido a que este es un subconjunto de MIMIC –aquellos sin sepsis–, deberíamos actualizar el objetivo del estudio a:

Estimar el efecto del uso de CAI durante una internación en UCI sobre la mortalidad a 28 días en pacientes sin sepsis dentro del estudio MIMIC II, que recibieron asistencia respiratoria mecánica, ajustado por la edad, el género, la severidad de la enfermedad y comorbilidades.

A su vez no querremos incluir la variable `sepsis_flg` como covariable en ninguno de nuestros modelos, ya que no hay pacientes con sepsis dentro del estudio para estimar su efecto. Ahora que hemos examinado las características basales generales de los pacientes podemos comenzar los siguientes pasos del análisis.

Los siguientes pasos variarán ligeramente, pero suele ser útil ponerse en el papel de un colega revisor. ¿Qué problemas podría encontrar un revisor en su estudio y cómo puede responder a ellos? Por lo general, el revisor querrá ver cómo difiere la población para diferentes valores de la covariable de interés. En nuestro caso de estudio, si el grupo tratado (CAI) difiere sustancialmente del grupo no tratado (no CAI), entonces esto podría explicar cualquier efecto que demostremos. Podemos hacer esto resumiendo ambos grupos de manera similar a lo que fue hecho en la Tabla 16.1. Podemos reutilizar la función `produce.table,1` pero la ejecutamos para ambos grupos de manera separada al dividir el dataframedat en dos utilizando la función `Split` (según la variable `aline_flg`), para luego combinarlos en una tabla utilizando `cbind` para lograr la Tabla 16.2. Es importante asegurar que se utilizan los mismos grupos de referencia a través de los dos grupos de

estudio y para esto se utiliza el argumento etiquetas (vea *?produce.table1* para más detalles).

```
datby.aline <- split(dat, dat$aline_flg)
reftable <- produce.table1(datby.aline[[1]])
tab2 <- cbind(produce.table1(datby.aline[[1]], labels = attr(reftable, "labels")),
    produce.table1(datby.aline[[2]], labels = attr(reftable, "labels")))
colnames(tab2) <- paste0("Average (SD), or N (%)", c(", No-IAC", ", IAC"))
kable(tab2, caption = "Patient characteristics stratified by IAC administration")
```

Tabla 16.2 Características de los pacientes estratificadas según la administración de CAI

	Promedio (DS) o N (%) sin IAC	Promedio (DS) o N (%) IAC
aline_flg==0	792 (100%)	0 (0%)
icu_los_day	2.1 (1.9)	4.3 (3.9)
hospital_los_day	5.4 (5.4)	10.3 (9.3)
Age	53 (21.7)	55.5 (20.5)
gender_num==1	447 (56.5) (missing: 1)	578 (58.7%)
weight_first	79.2 (22.6) (missing: 7)	80-7 (22.4) (missing: 39)
Bmi	28 (9.1) (missing: 220)	27.7 (7.5) (missing: 246)
sapsi_first	12.7 (3.8) (missing: 70)	15.2 (4) (missing: 15)
sofa_first	4.8 (2.1) (missing: 4)	6.6 (2.2) (missing: 2)
service_unit==MICU	480 (60.6%)	252 (25.6%)
service_num==0	504 (63.6%)	290 (29.5%)
day_icu_intime==Saturday	138 (17.4%)	140 (14.2%)
day_icu_intime_num	4 (2)	4.1 (2)
hour_icu_intime	9.9 (7.7)	11.2 (8.1)
hosp_exp_flg==0	702 (88.6%)	830 (84.3%)
icu_exp_flg==0	734 (92.7%)	872 (88.6%)

day_28_flg==0	679 (85.7%)	814 (82.7%)
mort_day_censored	619.1 (388.3)	610.5 (414.8)
censor_flg==1	579 (73.1%)	700 (71.1%)
sepsis_flg==0	792 (100%)	984 (100%)
chf_flg==0	695 (87.8%)	868 (88.2%)
afib_flg==0	710 (89.6%)	859 (87.3%)
renal_flg==0	764 (96.5%)	952 (96.7%)
liver_flg==0	754 (95.2%)	923 (93.8%)
copd_flg==0	711 (89.8%)	908 (92.3%)
cad_flg==0	741 (93.6%)	912 (92.7%)
stroke_flg==0	722 (91.2%)	832 (84.6%)
mal_flg==0	700 (88.4%)	820 (83.3%)
resp_flg==0	514 (64.9%)	697 (70.8%)
map_1st	87.5 (15.9)	88.9 (18.8)
hr_st	88.4 (18.8)	87.5 (18.7)
temp_1st	97.9 (3.8) (missing: 3)	97.7 (5.1)
spo2_1st	98.4 (5.7)	98.5 (5.4)
abg_count	1.4 (1.6)	9.7 (10.2)
wbc_first	11.7 (6.5) (missing: 6)	12.8 (6.6) (missing: 2)
hgb_first	12.7 (6.5) (missing: 6)	12.4 (2.2) (missing: 2)
platelet_first	254.3 (104.5) (missing: 6)	239.5 (95.6) (missing: 2)
sodium_first	139.8 (4.8) (missing: 3)	139.4 (4.7) (missing: 2)
potassium_first	4.1 (0.8) (missing: 3)	4.1 (0.8) (missing: 2)

tco2_first	24.7 (4.9) (missing: 3)	24.2 (5.1) (missing: 2)
chloride_first	103.3 (5.4) (missing: 3)	104.3 (5.9) (missing: 2)
bun_first	18.9 (14.5) (missing: 3)	19.6 (14.3) (missing: 2)
creatinine_first	1.1 (1.2) (missing. 4)	1.1 (1) (missing: 2)
po2_first	223.8 (152.9) (missing: 178)	230.1 (139.6) (missing: 8)
pco2_first	44.9 (15.9) (missing: 178)	42.5 (12.5) (missing: 8)
iv_day_1	1364.2 (1406.8) (missing: 110)	1808.4 (1825) (missing: 33)

Como se puede ver en la Tabla 16.2, el grupo CAI difiere en muchos puntos respecto del grupo no-CAI. Los pacientes que requirieron un CAI tendían a presentar mayor severidad de la enfermedad de base (*sapsi_first* y *sofa_first*), a tener edad ligeramente mayor, menor probabilidad de provenir de la UCI clínica (MICU) y a tener perfiles de comorbilidad ligeramente diferentes en comparación al grupo no-CAI.

A continuación, podemos ver cómo se distribuyen las covariables entre los diferentes resultados (muerte dentro de los 28 días contra sobrevida a los 28 días). Esto nos dará una idea de qué covariables pueden ser importantes para afectar el resultado. El código para generar esto es casi idéntico al utilizado para producir la Tabla 16.2, pero en su lugar reemplazamos *aline_flg* por *day_28_flg* (el resultado) para obtener la Tabla 16.3.

```
datby.28daymort <- split(dat, dat$day_28_flg)
reftablemort <- produce.table1(datby.28daymort[[1]])
tab3 <- cbind(produce.table1(datby.28daymort[[1]], labels = attr(reftablemort,
    "labels")), produce.table1(datby.28daymort[[2]], labels = attr(reftablemort,
    "labels")))
colnames(tab3) <- paste0("Average (SD), or N (%)", c(",Alive", ",Dead"))
kable(tab3, caption = "Patient characteristics stratified by 28 day mortality")
```

Como puede verse en la Tabla 16.3, aquellos pacientes que murieron dentro de los 28 días difirieron de muchas formas de aquellos que no lo hicieron. Aquellos que murieron tuvieron puntajes SAPS y SOFA mayores, eran en promedio mayores y tenían perfiles de comorbilidades diferentes.

16.5.3 Análisis de regresión logística

En la Tabla 16.3 vemos que de los 984 pacientes que requirieron un CAI, 170 (17,2%) murieron dentro de los 28 días, mientras que 113 de los 192

(14,2%) murieron en el grupo no-CAI.

Tabla 16.3 Características de los pacientes estratificada según la mortalidad a 28 días

	Promedio (DS), oN (%) vivo	Promedio (DS), o N (%) muerto
aline_flg==0	814 (54.5%)	170 (60.1%)
icu_los_day	3.2 (3.2)	4 (4)
hospital_los_day	8.4 (8.4)	6.4 (6.4)
Age	50.8 (20.1)	73.3 (15.3)
gender_num==1	886 (59.4%) (faltantes: 1)	139 (49.1%)
weight_first	81.4 (22.7) (faltantes: 77)	72.4 (19.9) (faltantes: 33)
FMI	28.2 (8.3) (faltantes: 392)	26 (7.2) (faltantes: 74)
sapsi_first	13.6 (3.9) (faltantes: 51)	17.3 (3.8) (faltantes: 34)
sofa_irst	5.7 (2.3) (faltantes: 3)	6.6 (2.4) (faltantes: 3)
service_unit==SICU	829 (55.5%)	153 (54.1%)
service_num==1	829 (55.5%)	153 (54.1%)
day_icu_intime==Saturday	235 (15.7%)	43 (15.2%)
day_icu_intime_num	4 (2)	4.1 (2)
hour_icu_intime	10.5 (7.9)	11 (8)
hosp_exp_flg==0	1490 (99.8%)	42 (14.8%)
icu_exp_flg==0	1493 (100%)	283 (100%)
day_28_flg==0	1493 (100%)	283 (100%)
mort_day_censored	729.6 (331.4)	6.1 (6.4)
censor_flg==1	1279 (85.7%)	0 (0%)
sepsis_flg==0	1493 (100%)	283 (100%)

chf_flg==0	1348 (90.3%)	215 (76%)
afib_flg==0	1372 (91.9%)	197 (69.6%)
renal_flg==0	1447 (96.9%)	269 (95.1%)
liver_flg==0	1413 (94.6%)	264 (93.3%)
copd_flg==0	1377 (92.2%)	242 (85.5%)
cad_flg==0	1403 (94%)	250 (88.3%)
stroke_flg==0	1386 (92.8%)	168 (59.4%)
mal_flg==0	1294 (86.7%)	226 (79.9%)
resp_flg==0	1056 (70.7%)	155 (54.8%)
map_1st	88.2 (17.5)	88.3 (17.9)
hr_st	88.3 (18.4)	85.8 (20.6)
temp_1st	97.8 (4.6) (faltantes: 1)	97.7 (4.5) (faltantes: 2)
spo2_1st	98.6 (5)	97.8 (7.6)
abg_count	5.7 (7.7)	7.5 (12.5)
wbc_first	12.2 (6.4) (faltantes: 6)	12.7 (7.5) (faltantes: 2)
hgb_first	12.7 (2.2) (faltantes: 6)	11.9 (2.1) (faltantes: 2)
platelet_first	246.8 (97.3) (faltantes: 6)	242.1 (112.6) (faltantes: 2)
sodium_first	139.6 (4.6) (faltantes: 4)	139.1 (5.4) (faltantes: 1)
potassium_first	4.1 (0.8) (faltantes: 4)	4.2 (0.9) (faltantes: 1)
tco2_first	24.3 (4.8) (faltantes: 4)	25 (5.8) (faltantes: 1)
chloride_first	104.1 (5.6) (faltantes: 4)	102.6 (6.4) (faltantes: 1)
bun_first	18 (12.9) (faltantes: 4)	26.2 (19) (faltantes: 1)
creatinine_first	1.1 (1.1) (faltantes: 5)	1.2 (0.9) (faltantes: 1)

po2_first	231.3 (146.3) (faltantes: 153)	207.9 (135.8) (faltantes: 33)
pco2_first	43.3 (12.9) (faltantes: 153)	43.8 (18.6) (faltantes: 33)
iv_day_1	1694.2 (1709.5) (faltantes: 127)	1258 (1449.4) (faltantes: 16)

En un análisis univariado podemos evaluar si la menor tasa de mortalidad es estadísticamente significativa ajustando en una regresión logística con la única covariable *aline_flg*:

```
uvr.glm <- glm(day_28_flg ~ aline_flg,data=dat,family="binomial")
exp(uvr.glm$coef[-1])
```

```
## aline_flg1
##   1.254919
```

```
exp(confint(uvr.glm)[-1,]);
```

```
##     2.5 %    97.5 %
## 0.9701035 1.6285165
```

Aquellos que recibieron un CAI tuvieron un incremento de más del 25% en las probabilidades de morir a los 28 días en comparación con aquellos que no recibieron un CAI. El intervalo de confianza incluye 1, por lo que esperaríamos que el valor p sea >0,05. Al ejecutar la función *summary* vemos que este es el caso.

```
##              Estimate Std. Error    z value     Pr(>|z|)
## (Intercept) -1.7932333  0.1015988 -17.650149 1.014880e-69
## aline_flg1   0.2270714  0.1320347   1.719786 8.547142e-02
```

En efecto el valor de p para *aline_flg* es aproximadamente 0,09. Como vimos en la Tabla 16.2 probablemente existen varias covariables importantes que difirieron entre quienes recibieron y quienes no recibieron un CAI. Estos pueden funcionar como confundidores y la asociación que observamos en el análisis univariado puede ser más fuerte, inexistente o contraria (es decir, teniendo el CAI menores tasas de mortalidad) dependiendo de la situación. Nuestro siguiente paso sería realizar el ajuste para estos confundidores. Este es un ejercicio dentro de lo que se conoce como construcción del modelo y existen varias formas en las que la gente lo realiza de acuerdo a la literatura.

Un abordaje común es ajustar todos los modelos univariados (una covariable a la vez, de la misma manera que lo hicimos para *aline_flg*, pero de manera separada para cada variable y sin *aline_flg*) y realizar una prueba de hipótesis en cada modelo. Cualquier variable que tuviera

significancia estadística en los modelos univariados sería incluida luego en un modelo multivariado. Otro abordaje comienza con el modelo que acabamos de ajustar (*uvr.glm* que solo tiene *aline_flg* como covariable) y luego agrega variables de a una a la vez. Este abordaje suele ser llamado selección *step-wise forward*. Haremos la elección de hacer la selección *stepwise backwards*, que tal como suena, sigue la dirección contraria de la selección *stepwise forward*. La elección del modelo es una tarea desafiante en el análisis de datos y existen muchos otros métodos [18] que no sería posible describir en detalle aquí. Como filosofía general, es importante delinear y describir el proceso por el cual realiza la selección de modelo antes de hacerlo y atenerse al proceso.

En nuestro procedimiento de eliminación *stepwise backwards*, ajustaremos un modelo conteniendo CAI (*aline_flg*), edad (*age*), género (*gender_num*), severidad de enfermedad (*sapsi_first* y *sofa_first*), tipo de servicio (*service_unit*) y comorbilidades (*chf_flg, afib_flg, renal_flg, liver_flg, copd_flg, cad_flg, stroke_flg, mal_flg y resp_flg*). Esto suele llamarse el modelo completo y se ajusta debajo (*mva.full.glm*). Del modelo completo continuaremos eliminando una variable a la vez, hasta que quedemos con un modelo que contenga únicamente covariables estadísticamente significativas. Debido a que *aline_flg* es la covariable de interés, esta permanecerá en el modelo más allá de su significancia estadística. A cada paso necesitaremos definir un criterio para elegir qué variable eliminaremos. Existen varias formas de hacer esto, pero una forma en la que podemos tomar esta decisión es realizando una prueba de hipótesis para cada covariable y eligiendo eliminar la covariable con el mayor valor de p, salvo que todos los valores p sean <0,05 ó el mayor valor p sea *aline_flg*, en cuyo caso nos detendríamos o eliminaríamos el siguiente valor de p más grande respectivamente.

La mayoría de las covariables son binarias o categóricas por naturaleza, por lo que ya las hemos convertido en factores. Los puntajes de severidad de enfermedad (SAPS y SOFA) son continuos. Podríamos agregarlos como hicimos con la edad, pero esto supone una tendencia lineal en la probabilidad de mortalidad a medida que los puntajes cambian. Esto puede o no ser apropiado (vea Fig. 16.8). En efecto, cuando graficamos el log odds de la mortalidad a 28 días según el puntaje SOFA notamos que, si bien el log

odds de mortalidad incrementa a medida que incrementa el puntaje SOFA, la relación puede no ser lineal (Fig. 16.8).

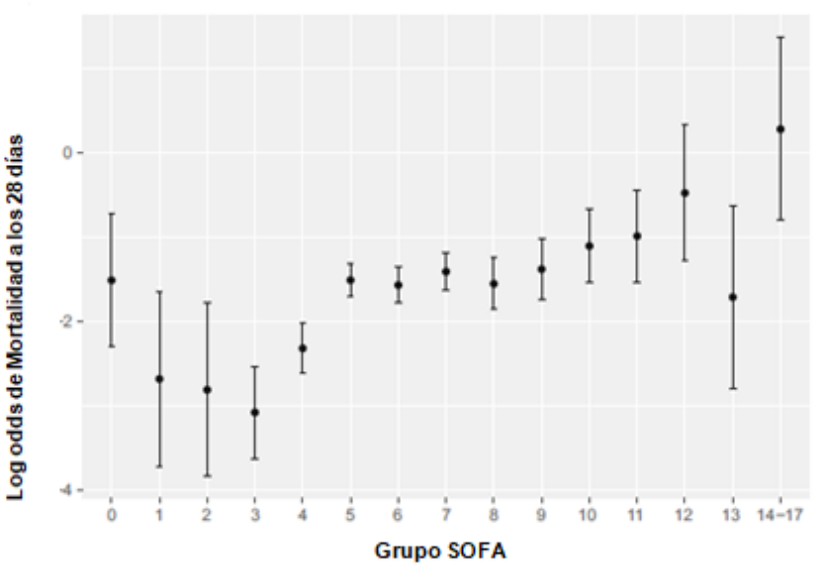

Fig. 16.8 Gráfico del log-odds de mortalidad para cada uno de los grupos SOFA. Las barras de error representan los intervalos de confianza 95% para los log-odds

Lo que puede hacerse en esta situación es transformar una covariable continua en una discreta. Una forma rápida de hacer esto es utilizando la función *cut2* en el paquete *Hmisc*[2]

Aplicar *cut2* (*sofa_first, g=5*) convierte la variable *sofa_first* en 5 grupos aproximadamente iguales según el score SOFA. Para mayor ilustración, SOFA se descompone en los siguientes grupos según el puntaje SOFA:

```
library(Hmisc)
table(cut2(dat$sofa_first,g=5))
```

```
##
## [0, 5)      5      6 [7, 9) [9,17]
##    523    346    294    391    216
```

con grupos no tan iguales, debido a la naturaleza discreta de SOFA con la que se comienza. Trataremos tanto SOFA como SAPS de esta manera para evitar cualquier especificación errónea que pueda ocurrir como resultado de asumir una relación lineal.

Regresando al ajuste del modelo completo, utilizaremos estos nuevos puntajes de severidad de enfermedad junto con el resto de las covariables

que identificamos para incluir en el modelo completo.

```
mva.full.glm <- glm(day_28_flg ~ aline_flg + age + gender_num + cut2(sapsi_first,
    g = 5) + cut2(sofa_first, g = 5) + service_unit + chf_flg + afib_flg + renal_flg +
    liver_flg + copd_flg + cad_flg + stroke_flg + mal_flg + resp_flg, data = dat,
    family = "binomial")
summary(mva.full.glm)
```

```
##
## Call:
## glm(formula = day_28_flg ~ aline_flg + age + gender_num + cut2(sapsi_first,
##     g = 5) + cut2(sofa_first, g = 5) + service_unit + chf_flg +
##     afib_flg + renal_flg + liver_flg + copd_flg + cad_flg + stroke_flg +
##     mal_flg + resp_flg, family = "binomial", data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2912  -0.4710  -0.2330  -0.1104   2.9640
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -7.61471    0.86262  -8.827  < 2e-16 ***
## aline_flg1                      0.01085    0.20443   0.053 0.957679
## age                             0.04020    0.00627   6.412 1.44e-10 ***
## gender_num1                     0.16214    0.17296   0.937 0.348527
## cut2(sapsi_first, g = 5)[12,14) 0.36961    0.40348   0.916 0.359637
## cut2(sapsi_first, g = 5)[14,16) 1.01794    0.36214   2.811 0.004940 **
## cut2(sapsi_first, g = 5)[16,19) 0.92803    0.36794   2.522 0.011662 *
## cut2(sapsi_first, g = 5)[19,32] 1.77615    0.37446   4.743 2.10e-06 ***
## cut2(sofa_first, g = 5)5        0.49761    0.30267   1.644 0.100159
## cut2(sofa_first, g = 5)6        0.58530    0.30300   1.932 0.053396 .
## cut2(sofa_first, g = 5)[7, 9)   0.68011    0.29439   2.310 0.020876 *
## cut2(sofa_first, g = 5)[9,17]   0.75134    0.34062   2.206 0.027397 *
## service_unitMICU                1.08086    0.67839   1.593 0.111100
## service_unitSICU                0.64257    0.67144   0.957 0.338562
## chf_flg1                        0.23350    0.23381   0.999 0.317962
## afib_flg1                       0.52408    0.21122   2.481 0.013092 *
## renal_flg1                     -0.76796    0.40904  -1.877 0.060452 .
## liver_flg1                      0.47238    0.34032   1.388 0.165125
## copd_flg1                       0.23440    0.24631   0.952 0.341287
## cad_flg1                       -0.25674    0.28823  -0.891 0.373065
## stroke_flg1                     2.04301    0.21966   9.301  < 2e-16 ***
## mal_flg1                        0.49319    0.20897   2.360 0.018274 *
## resp_flg1                       0.69330    0.19166   3.617 0.000298 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1400.58  on 1683  degrees of freedom
## Residual deviance:  954.39  on 1661  degrees of freedom
##   (92 observations deleted due to missingness)
## AIC: 1000.4
##
## Number of Fisher Scoring iterations: 6
```

El resultado *summary* muestra que algunas de las covariables son muy significativas estadísticamente, mientras que otras pueden ser prescindibles.

Idealmente querríamos el modelo más simple que pueda explicar la mayor parte posible de la variación en el resultado. Intentaremos remover la primer covariable según el procedimiento explicado anteriormente. Para cada una de las variables que consideramos remover podríamos ajustar un modelo de regresión logística sin la covariables y luego probarlo contra el modelo actual.

R tiene una función útil que automatiza este proceso llamada *drop1*. Ejecutamos *drop1* en nuestro objeto de regresión logística (*mva.full.glm*) y el tipo de prueba que queremos hacer. Si recuerda de la sección de regresión logística, utilizamos test = "*Chisq*" y esto es lo que utilizaremos para ejecutar la función *drop1*.

```
drop1(mva.full.glm,test="Chisq")
```

```
## Single term deletions
##
## Model:
## day_28_flg ~ aline_flg + age + gender_num + cut2(sapsi_first,
##     g = 5) + cut2(sofa_first, g = 5) + service_unit + chf_flg +
##     afib_flg + renal_flg + liver_flg + copd_flg + cad_flg + stroke_flg +
##     mal_flg + resp_flg
##                         Df Deviance     AIC    LRT  Pr(>Chi)
## <none>                       954.39 1000.39
## aline_flg                1   954.39  998.39  0.003 0.9576771
## age                      1  1000.60 1044.60 46.210 1.063e-11 ***
## gender_num               1   955.27  999.27  0.883 0.3475044
## cut2(sapsi_first, g = 5) 4   989.69 1027.69 35.304 4.023e-07 ***
## cut2(sofa_first, g = 5)  4   960.95  998.95  6.558 0.1611514
## service_unit             2   960.11 1002.11  5.716 0.0573820 .
## chf_flg                  1   955.38  999.38  0.990 0.3196816
## afib_flg                 1   960.47 1004.47  6.080 0.0136708 *
## renal_flg                1   958.20 1002.20  3.814 0.0508182 .
## liver_flg                1   956.23 1000.23  1.839 0.1750410
## copd_flg                 1   955.28  999.28  0.893 0.3445691
## cad_flg                  1   955.20  999.20  0.811 0.3678829
## stroke_flg               1  1045.22 1089.22 90.831 < 2.2e-16 ***
## mal_flg                  1   959.80 1003.80  5.410 0.0200201 *
## resp_flg                 1   967.57 1011.57 13.177 0.0002834 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como puede ver del resultado, cada covariable queda listada junto con su valor p *(Pr (>Chi))*. Cada fila representa una prueba de hipótesis siendo el mayor (modelo alternativo) el modelo completo (*mva.full.glm*) y siendo cada null el modelo completo sin la covariable de la fila. Los valores de p de la lista deberían ser iguales a aquellos que se obtendrían haciendo esta misma prueba con *anova*. Como puede verse de los valore de p listados, *aline_flg* tiene el mayor valor de p, pero

habíamos estipulado en nuestro plan de selección de modelo que conservaríamos esta covariable ya que es la covariable de interés. Procedemos luego al siguiente mayor valor de p que corresponde a la variable *cad_flg* (enfermedad de arteria coronaria). Actualizaremos nuestro modelo y repetiremos el paso de eliminación backwards en el modelo actualizado. Podríamos cortar y pegar el comando *mva.ful.glm* y remover + *cad_flg*, pero una forma más fácil y menos propensa al error es utilizar el comando *update*. La función *update* puede tomar un objeto *glm* o *lm* y alterar una de las covariables. Para hacer una eliminación *backwards*, el segundo argumento es. ~. – variable. La parte. ~. indica dejar el resultado y el resto de las variables tal cual y la parte – variable indica ajustar el modelo sin la variable llamada variable. Por lo tanto, para ajustar un modelo nuevo a partir del modelo completo pero sin la variable *cad_flg* ejecutaríamos:

```
mva.tmp.glm <- update(mva.full.glm, .~. - cad_flg)
```

Luego repetimos el paso *drop1*:

```
drop1(mva.tmp.glm,test="Chisq")
```

```
## Single term deletions
##
## Model:
## day_28_flg ~ aline_flg + age + gender_num + cut2(sapsi_first,
##     g = 5) + cut2(sofa_first, g = 5) + service_unit + chf_flg +
##     afib_flg + renal_flg + liver_flg + copd_flg + stroke_flg +
##     mal_flg + resp_flg
##                          Df Deviance     AIC    LRT  Pr(>Chi)
## <none>                        955.20  999.20
## aline_flg                 1   955.20  997.20  0.002 0.9674503
## age                       1  1000.92 1042.92 45.715 1.368e-11 ***
## gender_num                1   955.98  997.98  0.784 0.3760520
## cut2(sapsi_first, g = 5)  4   990.38 1026.38 35.180 4.266e-07 ***
## cut2(sofa_first, g = 5)   4   961.75  997.75  6.552 0.1615399
## service_unit              2   960.98 1000.98  5.782 0.0555160 .
## chf_flg                   1   955.92  997.92  0.719 0.3965762
## afib_flg                  1   961.32 1003.32  6.115 0.0134006 *
## renal_flg                 1   959.97 1001.97  4.774 0.0288966 *
## liver_flg                 1   957.06  999.06  1.862 0.1723427
## copd_flg                  1   956.02  998.02  0.824 0.3640764
## stroke_flg                1  1045.73 1087.73 90.526 < 2.2e-16 ***
## mal_flg                   1   960.64 1002.64  5.435 0.0197326 *
## resp_flg                  1   968.84 1010.84 13.638 0.0002217 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

y vemos que *aline_flg* sigue teniendo el mayor valor p, pero *chf_flag* tiene el segundo mayor, por lo que elegiríamos removerla a

continuación. Para actualizar el nuevo modelo y correr otro paso de eliminación, deberíamos ejecutar:

```
mva.tmp.glm2 <- update(mva.tmp.glm, .~. - chf_flg)
drop1(mva.tmp.glm2,test="Chisq")
```

```
## Single term deletions
##
## Model:
## day_28_flg ~ aline_flg + age + gender_num + cut2(sapsi_first,
##     g = 5) + cut2(sofa_first, g = 5) + service_unit + afib_flg +
##     renal_flg + liver_flg + copd_flg + stroke_flg + mal_flg +
##     resp_flg
##                           Df Deviance     AIC    LRT  Pr(>Chi)
## <none>                         955.92  997.92
## aline_flg                  1   955.93  995.93  0.016 0.9003547
## age                        1  1005.90 1045.90 49.976 1.556e-12 ***
## gender_num                 1   956.65  996.65  0.734 0.3916088
## cut2(sapsi_first, g = 5)   4   991.04 1025.04 35.121 4.387e-07 ***
## cut2(sofa_first, g = 5)    4   962.39  996.39  6.467 0.1669071
## service_unit               2   962.45 1000.45  6.529 0.0382253 *
## afib_flg                   1   963.01 1003.01  7.090 0.0077512 **
## renal_flg                  1   960.24 1000.24  4.321 0.0376445 *
## liver_flg                  1   957.70  997.70  1.780 0.1821692
## copd_flg                   1   956.95  996.95  1.035 0.3088774
## stroke_flg                 1  1045.73 1085.73 89.808 < 2.2e-16 ***
## mal_flg                    1   961.15 1001.15  5.231 0.0221921 *
## resp_flg                   1   970.13 1010.13 14.214 0.0001632 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

donde nuevamente *aline_flg* tiene el mayor valor de p y *gender_num* tiene el segundo mayor. Continuamos eliminando *gender_num*, *copd_flg*, *liver_flg*, *cut2 (sofa_first, g=5)*, *renal_flg* y *service_unit* en ese orden (los resultados se omitieron). La tabla producida por *drop1* a partir de nuestro modelo final es la siguiente:

```
drop1(mva.tmp.glm8,test="Chisq")
```

```
## Single term deletions
##
## Model:
## day_28_flg ~ aline_flg + age + cut2(sapsi_first, g = 5) + afib_flg +
##     stroke_flg + mal_flg + resp_flg
##                          Df Deviance    AIC    LRT  Pr(>Chi)
## <none>                        989.10 1011.1
## aline_flg                 1   989.10 1009.1  0.001  0.977380
## age                       1  1037.65 1057.7 48.556 3.209e-12 ***
## cut2(sapsi_first, g = 5)  4  1037.88 1051.9 48.788 6.465e-10 ***
## afib_flg                  1   995.60 1015.6  6.502  0.010777 *
## stroke_flg                1  1078.58 1098.6 89.485 < 2.2e-16 ***
## mal_flg                   1   997.37 1017.4  8.274  0.004021 **
## resp_flg                  1  1022.30 1042.3 33.200 8.317e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Todas las variables son estadísticamente significativas al nivel de significación 0,05. Mirando el resultado de *summary* podemos ver que *aline_flg* no es estadísticamente significativa (p = 0,98), pero todos los otros términos son estadísticamente significativos, con excepción de *cut2* (*sapsi_first*, g=5) [12, 14), lo que sugiere que el segundo grupo SAPS más bajo puede no ser diferente en forma estadísticamente significativo con el grupo de referencia (el grupo SAPS más bajo).

```
mva.final.glm <- mva.tmp.glm8;
summary(mva.final.glm)
```

```
##
## Call:
## glm(formula = day_28_flg ~ aline_flg + age + cut2(sapsi_first,
##     g = 5) + afib_flg + stroke_flg + mal_flg + resp_flg, family = "binomial",
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3025  -0.4928  -0.2433  -0.1289   3.1103
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -6.081944   0.445625 -13.648  < 2e-16 ***
## aline_flg1                      0.005078   0.179090   0.028  0.97738
## age                             0.037205   0.005644   6.592 4.33e-11 ***
## cut2(sapsi_first, g = 5)[12,14) 0.302084   0.391502   0.772  0.44035
## cut2(sapsi_first, g = 5)[14,16) 1.127302   0.344670   3.271  0.00107 **
## cut2(sapsi_first, g = 5)[16,19) 1.030901   0.347842   2.964  0.00304 **
## cut2(sapsi_first, g = 5)[19,32] 1.883738   0.347311   5.424 5.84e-08 ***
## afib_flg1                       0.522664   0.203485   2.569  0.01021 *
## stroke_flg1                     1.870553   0.199980   9.354  < 2e-16 ***
## mal_flg1                        0.592458   0.202297   2.929  0.00340 **
## resp_flg1                       0.976808   0.171629   5.691 1.26e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1413.4  on 1690  degrees of freedom
## Residual deviance:  989.1  on 1680  degrees of freedom
##   (85 observations deleted due to missingness)
## AIC: 1011.1
##
## Number of Fisher Scoring iterations: 6
```

Este sería nuestro modelo final y nos daría una tabla similar a la Tabla 16.4. Dado que el efecto de CAI era de particular interés, lo resaltaremos diciendo que no se asocia con mortalidad a los 28 días con un odds ratio de 1,01 (IC 95%: 0,71-1,43, p=0,98). Podemos concluir que luego de realizar el ajuste para los otros potenciales confundidores hallados en la Tabla 16.4 no encontramos ningún impacto estadísticamente significativo del uso de CAI sobre la mortalidad.

Tabla 16.4 Regresión logística multivariable para resultado de mortalidad a 28 días (modelo final)

Covariable	OR Ajustado	Limite Inferior de IC 95%	Limite Superior de IC 95%	Valor p

CAI	1.01	0.71	1.43	0.977
Edad (incremento por año)	1.04	1.03	1.05	<0.001
SAPSI (12-14) (Referencia SAPSI <2)	1.35	0.63	2.97	0.440
SAPSI (14-16)	3.09	1.61	6.28	0.001
SAPSI (16-19)	2.80	1.45	5.74	0.003
SAPSI (19-32)	6.58	3.42	13.46	<0.001
Fibrilación Articular	1.69	1.13	2.51	0.010
Accidente Cerebro Vascular	6.49	4.40	9.64	<0.001
Enfermedad Maligna	1.81	1.21	2.68	0.003
Enfermedad respiratoria no EPOC	2.66	1.90	3.73	<0.001

16.4 Conclusión y resumen

Esta breve revisión de las técnicas de modelado para datos de salud le ha provisto las bases para realizar los tipos de análisis más comunes en estudios en salud. Hemos mencionado cuán importante es tener un objetivo de estudio claro antes de conducir análisis de datos, ya que identifica todos los aspectos importantes que Ud. necesita para planificar y ejecutar su análisis. En particular, la identificación del resultado debería permitirle determinar qué metodología de análisis sería más apropiada. A menudo encontrará que estará utilizando múltiples técnicas de análisis para diferentes objetivos de estudio dentro del mismo estudio. La Tabla 16.5 resume algunos de los aspectos más importantes para cada abordaje de análisis

Afortunadamente, el marco de referencia de R para realizar estos análisis es bastante similar a lo largo de los diferentes tipos de técnicas y este marco de referencia podrá extenderse de manera general a otros modelos más complejos (incluyendo algoritmos de aprendizaje automático) y estructuras de datos (incluyendo datos dependientes/correlacionados como datos longitudinales).

Tabla 16.5 Resumen de diferentes métodos

	Regresión Lineal	Regresión Logística	Modelos de riesgo proporcional de COX
Tipos de datos del resultado	Continuo	Binario	Tiempo al evento (posiblemente censurado)
Análisis preliminar útil	Gráfico de dispersión	Contingencia y tablas 2x2	Estimación de la función de sobrevida de Kaplan-Meier
Presentación del Output	Coeficiente	Odds Ratio	Hazard Ratio
Output deR	Coeficiente	Log Odds Ratio	Relación Log de Log Hazard Ratio
Presentación Interpretación	Una estimación del cambio esperado en el resultado por unidad de aumento en la covariable, manteniendo todas las demás covariables constantes	Una estimación del cambio del odds del resultado por unidad de aumento en la covariable, manteniendo todas las demás covariables constantes	Una estimación del cambio en el Riesgo delos resultados por unidad de aumento en la covariable, manteniendo todas las demás covariables constantes

Hemos resaltado algunas áreas de interés a las que debe prestarse especial atención, incluyendo datos faltantes, colinealidad, especificaciones erróneas de modelos y outliers. Algunos de estos puntos serán revisados con más detalle en el Cap. 17.

Nota: Las imágenes de este capítulo se encuentran disponibles a color en el ebook; disponible en www.hardineros.com

Referencias

1. Hsu DJ, Feng M, Kothari R, Zhou H, Chen KP, Celi LA (2015) The association betweenindwelling arterial catheters and mortality in hemodynamically stable patients with respiratoryfailure: a propensity score analysis. CHEST J 148 (6): 1470-1476.
2. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng C-K, StanleyHE (2000) Physiobank, physiotoolkit, and physionet componentsof a new research resource for complex physiologic signals. Circulation 101 (23): e215-e220.
3. Indwelling arterial catheter clinical data from the MIMIC II database (2016) Disponible en http://physionet.org/physiobank/database/mimic2-iaccd/. [Consultado 02 de Junio2016].
4. Friedman J, Hastie T, Tibshirani R (2009) The elements of statistical learning: data mining, inference, and prediction. Springer series in statistics.
5. James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning, vol112. Springer, Berlin.
6. Harrell F (2015) Regressión modeling strategies: with applications to linear models, logisticand ordinal regressión, and survival analysis. Springer, Berlin.
7. Venables WN, Ripley BD (2013) Modern applied statistics with S-plus. Springer Science &Business Media.
8. Weisberg S (2005) Applied linear regressión, vol 528. Wiley, New York.
9. Diggle P, Heagerty P, Liang KY, Zeger S (2013) Analysis of longitudinal data. OUP Oxford.
10. McCullagh P, Nelder JA (1989) Generalized linear models, vol 37. CRC press, Boca Raton.
11. Hosmer DW, Lemeshow S (2004) Applied logistic regressión. Wiley, New York.
12. Kleinbaum DG, Klein M (2006) Survival analysis: a self-learning text. Springer Science &Business Media.
13. Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J AmStat Assoc 53 (282): 457-481.
14. CoxDR (1972) Regressión models and life-tables. JRStat Soc SerB (Methodol) 34 (2): 187-220.
15. Collett D (2015) Modelling survival data in medical research. CRC press, Boca Raton.

16. Kalbfleisch JD, Prentice RL (2011) The statistical analysis of failure time data, vol 360. Wiley, New York.
17. Therneau TM, Grambsch PM (2000) Modeling survival data: extending the Cox model. Springer Science & Business Media.
18. Dash M, Liu H (1997) Feature selection for classification. Intel Data Anal 1 (3): 131-156.

[1] Excede los objetivos de este texto la definición de “optimo”. Para aquellos interesados, estamos tratando de encontrar los valores de B_0 y B_1 que minimizan la distancia al cuadrado entre las lineas de ajuste y los puntos observados. Esta cantidad se conoce como la suma de errores cuadrados; o como la media del error cuadrado cuando se divide por el número de observaciones.

[2] Ud podrá necesitar instalar Hmisc, lo cual puede hacerse ejecutando install. packages (Hmisc) de la consola de comandos R.

Capítulo 17

ANÁLISIS DE SENSIBILIDAD Y VALIDACIÓN DEL MODELO

Justin D. Salciccioli, Yves Crutain,
Matthieu Komorowski y Dominic C. Marshall

Objetivos de aprendizaje

- Comprender que todos los modelos tienen limitaciones implícitas para la generalización.
- Entender los supuestos para hacer inferencias causales a partir de los datos disponibles.
- Controlar el ajuste del modelo y su desempeño.

17.1 Introducción

Imagine que Ud. ha finalizado el primer análisis de su actual investigación y que pudo rechazar la hipótesis nula. Aún luego de haber aplicado los métodos seleccionados y de haber desarrollado modelos robustos, pueden persistir algunas dudas. "¿Qué tanta confianza tiene usted en los resultados obtenidos? ¿Cuánto podrían cambiar los mismos si sus datos crudos fueran ligeramente incorrectos? ¿Tendría esto un impacto mínimo en sus resultados? ¿O se obtendría un resultado completamente diferente?" La inferencia causal está a menudo limitada por los supuestos realizados en el diseño y análisis del estudio, y esto tiene más peso aún cuando se trabaja con datos observacionales en salud. Una recomendación importante para cualquier investigador es evitar la dependencia de cualquier abordaje analítico único para evaluar la hipótesis en estudio y como tal, el siguiente paso crítico será probar los supuestos asumidos en el análisis.

El análisis de sensibilidad y la validación del modelo están relacionados entre sí por el hecho de que ambos constituyen una tentativa de evaluar la adecuación de las especificaciones de un modelo y de apreciar la fuerza de las conclusiones extraídas a partir del mismo. Si bien la validación del modelo es útil para evaluar el ajuste del mismo en un set de datos específico, el análisis de sensibilidad es particularmente apropiado para brindar confianza en los resultados del análisis primario y especialmente importante en situaciones donde es probable que el modelo sea usado en

futuras investigaciones o en la práctica clínica. Aquí discutiremos conceptos relacionados con la evaluación del ajuste del modelo y resumiremos a grandes rasgos los pasos relacionados con la validación cruzada y la validación externa aplicándolas directamente sobre el proyecto de la vía arterial invasiva. Discutiremos brevemente algunas de las razones comunes por las cuales los modelos fracasan en las pruebas de validación y las potenciales implicancias de esas fallas.

17.2 Parte 1 - Conceptos teóricos

17.2.1 Sesgo y varianza

En estadística y aprendizaje automático, la negociación (o dilema) sesgo-varianza es el problema de minimizar simultáneamente dos fuentes de error que impiden la generalización de los algoritmos supervisados más allá del set de datos de entrenamiento. Un modelo con sesgo alto falla para estimar los datos en forma precisa. Por ejemplo, una regresión lineal tendrá un sesgo alto cuando trata de modelar una relación cuadrática sin importar cómo se hayan fijado los parámetros (figura 17.1). La varianza, en el otro extremo, se relaciona con la estabilidad del modelo cuando se lo aplica en nuevos escenarios de entrenamiento. Decimos que un algoritmo que ajusta muy bien sobre los datos de entrenamiento pero que tiene un desempeño pobre en nuevo ejemplo (mostrando sobre-ajuste) tiene alta varianza.

Se detallan a continuación algunas estrategias frecuentes para tratar el sesgo y la varianza:

- Sesgo alto:
  - El agregado de variables (predictoras) tiende a reducir el sesgo con el costo de introducir más varianza.
  - El agregado de ejemplos o escenarios de entrenamiento no resolverá el sesgo porque el modelo desarrollado no será capaz de aproximar la función correcta.
- Alta varianza:
  - Reducir la complejidad puede ayudar a reducir la varianza. La reducción de la dimensionalidad y la selección de atributos son dos ejemplos de métodos para reducir la cantidad de parámetros en el modelo y así reducir la varianza (la selección de parámetros se discute más adelante).

- Un set de datos de entrenamiento más grande ayuda a disminuir la varianza.

Fig. 17.1 Comparación entre sesgo y varianza en el desarrollo de modelos.

17.2.2 Herramientas habituales de evaluación

Existe una variedad de técnicas estadísticas para evaluar cuantitativamente el desempeño de los modelos estadísticos. Si bien estas técnicas son importantes, exceden el alcance de este texto. De todas maneras, mencionaremos brevemente dos de las más comunes: el estadístico R^2 usado en regresiones y la Curva Operativa del Receptor, que llamaremos ROC de aquí en más (del inglés "Receiver Operating Characteristic curve") para clasificación binaria (en resultados dicotómicos).

El estadístico R^2 es una métrica de resumen que representa la proporción de la varianza total de la variable resultado que es capturada por el modelo. El R^2 tiene un rango que va de 0 a 1, donde valores cercanos a 0 reflejan situaciones en las cuales el modelo no resume adecuadamente la variación del resultado de interés, y valores cercanos a 1 indican que el modelo captura la casi totalidad de la variación del mismo. Valores altos de R^2 se interpretan como que una proporción alta de la varianza se explica por el modelo de regresión. En R, el R^2 se calcula usando la función de regresión lineal. Para ver un ejemplo de código en R para generar el R^2 consultar la función "R^2".

El valor R^2 es una medida general de la fuerza de asociación entre el modelo y el resultado, sin reflejar la contribución individual de ningún predictor independiente. Más aún, si bien podríamos esperar intuitivamente que hubiera una relación proporcional entre el número de variables predictoras y el R^2 general del modelo, en la práctica agregar predictores no necesariamente incrementa el R^2 en un nuevo set de datos.

Más aún, es posible que un predictor individual disminuya el R^2 dependiendo de cómo interactúa esta variable con otros parámetros del modelo.

Para el objetivo de esta discusión, esperamos que el lector se familiarice con el manejo computacional y la utilidad de los valores de sensibilidad y especificidad. En situaciones como el desarrollo de una nueva prueba diagnóstica, los investigadores pueden definir un umbral específico para clasificar una prueba como positiva. Cuando tratamos con un resultado dicotómico, la curva ROC brinda una descripción más completa de la capacidad del modelo para clasificar los resultados. La curva ROC es un método de uso frecuente para mostrar la relación entre la sensibilidad de un modelo de clasificación y su tasa de falsos positivos (1 – especificidad). El área bajo la curva ROC refleja la predicción estimada por el modelo y puede tomar valores en un rango de 0.5 a 1, donde aquellos valores cercanos a 0.5 no aportan más información que el azar, y los cercanos a 1 representan una predicción perfecta. Para ver ejemplos de curvas ROC en R, consultar la función “ROC” en el código del material adjunto. Para profundizar la lectura sobre curvas ROC consultar, por ejemplo, el artículo de Fawcett [1] (Fig. 17.2).

17.2.3 Análisis de sensibilidad

El análisis de sensibilidad incluye una serie de métodos para cuantificar la magnitud en la cual la incertidumbre en los datos de origen pueden afectar las estimaciones realizadas por el modelo. En otras palabras, el análisis de sensibilidad evalúa cuán “sensible” es el modelo a las fluctuaciones en los parámetros y los datos sobre los cuales se construye. Los resultados del análisis de sensibilidad pueden tener importantes implicancias en varias etapas del proceso de modelado, incluyendo en la identificación de errores en el modelo en sí, informando la calibración de los parámetros del modelo y explorando más ampliamente la relación entre los datos de origen y las estimaciones del modelo.

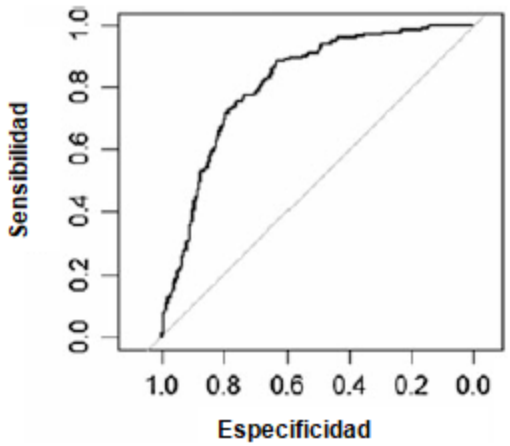

Fig. 17.2 Ejemplo de una curva ROC que puede usarse para evaluar la capacidad de un modelo para discriminar resultados dicotómicos.

Los principios de un análisis de sensibilidad son: (a) permitir al investigador cuantificar la incertidumbre en un modelo, (b) evaluar un modelo de interés usando un diseño experimental secundario, (c) calcular la sensibilidad global del modelo de interés usando el resultado del diseño experimental secundario.

La justificación de un análisis de sensibilidad es que un modelo siempre tendrá mejor desempeño (i.e., se "sobre-desempeñará") cuando se lo prueba en el set de datos del cual fue derivado. El análisis de subgrupos es una variante frecuente del análisis de sensibilidad [2].

17.2.4 Validación

Como fuera discutido en el capítulo 16, la validación del análisis de datos se usa para confirmar que el modelo de interés se desempeñará de manera similar cuando se lo aplique en condiciones de prueba diferentes. Así, es una responsabilidad fundamental del investigador evaluar la adecuación del ajuste del modelo a los datos. Esto puede lograrse con una variedad de abordajes metodológicos, cuyo tratamiento en profundidad escapa a los objetivos de este texto y puede consultarse en otras fuentes [3]. Aunque la discusión en detalle no será tratada aquí, la teoría general marca que un modelo debería seleccionarse basándose en dos principios: parsimonia del mismo y relevancia clínica. Una cantidad de abordajes predefinidos basados en algoritmos para la selección de modelos incluye: selección hacia

adelante (*Forward),* hacia atrás (*Backward)* y paso-a-paso (S*tepwise*), pero también Lasso y algoritmos genéticos, disponibles en los paquetes estadísticos más difundidos. Consultar el capítulo 16 para más información sobre selección de modelos.

La validación cruzada es una técnica que se usa para evaluar la capacidad predictiva de un modelo de regresión. Para más detalle se recomiendan otras fuentes [4]. El concepto de validación cruzada se basa en el principio de que un set de datos lo suficientemente grande puede dividirse en dos o más subgrupos (no necesariamente del mismo tamaño), usando el primero para derivar (entrenar) el modelo y el/los restante/s para las pruebas (testing) y validación. Para evitar la pérdida de información por usar un subgrupo y no la totalidad de los datos del set de entrenamiento, existe una variante llamada validación cruzada con k-réplicas (k-fold), que no será discutida aquí.

La validación externa se define como la prueba del modelo en una muestra de sujetos tomados de una población diferente de la cohorte original. La validación externa habitualmente es el abordaje más robusto para probar un modelo porque involucra: 1) la extracción de la máxima cantidad de información del set de datos inicial para derivar un modelo y 2) el uso de un set de datos completamente independiente para verificar la aplicabilidad del modelo de interés. Aunque la validación externa es un método esencial y es el más riguroso, poder contar con un set de datos con características similares y que además sea completamente independiente, es a menudo imposible para la mayoría de los investigadores. De todas maneras, con el notable incremento del volumen de datos relacionados con el cuidado de la salud que se capturan electrónicamente, es probable que asistamos a un incremento proporcional en la disponibilidad de los mismos para la realización de validaciones externas.

17.3 Caso de estudio: Ejemplos de validación y análisis de sensibilidad

En este caso de estudio se usa del set de datos del estudio CAI, que evaluó el impacto de colocar una vía arterial en pacientes con fallo respiratorio internados en la unidad de cuidados intensivos. Se llevaron a cabo tres análisis de sensibilidad diferentes:

1. Probar los efectos de variar los criterios de inclusión relacionados con el tiempo hasta el inicio de la ventilación mecánica y la mortalidad;
2. Probar los efectos de cambios en el nivel de calibración para la aplicación del score de propensión en la asociación de inserción de una vía arterial y la mortalidad.
3. Aplicar el test de bondad de ajuste de Hosmer-Lemeshow para evaluar el ajuste global de los datos al modelo de interés.

Se usaron los siguientes paquetes de R en CRAN: pareamiento por score de propensión [5], análisis de muestras de encuestas complejas [6], ggplot2 para generar gráficos [7], pROC para curvas ROC [8] y Twang para ponderar y analizar grupos no equivalentes [9].

17.3.1 Análisis 1: variando los criterios de inclusión del tiempo al inicio de la ventilación mecánica

El primer análisis de sensibilidad evalúa el efecto de variar el criterio de inclusión del tiempo hasta el inicio de la ventilación mecánica y la mortalidad. La ventilación mecánica es una de las intervenciones más comunes en la unidad de cuidados intensivos, y el tiempo transcurrido hasta el inicio de la misma puede servir como un subrogante de la severidad de la enfermedad crítica, dado que puede asumirse que pacientes con enfermedad más grave requerirán un inicio más temprano de la misma. Por eso, la ventilación mecánica junto con la colocación de un catéter arterial invasivo (CAI), otra intervención invasiva, pueden relacionarse con el resultado de interés, que es la mortalidad a los 28 días. En el documento de funciones de R adjunto, en la función “Cohort” (cohorte), se muestra un ejemplo de código en R para analizar la distribución entre grupos de acuerdo con su status ventilatorio (Fig. 17.3).

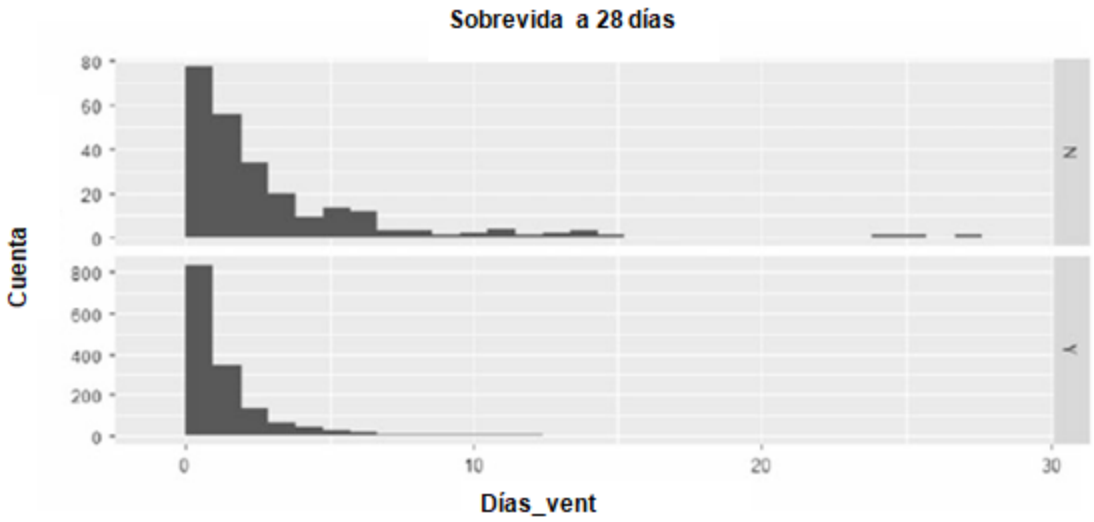

Fig. 17.3 Análisis simple de sensibilidad para comparar los resultados entre grupos variando los criterios de inclusión. La modificación de los criterios de inclusión de sujetos ingresados en el modelo es un análisis de sensibilidad muy usado.

A través de la modificación del tiempo de inicio de la ventilación mecánica podemos también obtener información importante del efecto de la exposición primaria sobre el resultado. Se muestra un ejemplo de código en R para este análisis en la función “Ventilation” (ventilación).

17.3.2 Análisis 2: Cambiando el nivel de calibración para el pareamiento por puntaje de propensión

El segundo análisis de sensibilidad evalúa el impacto de diferentes niveles de calibración para el puntaje de propensión en la asociación entre catéter arterial y mortalidad. En este estudio, el puntaje de propensión hace coincidir un sujeto que no tuvo catéter arterial con otro que sí lo tuvo. El algoritmo de pareamiento crea un par de sujetos independientes cuyos puntajes de propensión sean los más parecidos. Sin embargo, el investigador es responsable de fijar una diferencia máxima razonable en el puntaje de propensión, lo que permitirá al algoritmo generar una coincidencia posible; esta diferencia máxima razonable se conoce como “calibre”. La elección del calibre para el pareamiento por puntaje de propensión influirá directamente la negociación sesgo-varianza en la medida de que un calibre más amplio resultará en el pareamiento de sujetos menos parecidos con respecto a la probabilidad del tratamiento o

exposición. Se muestra un ejemplo de código en R para llevar a cabo un análisis de sensibilidad variando el nivel de calibre del puntaje de propensión en el documento de funciones R adjunto bajo la función "Caliper".

La figura 17.4 muestra el efecto de los ajustes en el nivel de calibre del puntaje de propensión. El modelo completo muestra un coeficiente más bajo debido a la presencia de variables adicionales.

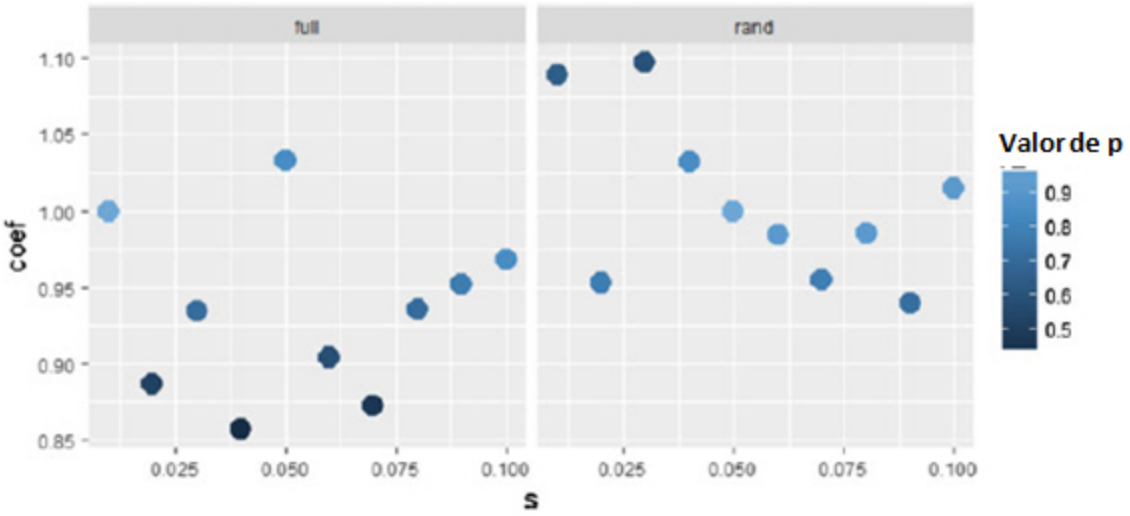

Fig. 17.4 Un análisis de sensibilidad para evaluar el efecto de modificar el nivel de calibre del puntaje de propensión

17.3.3 Análisis 3: prueba de Hosmer-Lemeshow

La prueba de bondad de ajuste de Hosmer-Lemeshow puede usarse para evaluar el ajuste global de los datos al modelo de interés [10]. Para esta prueba, los sujetos se agrupan de acuerdo con un percentil de riesgo (habitualmente, deciles) y se calcula el estadístico Chi^2 de Pearson para comparar la coincidencia dentro de cada grupo entre el valor observado y el riesgo calculado por el modelo. Se muestra un ejemplo de código en R para llevar a cabo este test en el documento adjunto bajo la función "HL".

17.3.4 Implicancias de un modelo "fallido"

En la situación favorable de un modelo robusto, cada análisis de sensibilidad y técnica de validación apoyan la idea de que el modelo sirve como un resumen apropiado de los datos. De todas maneras, en algunas

situaciones, el método de validación o el análisis de sensibilidad escogido revelan un ajuste inadecuado del modelo a los datos, de manera tal que el mismo fracasa en predecir el resultado de interés en forma precisa. Un modelo "fallido" puede ser el resultado de gran cantidad de factores diferentes. Ocasionalmente, es posible modificar el proceso de derivación del modelo para lograr un mejor ajuste a los datos. En aquellas situaciones en las que la modificación del modelo no permite lograr un nivel aceptable de error, es de buena práctica renunciar a la investigación y reiniciar la misma con una evaluación crítica de los supuestos asumidos *a priori* en una tentativa de desarrollar un modelo diferente.

17.4 Conclusión

El análisis de datos observacionales relacionados con la salud tiene la limitación inherente de conllevar confundidores no medidos. Luego del desarrollo del modelo y del análisis primario, un paso importante es confirmar el desempeño del mismo con una serie de pruebas confirmatorias para verificar su validez. Mientras que la validación puede usarse para controlar que el modelo brinde un ajuste apropiado a los datos y que es probable que se desempeñe de manera similar en otras cohortes, el análisis de sensibilidad puede usarse para interrogar sobre supuestos inherentes al análisis primario. Cuando se llevan a cabo adecuadamente, estos pasos adicionales contribuyen a mejorar la robustez del análisis global y ayudan al investigador a hacer inferencias valiosas a partir de datos observacionales.

Puntos clave

1. La validación y los análisis de sensibilidad evalúan la robustez de los supuestos del modelo y son un paso clave en el proceso de modelado;
2. El principio clave de estos análisis es variar los supuestos del modelo y observar cómo responde;
3. Cuando fallan la validación o los análisis de sensibilidad puede ser necesario comenzar con un nuevo modelo.

Nota: Las imágenes de este capítulo se encuentran disponibles a color en el ebook; disponible en www.hardineros.com

Apéndice: Código

El código usado en este capítulo se encuentra disponible en el repositorio GitHub de este libro: https://github.com/MIT-LCP/critical-data-book. En dicho sitio web se encuentra disponible información adicional sobre el código.

Referencias

1. Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27 (8): 861-874.
2. Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ (2004) Subgroupanalyses in randomized trials: risks of subgroup-specific analyses; power and sample size forthe interaction test. J Clin Epidemiol 57 (3): 229-236.
3. Pregibon D (1981) Logistic regressión diagnostics. Ann Stat 9 (4): 705-724.
4. Picard RR, Cook RD (1984) Cross-validation of regressión models. J Am Stat Assoc 79 (387): 575-583.
5. Sekhon JS (2011) Multivariate and propensity score matching software with automatedbalance optimization: the matching package for R. J Stat Softw 42 (i07).
6. Lumley T (2004) Analysis of complex survey samples. J Stat Softw 09 (i08).
7. Wickham H (2009) ggplot2. Springer, New York.
8. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M (2011) pROC: anopen-source package for R and S + to analyze and compare ROC curves. BMC Bioinf 12:77.
9. Ridgeway G, Mccaffrey D, Morral A, Burgette L, Griffin BA (2006) Twang: toolkit forweighting and analysis of nonequivalent groups. R package versión 1.4-9.3. In: R Foundationfor Statistical Computing, 2006. Disponible en http://www.cran.r-project.org. [Consultado 2015].

10. Hosmer DW, Lemesbow S (1980) Goodness of fit tests for the multiple logistic regressión model. Commun Stat Theory Methods 9 (10): 1043-1069.

PARTE III

ESTUDIO DE CASOS UTILIZANDO MIMIC

INTRODUCCIÓN

En esta sección se presentan doce estudios de caso de análisis secundario de historias clínicas electrónicas (HCEs). Los estudios de caso exponen un amplio rango de tópicos de investigación y de metodologías, resultando de esta manera interesantes para un gran número de investigadores. Se encuentran escritos, fundamentalmente, para el principiante, aunque el investigador experimentado también se beneficiará de las explicaciones detalladas ofrecidas por los expertos en la materia. Los estudios de caso brindan la oportunidad de involucrarse en profundidad en estudios de investigación de alto nivel, dado que se brindan tanto datos de acceso público como el código usado para el análisis. Esta sección no debe ser interpretada como una narración continua. Más bien, cada estudio de caso puede ser leído de forma independiente. De hecho, es aconsejable comenzar por aquellos que le resulten más interesantes. A continuación, se proporcionará una visión general de las áreas de investigación y las metodologías de los estudios de caso.

Los estudios de caso se ordenan según sus áreas de investigación. Los dos primeros casos se ocupan del análisis a nivel de sistema, comenzando con un análisis de las tendencias en la práctica clínica relacionadas con la ventilación mecánica (capítulo 18). Seguido de esto, se presenta una investigación sobre el efecto en la mortalidad en la internación de pacientes críticamente enfermos en UCIs no acordes a su patología, conocidas como UCI "*non target*" o "*boarding*" (por ejemplo, paciente con patología neurocrítica internado en una unidad coronaria) (Capítulo 19). Los siguientes tres casos se centran en la predicción de la mortalidad usando una gran variedad de predictores como por ejemplo, variables demográficas, signos vitales y resultados de laboratorio (Capítulos 20-22). Dos estudios de caso investigan la efectividad de una intervención clínica evaluando su efectividad clínica (Capítulo 23) y su costo-efectividad (Capítulo 24). Se presenta un estudio que evalúa la relación entre la presión arterial y el riesgo de insuficiencia renal aguda, ilustrando los hallazgos fisiológicos que pueden resultar del análisis de las HCEs (Capítulo 25). Luego

se presentan dos estudios de técnicas de monitoreo: una investigación sobre la estimación de la frecuencia respiratoria, un parámetro fisiológico clave, a partir de señales fisiológicas adquiridas en forma rutinaria (Capítulo 26); y un estudio detallado sobre el potencial de reducción de falsas alarmas usando técnicas de clasificación de aprendizaje automático (Capítulo 27). Para finalizar, dos estudios consideran aspectos particulares de la metodología de la investigación, focalizándose en la identificación de cohortes de pacientes (Capítulo 28) y en la selección de hiperparámetros a través de técnicas matemáticas (Capítulo 29).

A través del estudio de casos se muestra una gran cantidad de metodologías. Las técnicas de aprendizaje automático usadas incluyen: regresión, máquinas de vector soporte, árboles de decisión (Capítulo 21), clasificación de *random forest* (capítulo 27), modelos de Markov (Capítulo 24) y un algoritmo de Súper Aprendizaje para integrar múltiples técnicas (Capítulo 20). Otros abordajes analíticos incluyen análisis de variables instrumentales (Capítulo 19), pareamiento por puntaje de propensión (Capítulo 23), diseños de casos y controles y diseños con entrecruzamiento de casos (Capítulo 25), procesamiento de señales (Capítulos 26 y 27), y procesamiento de lenguaje natural (Capítulo 28).

El objetivo de esta sección es proveer a los lectores ejemplos de análisis secundario de la HCE para alentarlos en su propia investigación. Esperamos que la relevancia clínica de las investigaciones inspire a investigadores a explotar al máximo el potencial de las HCE en beneficio de futuros pacientes. La descripción detallada de la metodología empleada en los estudios pretende aportar un mejor entendimiento de los distintos matices del análisis de las HCE. Finalmente, una amplia gama de herramientas se encuentra disponible para guiar a los nuevos investigadores: los datos y el código usado para el análisis en esta sección son públicos.

Puede encontrarse más información sobre estas herramientas en el repositorio de GitHub: https://github.com/MIT-LCP/critical-data-book.

Capítulo 18

ANÁLISIS DE TENDENCIAS: EVOLUCIÓN DEL VOLUMEN CORRIENTE EN EL TIEMPO EN PACIENTES QUE RECIBEN VENTILACIÓN MECÁNICA INVASIVA

Anuj Metha, Franck Dernoncourt y Allan Walkey

Objetivos de aprendizaje

Aprender la importancia del análisis de tendencia:

- Para comprender los cambios epidemiológicos en la salud y la atención sanitaria brindada.
- Para evaluar la implementación de nueva evidencia en la práctica clínica.
- Comprender la efectividad en el mundo real de los descubrimientos (diseños de series temporales interrumpidas, diferencias en diferencias, regresión discontinua)

Aprender los métodos para realizar un análisis de tendencia:

- Test de Cochrane–Armitage para tendencias
- Diferencias entre regresión lineal y logística tomando al tiempo como variable independiente.

Identificar cambios en la población de estudio a través del tiempo en relación con las variables dependientes e independientes más importantes:

- Ajustes/Confundidores
- Interacción de covariables con el tiempo y con los resultados.

Redefinir la pregunta de investigación:

- Identificar limitaciones de los datos.

18.1 Introducción

La atención de la salud es un área dinámica que evoluciona continuamente en respuesta a los cambios epidemiológicos de las enfermedades, variables demográficas poblacionales y nuevos descubrimientos. Los cambios epidemiológicos en la prevalencia de las enfermedades y los resultados tienen implicancias importantes para determinar la distribución de los recursos sanitarios. Por ejemplo, identificar

tendencias que muestren un aumento en la utilización de ventilación mecánica podría sugerir que son necesarias más camas de cuidados intensivos, más enfermeros y médicos especialistas y más ventiladores mecánicos. Además, los cambios en los resultados sanitarios a través del tiempo permitirían comprender la adopción de nuevo conocimiento científico e identificar objetivos de mejora de la calidad donde la implementación de la evidencia haya sido lenta o donde no es posible dar cuenta en el "mundo real" de los resultados de investigaciones estrictamente controladas. El análisis de tendencias utiliza métodos estadísticos para cuantificar cambios que permitan entender mejor la evolución de la salud y la atención de la misma.

Para destacar los usos del análisis de tendencias, presentamos un estudio que evalúa cómo la evidencia científica que avala el tratamiento de una enfermedad determinada puede ser generalizada por los profesionales de la salud a otras condiciones en las cuales el tratamiento no fue evaluado. Investigamos la adopción de la evidencia que apoya utilizar bajo volumen corriente durante la ventilación mecánica en pacientes ingresados en una Unidad de Cuidados Intensivos (UCI) comparada con una Unidad Coronaria (UCO).

Los pacientes críticamente enfermos pueden desarrollar dificultad respiratoria grave y requerir asistencia a través de una máquina (ventilador) mediante un proceso llamado ventilación mecánica invasiva. Los pacientes pueden requerir ventilación mecánica invasiva por múltiples causas, como puede ser neumonía, asma e insuficiencia cardíaca.

En algunos casos, los pulmones presentan una inflamación masiva desencadenada por enfermedades sistémicas como infecciones, trauma o aspiración. Esta inflamación conduce a una acumulación de líquido en los pulmones (edema pulmonar), situación conocida con el nombre de Síndrome de Distress Respiratorio Agudo (SDRA).

El SDRA se define por cuatro criterios [1]:

1. Inicio agudo
2. Infiltrados bilaterales en la radiografía de tórax
3. No es causado por insuficiencia cardíaca (dado que la insuficiencia cardíaca también puede causar edema pulmonar)
4. Hipoxia severa definida como la ratio entre la presión parcial arterial de oxígeno y la fracción inspirada de oxígeno (PA/FI)

Más allá de la causa de la falla respiratoria, muchos pacientes que reciben ventilación mecánica invasiva desarrollan SDRA.

Los ventiladores mecánicos generalmente se programan para entregar un volumen de aire en cada ciclo ventilatorio (volumen corriente). La entrega de un volumen exagerado en cada ciclo ventilatorio puede causar daño por sobre-estiramiento y lesión en pulmones que ya se encuentran dañados, resultando en un daño mayor a través de la liberación de mediadores inflamatorios.

Durante el SDRA, la entrega de grandes volúmenes corrientes genera que los pulmones, que ya se encuentran inflamados, liberen aún más mediadores inflamatorios lo que puede causar mayor daño pulmonar y daño de otros órganos.

Basándose en la teoría de que el bajo volumen corriente durante la ventilación mecánica invasiva podría actuar de forma protectora tanto para el pulmón como para otros órganos al disminuir la sobredistensión pulmonar y la liberación de mediadores inflamatorios, un reconocido estudio demostró que el uso de bajos volúmenes corrientes en pacientes que se encontraban recibiendo ventilación mecánica por SDRA resultó en una reducción absoluta de la mortalidad de 8,8% [2]. A partir de ese momento, muchos trabajos han demostrado mejoras en la mortalidad a través del tiempo en pacientes con SDRA [3-6], así como la reducción del volumen corriente usado en todos los pacientes internados en UCIs médicas.

Debido a que los pacientes con insuficiencia cardíaca se encuentran excluidos de la definición de SDRA, los mismos no han sido incorporados a estudios de investigación que evalúen los efectos y la epidemiología de la reducción del volumen corriente. Para llenar este bache de conocimiento actual en relación con la selección de volumen corriente en pacientes con insuficiencia cardiaca, realizamos un análisis de tendencia explorando los cambios temporales en la asignación de volumen corriente entre pacientes con insuficiencia cardíaca en comparación con pacientes con SDRA. Para enfrentar las dificultades de identificar el motivo de inicio de la asistencia ventilatoria mecánica en la historia clínica electrónica, ajustamos nuestro plan de análisis focalizándonos en las tendencias de selección del volumen corriente en las UCOs (donde la insuficiencia cardíaca en la causa más común

de inicio de la ventilación mecánica) comparadas con las UCIs (donde son ingresados la mayoría de los pacientes con SDRA).

18.2 Estudio del Set de Datos

En este caso de estudio, utilizamos la base de datos Medical Information Mart for Intensive Care II (MIMIC-II) versión 3 [8], que contiene información *deidentificada* y granular a nivel paciente, de 57,995 hospitalizaciones en UCI correspondientes a 48,018 pacientes de un único centro académico desde el año 2002 al año 2011. La base de datos MIMIC II es una base de datos relacional que contiene registros individuales de una variedad de variables de los pacientes, como por ejemplo resultados de laboratorios, signos vitales y códigos de facturación.

18.3 Preprocesamiento

Identificamos a pacientes ingresados en MIMIC-II que hubieran recibido ventilación mecánica invasiva. Excluimos pacientes menores de 18 años de edad; la práctica clínica y la fisiología en pacientes críticos pediátricos difiere de la de adultos.

Aunque inicialmente buscamos comparar pacientes con SDRA con pacientes con insuficiencia cardíaca, la identificación certera de la indicación especifica de ventilación mecánica fue difícil en las HCE y sujeto a errores de clasificación. Por este motivo, seleccionamos pacientes admitidos en la UCI médica, entendiéndolo como un subrogante de pacientes con SDRA [3,7] y pacientes admitidos en UCO como un subrogante de pacientes con insuficiencia cardíaca. Excluimos a los pacientes ingresados inicialmente a la UCI quirúrgica, debido a que probablemente la mayoría de estos pacientes sólo recibió ventilación mecánica como parte de los cuidados post operatorios habituales.

Para pacientes que fueron atendidos en múltiples UCIs durante la misma internación, basamos los criterios de inclusión/exclusión de acuerdo con la admisión inicial. Además, excluimos aquellos pacientes en los que no se disponía el dato de volumen corriente.

18.4 Métodos

Nuestro resultado primario fue la media de volumen corriente proporcionado por los médicos durante la ventilación asistida-controlada.

Utilizamos el test Cochrane-Armitage para tendencias para evaluar cambios a través del tiempo en el porcentaje de pacientes en cada unidad que requirieron ventilación mecánica invasiva. Calculamos la media de volumen corriente para el período completo en que cada paciente requirió ventilación mecánica invasiva y luego calculamos la media de volumen corriente para UCI y para UCO cada año. Para evaluar cambios en la tendencia temporal en el uso de volumen corriente, hicimos una regresión lineal multivariable (ver Sección 5.2 del Capítulo 5 Análisis de Datos para más detalles) estratificada por el tipo de UCI. El análisis de cambios en la tendencia del volumen corriente a través del tiempo incluyó una variable dependiente (resultado): el volumen corriente y una independiente (o exposición) que fue el tiempo (obtenido del año de ingreso a la UCI). El año de ingreso es una variable de tiempo usada frecuentemente en los análisis de tendencia. El análisis de muestras más pequeñas usando períodos de tiempo más breves, como "meses", puede resultar en una gran cantidad de ruido y fluctuaciones. Elegimos un modelo de regresión multivariable porque el volumen corriente es una variable continua y porque los métodos de regresión permitieron ajustar los estimadores de efecto por posibles confundidores de la relación entre el tiempo y el volumen corriente. Realizamos ajustes por edad y género, ya que ambas variables podrían afectar la selección del volumen corriente. Para determinar diferencias en las tendencias de volumen corriente entre UCI y UCO, incluimos en los modelos de regresión un término de interacción entre el tiempo y la localización del paciente. Para determinar si la variabilidad del volumen corriente promedio cambió a través del tiempo, comparamos el coeficiente de variación (desvío estándar normalizado según la media muestral) desde el inicio del estudio hasta el final del mismo en cada unidad [9]. Para todos los tests se fijó un nivel alfa = 0,05.

Todos los estudios se consideraron exentos de requerimiento de consentimiento informado por los Comités de Revisión Institucional del Boston Medical Center y Beth Israel Deaconess. Todas las pruebas estadísticas se realizaron con SAS 9.4 (Cary, NC).

18.5 Análisis

Se identificaron 7083 pacientes que recibieron ventilación mecánica invasiva en la UCI y 3085 en la UCO desde el año 2002 al año 2011. El número de pacientes que recibió ventilación mecánica en la UCI fluctuó en

este período con un cambio neto consistente en un incremento de 20,2%. El porcentaje de pacientes en UCI que recibió ventilación mecánica invasiva disminuyó de 48,1% en 2002 a 30,8% en 2011 (p<0,0001 *para tendencia)* (Fig. 18.1). Por lo tanto, el aumento del uso de ventilación mecánica en UCI fue dado por un aumento de los sujetos que recibieron esta práctica y no por una mayor probabilidad de recibir ventilación mecánica. Esto se contrapone con la tendencia en UCO, donde la ventilación mecánica disminuyó en un 35,6%, tendencia dominada por un menor número de pacientes y una disminución en la proporción que recibieron ventilación mecánica invasiva (de 58,4% en 2002 a 46,8% en 2011) (p<0,0001 para tendencia) (Fig. 18.2)

La media de volumen corriente en UCO disminuyó un 24.4% durante el período estudiado, de 661 mL (DE=132 mL) en el año 2002 a 500 mL (DE=59) en 2011 (p<0,0001).

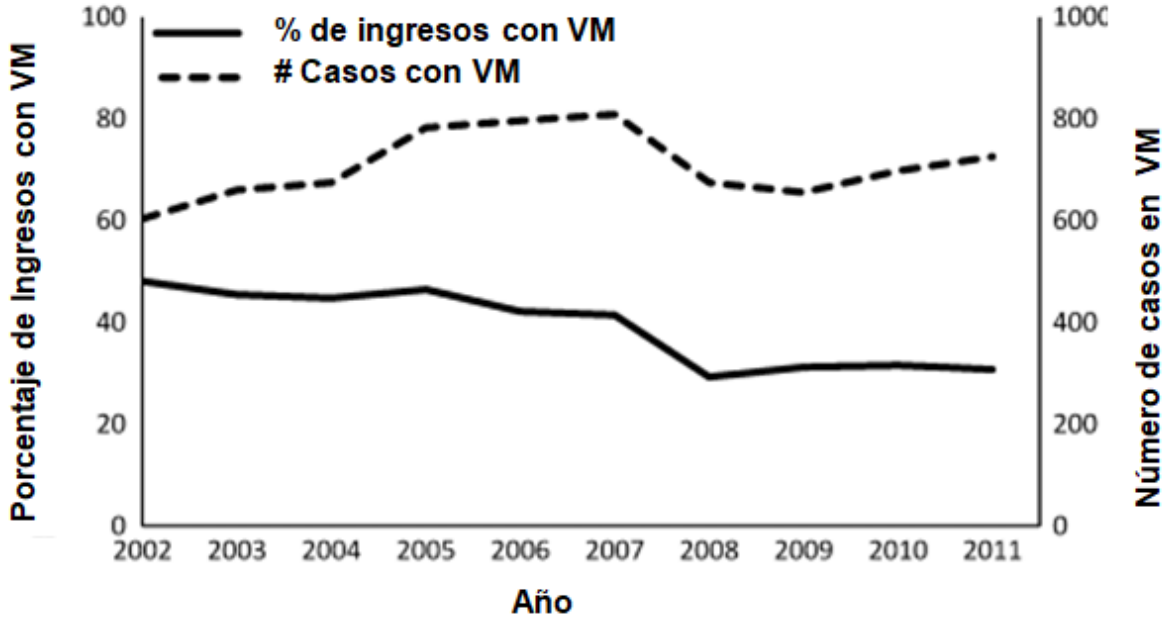

Fig. 18.1: Porcentaje de todos los ingresos (eje Y, escala izquierda) y número de casos (eje Y, escala derecha) que recibieron ventilación mecánica invasiva en la UCI. *VM: ventilación mecánica invasiva, UCI: Unidad de Cuidados Intensivos.*

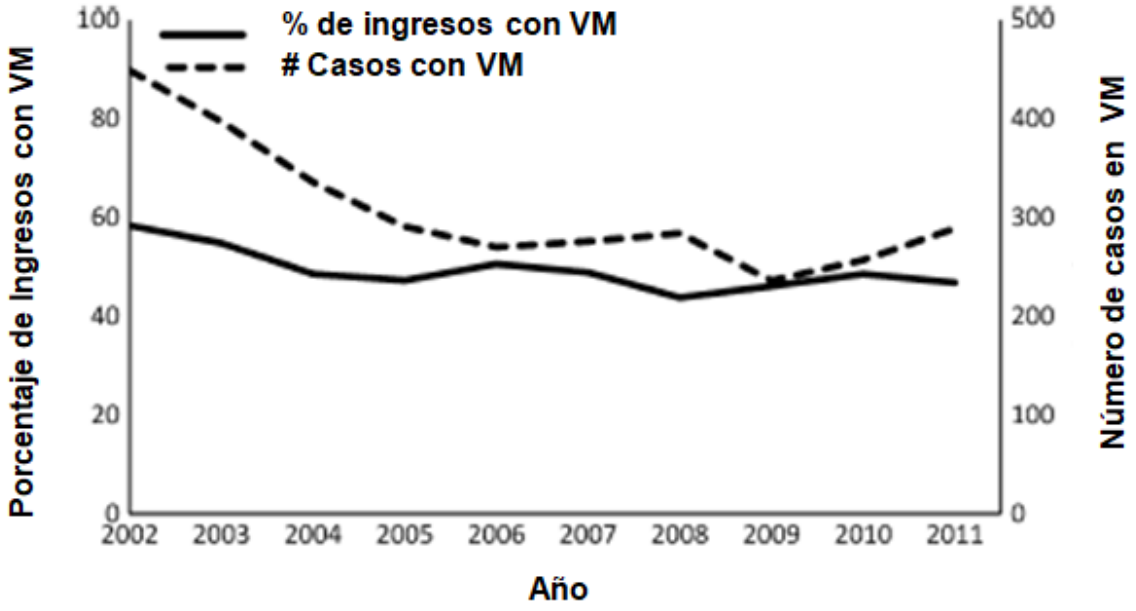

Fig. 18.2: Porcentaje de todos los ingresos (eje Y, escala izquierda) y número de casos (eje Y, escala derecha) que recibieron ventilación mecánica invasiva en UCO. *VM: ventilación mecánica invasiva, UCO: Unidad Coronaria.*

El volumen corriente en la UCI disminuyó un 17.6%, de 568 mL (DE= 121 mL) en el año 2002 a 468 mL (DE=65 mL) en 2011 (p<0,0001) (Fig. 18.3). Durante cada año del período estudiado, la UCO utilizó volúmenes corrientes más altos que la UCI (p<0,0001 para comparación entre unidades por año). Luego de ajustar por edad y género, se encontró una disminución del volumen corriente usado en la UCO a razón de una media de 18 mL por año (IC 95% 16-19 mL, p<0,0001) mientras que los volúmenes corriente en la UCI disminuyeron en promedio 11 mL por año (IC 95% 10-11, p<0,0001). La disminución del volumen corriente fue mayor en la UCO que en la UCI (p para el término de interacción < 0,0001). Adicionalmente, el coeficiente de variación disminuyó en ambas unidades durante el periodo estudiado (UCI: 20,0% en el año 2002 a 11.8% en el año 2011, p <0,0001; UCO: 21.3% en el año 2002 a 13.9% en el año 2011, p<0,0001).

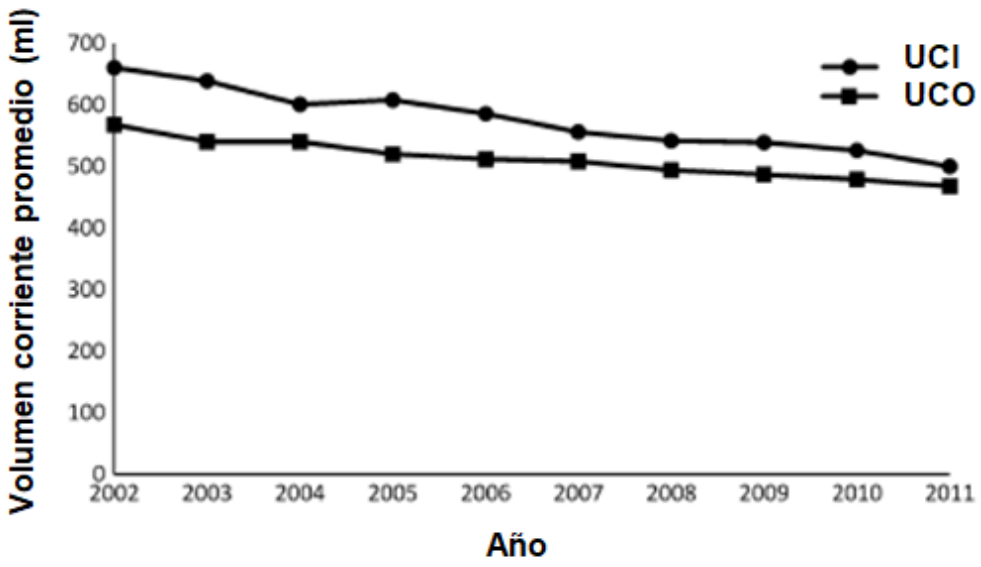

Fig. 18.3: Volumen corriente promedio en la UCI y la UCO por año. Para cada año, el promedio de volumen corriente fue más alto en la UCO, p<0,0001. La pendiente de descenso en el volumen corriente fue mayor para la UCO, p< 0,001. *UCI: Unidad de cuidados Intensivos. UCO: Unidad Coronaria.*

18.6 Conclusiones

Mientras que hay fuerte evidencia que indica los beneficios en la sobrevida del uso de volúmenes corrientes menores en pacientes con edema pulmonar no cardiogénico (SDRA) [2] existe muy poca evidencia para su aplicación en pacientes con edema pulmonar cardiogénico (insuficiencia cardíaca). Utilizando la base de datos MIMIC-II, identificamos una disminución en las tasas de utilización de ventilación invasiva tanto en UCI como en UCO, a pesar de observarse un aumento en el número de pacientes ventilados en UCI. El volumen corriente disminuyó en ambas unidades durante el período de estudio. Es interesante ver que el volumen corriente disminuyó de forma más rápida en la UCO que en la UCI, llegando a volúmenes corrientes casi equivalentes a los de UCI en el año 2011. Esta reducción más rápida del volumen corriente usado en la UCO ocurrió a pesar de no existir evidencia fuerte del uso de volúmenes corrientes bajos en pacientes con edema pulmonar de origen cardiogénico o insuficiencia cardíaca. Además de la disminución del volumen corriente, la variación en la selección del mismo también disminuyó a través del tiempo, demostrando una tendencia a la uniformidad en su selección. Nuestros hallazgos

demuestran una generalización de la evidencia para SDRA, aplicándose también a pacientes cuyas características los excluyeron de los estudios existentes al respecto.

18.7 Próximos Pasos

Nuestro análisis presenta varias limitaciones. Primero, existen muchos factores que afectan la elección del volumen corriente en las distintas unidades, como pueden ser el peso corporal, el *drive* respiratorio y el estado ácido base. Si estos factores no medidos se hubieran modificado a través del tiempo, entonces podrían haber actuado como confundidores del cambio observado en el volumen corriente a través del tiempo. Incluir estas covariables en el análisis de regresión podría reducir el efecto confundidor. Para el propósito de este caso de estudio, hemos limitado nuestras covariables a las características demográficas, pero podrían agregarse otras al modelo en futuros análisis. Segundo, la variable resultado de nuestro objetivo primario fue la media de volumen corriente. No buscamos los cambios en los volúmenes corrientes de cada paciente durante su internación, lo cual podría realizarse en futuros estudios. Tercero, los volúmenes corrientes son generalmente normalizados al peso corporal ideal, debido a que el tamaño pulmonar se correlaciona bien con este parámetro. No disponíamos de los pesos ideales en la base de datos MIMIC-II.

El próximo paso de este estudio sería determinar la asociación entre los cambios de volumen corriente y los cambios en los resultados clínicos. Los estudios que intenten evaluar esta asociación deberán ser cautos e incluir las múltiples variables confundidoras que pudieran mostrar colinealidad con la reducción del volumen corriente. Además, utilizamos a los pacientes ingresados en la UCI como un subrogante de pacientes con SDRA y a los ingresados en la UCO como subrogante de pacientes con insuficiencia cardíaca. Esperamos en futuros estudios poder redefinir nuestros algoritmos de búsqueda en la base de datos de nuestra HCE de forma tal que nos permita diferenciar pacientes con SDRA de aquellos con insuficiencia cardíaca con un riesgo mínimo de sesgo de clasificación. El poder de las bases de datos basadas en HCE, como MIMIC-II, yace en su singular granularidad, brindando una amplia posibilidad de registrar detalles clínicos tales como datos de farmacia, resultados de laboratorio, notas clínicas (a

través de procesamiento de lenguaje natural), etc., lo que permitirá una mayor posibilidad de disminuir el efecto de confundidores.

18.8 Conexiones

El análisis de tendencia permite evaluar cambios en el tiempo de la asistencia sanitaria. En nuestro caso de estudio, utilizamos análisis de regresión lineal para determinar la asociación del tiempo con una variable continua (volumen corriente). Los modelos de regresión permiten a los investigadores incorporar variables confundidoras que podrían haberse modificado en el tiempo junto con la exposición y el resultado de interés. Sin embargo, las técnicas de regresión lineal se limitan a datos que poseen una relación lineal. Para datos no lineales, pueden utilizarse técnicas de transformación (ej transformación logarítmica) que permiten convertir una distribución no lineal a una relación más lineal; también pueden usarse regresión polinómica de alto orden o regresión por *splines*; de forma alternativa puede utilizarse regresión de Poisson.

Para variables categóricas deben utilizarse otras técnicas. El test de Cochrane-Armitage para tendencias es un test de Chi-cuadrado de Pearson modificado que permite ordenar a partir de una de las variables (por ej. la variable tiempo). De forma adicional, la regresión logística multivariable permite el análisis de tendencia para datos categóricos con el beneficio potencial de permitir la incorporación de posibles confundidores como covariables.

Estos métodos de análisis pueden ser ampliamente aplicados, más allá de nuestro caso de estudio. El aspecto fundamental del análisis de tendencia recae en el hecho de que la variable de exposición o independiente es el tiempo. Teniendo en cuenta este concepto, pueden estudiarse múltiples condiciones y tratamientos para ver cómo cambia su utilización a través del tiempo, por ejemplo subgrupos de pacientes que haya recibido ventilación mecánica invasiva [10], pacientes con traqueostomía [11], etc. El análisis de tendencia es una herramienta interesante para evaluar el grado de impacto que han tenido los ensayos clínicos en la práctica clínica cotidiana, evidenciando cambios en las tendencias en relación con nuevos descubrimientos o con la publicación de nuevas guías. De forma adicional, el análisis de tendencia es útil para determinar si ciertas intervenciones o procesos han cambiado de forma significativa los resultados. Como en toda

la estadística, debemos comprender los supuestos involucrados en el tipo de test que utilicemos y asegurarnos de que los datos cumplan con estos criterios.

Apéndice: Código

El código utilizado en este caso de estudio se encuentra disponible en el repositorio de GitHub que acompaña este libro https://github.com/MIT-LCP/critical-data-book. En el sitio web se encuentra disponible mayor información sobre el código.

Referencias

1. The ARDS Definition Task Force (2012) Acute respiratory distress syndrome: the Berlin definition. JAMA 307 (23): 2526-2533.
2. Amato MB, Barbas CS, Medeiros DM, Laffey JG, Engelberts D, Kavanagh BP (2000) Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. The acute respiratory distress syndrome network. N Engl J Med 342 (18): 1301-1308.
3. Esteban A, Frutos-VIvar F, Muriel A et al (2013) Evolution of mortality over time in patients receiving mechanical ventilation. Am J Respir Crit Car Med 188 (2): 220.
4. Rubenfeld GD, Caldwell E, Peabody E et al (2005) Incidence and outcomes of acute lung injury. N Engl J Med 353 (16): 1685-1693.
5. Erickson SE, Martin GS, Davis JL et al (2009) Recent trends in acute lung injury mortality: 1996-2005. Crit Care Med 37 (5): 1574-1579.

6. Zambon M, Vincent JL (2008) Mortality rates for patients with acute lung injury/ARDS have decreased over time. Chest 133 (5): 1120-1127.
7. Esteban A, Ferguson ND, Meade MO et al (2008) Evolution of mechanical ventilation in response to clinical research. Am J Respir Crit Care Med 177 (2): 170-177.
8. Scott DJ, Lee J, Silva I et al (2013) Accessing the public MIMIC-II intensive care relational database for clinical research. BMC Med Inform Decis Mak 13:9. doi: 10.1186/1472-6947-13-9.
9. United States Forest Service (2015) A likelihood ratio test of the equality of the coefficients of variation of k normally distributed populations. http://www1.fpl.fs.fed.us/covtestk.html. [Consultado 28 Julio 2015].
10. Mehta AB, Syeda SN, Wiener RS et al (2015) Epidemiological trends in invasive mechanical ventilation in the United States: a population-based study. J Crit Care 30 (6): 1217-1221.
11. Mehta AB, Syeda SN, Bajpayee L et al (2015) Trends in tracheostomy for mechanically ventilated patients in the United States, 1993-2012. Am J Respir Crit Care Med 192 (4): 446-454.

Capítulo 19

ANÁLISIS DE VARIABLES INSTRUMENTALES DE HISTORIAS CLÍNICAS ELECTRÓNICAS

Nicolás Della Penna, Jennifer P. Stevens y Robert Stretch

Objetivos de Aprendizaje

En este caso de estudio aprenderemos como:

- Estimar los efectos causales de una potencial intervención cuando hay una variable instrumental disponible.
- Identificar los modelos apropiados con los cuales estimar los efectos utilizando variables instrumentales.
- Examinar las posibles fuentes de heterogeneidad de los efectos de los tratamientos.

19.1 Introducción

El objetivo de los estudios observacionales es identificar el efecto causal de una exposición o de tratamientos en un resultado clínico de interés. La disponibilidad de datos derivados de las historias clínicas electrónicas (HCEs) ha mejorado la factibilidad de realizar estudios observacionales a gran escala. Sin embargo, tanto los tratamientos como las características de los pacientes (covariables) afectan el resultado final. Como, en general, ambas son dependientes, no es correcto comparar los resultados de pacientes que reciben distintos tratamientos para decidir cuál de ellos es el más efectivo. Mientras que un análisis de regresión puede incorporar estas covariables, las estimaciones pueden encontrarse sesgadas por covariables que no son observadas y que afectan la indicación de tratamiento y los resultados.

Los estudios controlados, randomizados, solucionan este problema de covariables no identificadas, debido a que mediante la randomización se balancean estas covariables de forma homogénea tanto en el grupo tratamiento como en el control, a medida que la muestra se va tornando más grande.

En la práctica, sin embargo, estos estudios se encuentran afectados por la no adherencia de los pacientes al tratamiento. Las técnicas de variable

instrumental, que utilizan la asignación del tratamiento (como el instrumento) y al tratamiento, tomados como variables endógenas (aquellas que resultan de elecciones que pueden verse afectadas por variables no observadas), son útiles en este escenario.

El análisis de variables instrumentales (AVIs) intenta utilizar "experimentos naturales", fuente de randomización de sujetos a tratamientos, no intencional, pero efectiva. Para tomar ventaja de estos experimentos naturales, los sujetos deben poseer alguna característica observable que los haga más propensos a recibir determinado tratamiento sin que esto modifique el resultado de interés y sea independiente de las variables no observables. (ver Fig. 19.1). La estimación entonces se realiza utilizando únicamente la variación causada por esta característica observable llamada *instrumento* o *variable instrumental (VI),* pudiendo identificar de esta forma el efecto. Hay tres consideraciones principales a tener en cuenta en la selección de los controles apropiados e instrumentos válidos:

1. **Las variables de control deben ser características de los pacientes o del personal de salud previas al tratamiento:** No se deben controlar los resultados o las decisiones que ocurran después del tratamiento, incluso si estos no son el resultado de interés, ya que esto sesgaría los resultados. Dibujar el modelo causal y analizar los caminos permite entender los supuestos que se están realizando. En la web se encuentra disponible un software disponible [1] para facilitarlo.

2. **El instrumento debe estar relacionado con el tratamiento y debe explicar una parte substancial de la variación del tratamiento:** Entre menor sea la variación en el tratamiento explicada por el instrumento (entre "más débil" sea el instrumento), mayor será la varianza de la estimación. Esta mayor varianza puede inhabilitar cualquier beneficio en la reducción del sesgo.

3. **El instrumento debe ser independiente del resultado a través de cualquier otro mecanismo que sea distinto al tratamiento:** Este es el mayor desafío de usar AVIs, puesto que puede resultar complejo

identificar instrumentos que no guarden relación con ninguna otra variable clínica no observable por fuera del tratamiento.

Para entender estos conceptos, proponemos el uso de AVIs para estimar el efecto en la mortalidad en una unidad de cuidados intensivos (UCI) *"non-target",* definida como una unidad que tiene un enfoque de especialidad diferente al de la UCI a la que los pacientes habrían sido asignados, teniendo en cuenta que no existieran limitaciones de capacidad. Por ejemplo, los pacientes que son atendidos por un equipo clínico de UCI, cuidan idealmente a sus pacientes en un área geográfica denominada Unidad de Cuidados Intensivos Médicos (UCIM), pero cuando no existen camas disponibles, un paciente puede ser ingresado en una UCI "*non-target*", como por ejemplo la Unidad de Cuidados Intensivos Quirúrgicos (UCIQ). En este estudio, definimos a estos pacientes asignados en UCI "*non-target"* como *"boarders".*

Aunque los médicos de las UCIM siguen siendo responsables de los pacientes internados fuera de área o *"boarders",* el resto del personal involucrado en la atención del paciente (ej. enfermeros, kinesiólogos respiratorios, especialistas en rehabilitación), será distinto. Esto sucede porque este personal está designado para un área geográfica específica, por ejemplo UCIQ. Como resultado, los pacientes "boarders" son atendidos por enfermeros y otro personal que tiene mayor experiencia en pacientes quirúrgicos que en pacientes clínicos. Además, como los médicos y enfermeros que trabajan en otras UCIs pueden no estar familiarizados con las prácticas clínicas del lugar, pueden ocurrir errores en la comunicación. Por último, existen también grandes distancias geográficas entre los pacientes "boarders" y sus médicos en comparación con los pacientes internados en el area correspondiente a su patología o "no boarders". Esto puede contribuir a demoras en la atención y desconocimiento por parte del médico tratante de la situación de su paciente. Con todo lo expuesto, es lógico hipotetizar que la internación fuera de área podría impactar de forma negativa en el resultado clínico en los pacientes, incluyendo su supervivencia.

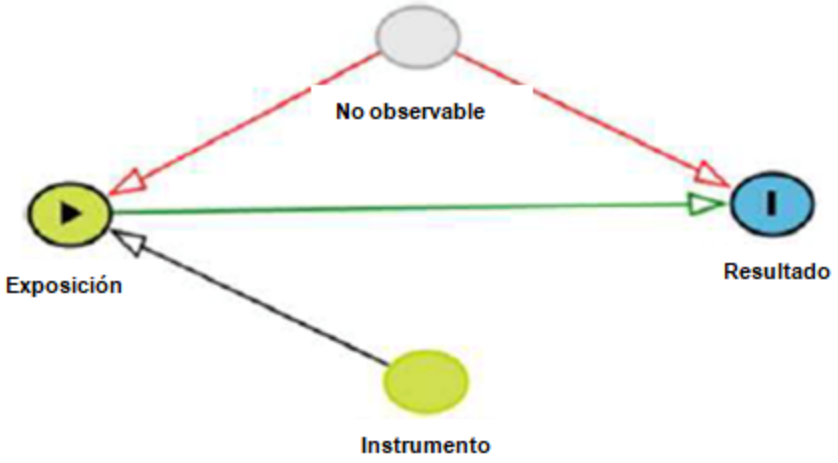

Fig 19.1: El análisis de variable instrumental emplea instrumentos que pueden afectar la probabilidad de exposición, pero no así el resultado.

19.2 Métodos

19.2.1. Set de Datos

La base de datos Medical Information for Intensive Care (MIMIC-III) contiene datos clínicos y administrativos de más de 60.000 estancias en UCI en el hospital Beth Israel Deaconess Medical Center (BIDMC) entre 2001 y 2012. Esta incluye desde datos operativos como asignación de camas y transferencias, hasta datos de diagnósticos CIE-9 y varias medidas de mortalidad (mortalidad en la UCI, mortalidad hospitalaria y supervivencia hasta por un año).

19.2.2 Metodología

Selección de la cohorte

Incluimos todos los pacientes adultos, de 18 años o más, atendidos por la UCIM en cualquier momento de su admisión. El período de estudio fue definido entre junio de 2002 a diciembre de 2012. Para asegurarnos la independencia en las observaciones, sólo tuvimos en cuenta el último ingreso a UCI para cada sujeto incluido en el análisis.

Los criterios de exclusión fueron: pacientes cuyo equipo tratante en cualquier momento de su internación haya sido un equipo no clínico (por ej. quirúrgico o cardiovascular, ya que esto podría implicar una razón específica, aparte de la limitante en capacidad de la Unidad, para que el

paciente sea un "*boarder"* en una UCI no médica (por ej. un paciente postoperatorio en una UCI quirúrgica que haya sido transferido de una UCI quirúrgica a una UCI médica por persistencia de falla respiratoria).

Finalmente, el estudio incluyó 8442 pacientes, de los cuales 1881 (22%) resultaron expuestos a los efectos de encontrarse internados fuera del área especializada para su patología o "boarding".

Análisis estadístico

Una forma básica de estimar el efecto en la mortalidad de encontrarse internado fuera de área ("*boarding"*), sería comparar los resultados en los pacientes que fueron "boarders" y los que no. Sin embargo, la decisión de trasladar a un paciente no se toma de manera arbitraria. Se tiene en cuenta el nivel de severidad de la condición del paciente, así como la de otros pacientes que también requieren de una cama en UCI. Es muy probable que la información sobre esta toma de decisión sea inobservable. Como consecuencia, si realizaramos este análisis como una regresión simple obtendríamos información sesgada sobre el efecto de *"boarding".*

Por ejemplo, asumamos que la internación fuera de área especializada, aumenta la mortalidad, pero también que el personal de la UCI selecciona los pacientes menos graves para pasarlos a otra unidad. En este escenario hipotético, la asociación observada entre *boarding* y mortalidad podría mostrase como protectora si el efecto negativo de *boarding* en mortalidad es menor que el efecto positivo de seleccionar a pacientes más sanos. Por más que uno pueda y deba controlar el nivel de severidad de la enfermedad y condiciones de salud pre existentes, no siempre es posible hacerlo con el mismo detalle y precisión que lo hace el equipo de salud tratante que decide quién se convertirá en un *boarder.* Como resultado, los pacientes "*boarders"* pueden ser más sanos que los *"no boarders",* incluso después de ajustar por una medida de severidad de enfermedad.

Un AVI se presenta como una solución atractiva frente a esta situación. En este estudio, nos hemos focalizado en los pacientes de UCIM. Hemos propuesto que el número de camas remanentes disponibles en el Pabellón Oeste de la UCI médica al momento del ingreso del paciente *(west_initial_remainig_beds)* podrían servir de un instrumento válido para determinar el estado de *boarding.* Es importante remarcar que *west_initial_remainig_beds* no incluye las camas disponibles por fuera de la

UCIM (por ej. camas en los que los pacientes *"boarders"* pueden ser asignados). El estado de "*boarder"* de un paciente es la *variable causal* y el *resultado* es la muerte durante su internación en UCI (Fig. 19.2).

El Oxford Acute Severity of Illness Score (OASIS) se utiliza para analizar las diferencias residuales entre el estado de salud de los pacientes *"boarders"* y de los "no boarders" al momento de ser ingresados en la UCI. OASIS es un puntaje de UCI que ha demostrado un desempeño que no es inferior al de APACHE (Acute Physiology and Chronic Health Evaluation), MPM (Mortality Probability Model) y al SAPS *(Simplified Acute Physiology Score)* [2]. Preferimos utilizar OASIS para severidad de enfermedad porque es el puntaje que mejor se ajusta a MIMIC-III y nos permite reconstruirlo de forma retrospectiva.

En los momentos en que el hospital presenta muchos pacientes internados, es más probable que la UCI presente muchos pacientes internados también *(west_initial_team_census*) y el *west_initial_remaining_beds* probablemente sea bajo. Además es plausible que un valor más alto de *west_initial_team_census* pueda afectar la mortalidad ya que una cantidad relativamente fija de recursos de la UCI (ej. médicos) se extiende a un mayor número de pacientes.

Al principio podría no parecer claro porque existe una correlación imperfecta entre *west_initial_team_census* y *west_initial_remaining_beds,* debido a que se podría anticipar que el número de camas restantes es simplemente inversamente proporcional al total de pacientes que están siendo atendidos en UCI. La fuente de variabilidad entre estas variables es doble. El principal impulsor es el patrón estocástico de egresos en UCI. Es improbable que todos los pacientes *"boarders"* sean egresados antes que los *"no boarders".* Darle de alta a un "non boarder" mientras otro paciente continúa como "boarder" genera una situación donde el censo total de recursos médicos va a continuar siendo mayor que la capacidad de camas en UCIM, entonces el número de camas de UCIM es mayor a cero. En segundo lugar, una fuente de variación más pequeña es la ocupación de camas de UCIM por pacientes que están siendo atendidos por otros equipos de UCI (por ej. un paciente de UCIQ internado en la UCIM).

Es válido usar *west_initial_remaining_beds* como un instrumento, pero debemos controlar a *west_initial_team_census*. Para chequear que *west_initial_remaining_beds* se correlaciona con la probabilidad de los

pacientes a internarse fuera de área, creamos un modelo aditivo generalizado (generalized additive model) con una función de enlace logística (logistic link function).

Una vez que se ha identificado el experimento natural y se confirmó la validez del instrumento, puede realizarse un AVI para estimar el efecto causal del tratamiento. El estándar en la literatura econométrica ha sido utilizar una regresión de mínimos cuadrados ordinarios (ordinary least square regression) en dos etapas. Existen dos limitantes importantes para este enfoque en entornos biomédicos. Primero, se requiere de variables continuas para tratamiento y resultado, cuando ambas en medicina tienden a ser discretas o binarias. Segundo, se requiere el conocimiento de la forma funcional de las relaciones subyacentes, de tal manera que los datos puedan ser transformados para crear una relación lineal. Esto, a menudo, va más allá de los conocimientos del ámbito biomédico.

Se han desarrollado varios enfoques para abordar estas limitaciones. Los modelos Probit forman parte de la familia de modelos lineales generalizados (GLM, por sus siglas en inglés) que se adaptan bien a trabajar con datos discretos, afrontando de este modo la primer limitante antes mencionada. Más aún, el uso de una expansión de base permitiría que la forma funcional sea aproximadamente flexible usando splines penalizados, relajando sustancialmente la segunda limitación relacionada con el conocimiento de la forma de la función. Por lo menos el paquete estadístico, *SemiParBIVProbit* para R, combina estos dos enfoques con una forma accesible de implementación.

Además del modelo Probit, utilizamos el paquete R de sobrevida para estimar un modelo no instrumental de riesgos proporcionales de Cox, como forma para verificar la robustez. Para disminuir el sesgo de selección en ese modelo no instrumental, utilizamos un subconjunto de la base de datos que intuitivamente reducirá la presión selectiva: *west_initial_remainig_beds* igual a cero (todos los pacientes deberán internarse fuera de área independientemente de la severidad de su enfermedad) o *west_initial_remaining_beds* mayor o igual a tres (sin restricción de capacidad inminente que presione a los médicos a internar pacientes fuera de área). Las asunciones de linealidad del modelo Cox son fuertes y no están justificadas *a priori,* sin embargo, para evaluar el potencial no lineal

del modelo instrumental utilizamos el test *Vuong and Clarke* del paquete *SemiParBIVProvit.*

Todos nuestros modelos incluyen controles para edad, género, OASIS y para score de comorbilidad de Elixhauser, tiempo de estadía en el hospital previo al ingreso a UCI y año calendario. Además para controlar *west_initial_team_census,* hemos controlado también el número total de pacientes *"boarders"* bajo el cuidado del personal de UCIM.

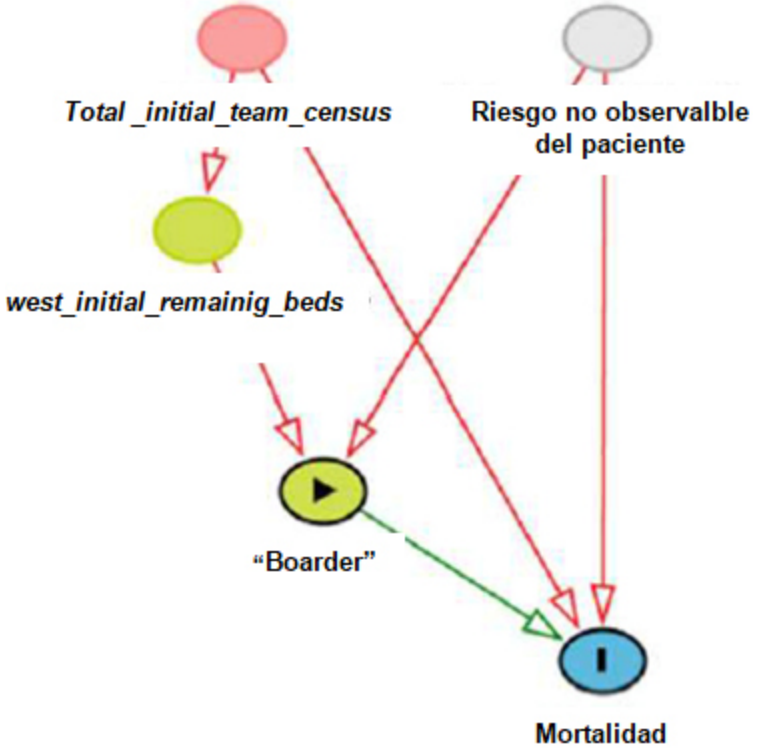

Fig. 19.2: Diagrama causal simplificado que ilustra al confundidor entre la relación de internación fuera de área boarding y mortalidad, dado por una variable heterogénea no observable del riesgo del paciente y un potencial instrumento condicional *west-initial-remainig-beds.* El diagrama puede ser manipulado en http://dagitty.net/dags. html?id= AVKMi0

19.2.3 Preprocesamiento

Utilizamos un paquete de software llamado *Chatto-Transform* [3] que se conecta a una instancia local PostgreSQL del MIMIC-III y simplifica el proceso de importación de tablas de datos a una librería interactiva *Jupyter* [4]. Las librerías usadas para la extracción y análisis de datos fueron Python 3 y Pandas (ver código suplementario).

La versión disponible públicamente de MIMIC-III aplica cambios aleatorios de tiempo a los registros para evitar que los sujetos sean identificados. Después de la aprobación de la junta de revisión institucional, obtuvimos las fechas exactas y las asignaciones de camas para la estadía en la UCI de cada sujeto y lo usamos para reconstruir todo el censo de la UCI del hospital.

La tabla de servicios en MIMIC-III documenta el servicio específico (por ejemplo, medicina, cirugía general, cardiología) responsable de un paciente en un momento dado. El servicio brindado por UCIM es clasificado como "medicina". Por lo tanto, los pacientes de medicina general que ingresan inicialmente en una sala y luego requieren una cama en UCIM solo tendrán una entrada por admisión en esta tabla, siempre que el cuidado no sea transferido a un servicio diferente.

Realizamos una copia depurada de la tabla de servicios (*"med_service_only"*) que incluye solo aquellas filas pertenecientes a pacientes atendidos de forma exclusiva por servicios de medicina durante su estadía. La tabla resultante posee una sola fila por admisión hospitalaria.

La tabla de *transferencias* documenta todos los cambios de locación de un paciente durante su estadía en el hospital, incluyendo la cama exacta y el tiempo que permaneció en ella. Puede ser creada una tabla nueva *df realizando un left join* entre *transfers* y *med_services_only.* En la tabla resultante, las filas que pertenecen a la población de interés (por ejemplo, pacientes de medicina que ingresaron a UCIM en algún momento de su internación) tendrán información de ambos lados de la tabla a la izquierda *(transfers*) y a la derecha *(med_service_only*). Todos los otros pacientes, solo tendrán información del lado de *transfers.* Luego, subdividimos esta tabla en *inboarders* (que contiene filas pertenecientes a pacientes que no son de UCIM ocupando camas de UCIM) y *df5* (que contiene filas pertenecientes a nuestra población de interés).

Recorriendo cada fila de *df5*, identificamos las filas de *inboarders* que representan una cama de UCIM ocupada por un paciente no UCIM al momento en el que un paciente UCIM comienza su estadía en UCI. También determinamos si un nuevo paciente de UCIM era asignado a una cama geográficamente distinta a UCIM, en ese caso, se lo clasificó como "*boarder".* Por último, se realizó un recuento del número total de pacientes que fue atendido por el personal de UCIM y se agregó a cada fila de *df5*.

Estas variables permiten el cálculo de las camas restantes de UCIM a través de la siguiente fórmula:

Camas restantes= (Capacidad de la UCIM– Nº de *inboarders*) – (Censo – Nº de *boarders*)

Se determinó *a priori* que la mortalidad durante la estadía en UCI era nuestro resultado de interés. Identificamos algunos casos en el set de datos donde la muerte ocurrió luego de minutos u horas de ser egresado de la UCIM. Esto probablemente se deba a una combinación entre muertes esperadas (pacientes con cuidados centrados en confort que fueron transferidos fuera de la UCIM previo a su muerte), muertes inesperadas y una pequeña discrepancia en el tiempo en el que se cargan los datos, lo cual incluye el trabajo administrativo. Previo al análisis de datos, se decidió que preferíamos que la definición de *muerte durante la estancia en UCI* incluyera a las muertes que se sucedieran hasta 24 hrs después de haber salido de la UCI.

19.3 Resultados

Observando los modelos ajustados, observamos un aumento de mortalidad en los pacientes internados fuera de área a lo largo de las distintas especificaciones. En el modelo probit semiparamétrico bivariado, usando west_initial_remaining_beds como un instrumento, el riesgo relativo estimado [6] fue de 1.44 (95% IC: 1.17-1.79). En el modelo de riesgos proporcionales no instrumental de Cox observamos una estimación similar de 1.34 (1.06, 1.70).

A menudo, los tratamientos presentan distintos resultados en distintos pacientes, por lo que parece sensato pensar en una media de efectos de tratamiento (MET). El análisis de variable instrumental, sin embargo, restringe la estimación a la variación de los datos que es atribuible al instrumento. Esto quiere decir que el efecto que se estima es el efecto *local* en esos pacientes en los que su tratamiento es afectado por el instrumento. Esto se denomina Efecto Local Promedio del Tratamiento (ELPT) y es lo que se estima con un AVI cuando existe heterogeneidad en los efectos de tratamiento.

19.4 Próximos pasos

Gran parte de la literatura médica existente que utiliza AVI ha abordado cuestiones relacionadas con políticas en lugar del efecto de los tratamientos médicos. Esto ha sido impulsado por el interés en tales preguntas por parte de los economistas de la salud, así como por la mayor disponibilidad de datos administrativos por sobre los clínicos dentro del campo médico.

En contraste, la creciente adopción y sofisticación de las HCEs presenta la oportunidad de investigar los efectos de los tratamientos médicos a través de la provisión de una rica fuente de variables observables y potenciales instrumentos. Los ejemplos incluyen las variaciones medibles en el número y características del personal hospitalario, así como los niveles de carga laboral que causan el desborde entre las diferentes unidades del hospital y, por lo tanto, son externos a un paciente en particular en una unidad dada. Existe también una gran cantidad de literatura que ha explorado la aleatorización mendeliana como fuente de instrumentos, sin embargo, estos generalmente crean una variación limitada, por lo tanto, la debilidad del instrumento es una preocupación sustancial.

Además de servir como candidatos para instrumentos o controles, algunas variables fácilmente extraídas de las HCEs pueden ser útiles para verificar la plausibilidad de un proceso propuesto de pseudo-aleatorización: si un instrumento está realmente aleatorizando pacientes con respecto a un tratamiento, entonces esperaríamos una distribución equilibrada de una amplia gama de variables observables (por ejemplo, datos demográficos del paciente).

Esto es similar a las tablas que comparan las características basales entre los grupos en los resultados de un ensayo clínico controlado aleatorizado. La estimación de los efectos causales a partir de experimentos naturales es una parte importante de la literatura econométrica. Como referencia importante, véase *Mostly Harmless Econometrics* [7]. Se puede encontrar una excelente contraposición en la parte III del trabajo de *Shalizi* [8].

Las variables instrumentales son herramientas poderosas en la identificación de las relaciones causales, pero es fundamental tener en cuenta las posibles fuentes de confundidores. Garabedian y col. revisaron los estudios publicados en la literatura médica utilizando AVIs y observaron que las cuatro categorías de instrumentos más utilizadas (distancia al establecimiento, variación regional, variación del establecimiento y variación del médico) sufrían de “posibles confundidores de resultado no

ajustados… incluyendo raza del paciente, estado socioeconómico, factores de riesgo clínico, estado de salud y residencia urbana o rural, volumen de instalaciones y procedimientos, y tratamientos concurrentes "[9].

19.5 Conclusiones

Este estudio de caso muestra los pasos involucrados en la identificación y validación de una variable instrumental. También muestra el proceso por el cual se lleva a cabo un AVI para medir el tamaño del efecto e inferir causalidad desde datos observacionales.

Los resultados de nuestro estudio respaldan la hipótesis de que la asistencia de pacientes críticos fuera de una unidad especializada para su patología (boarding) tiene efectos nocivos sobre la sobrevida en la UCI. Recomendamos que las instituciones tomen medidas para minimizar el ingreso de pacientes de UCI fuera de área y que se realicen más estudios para caracterizar con mayor precisión el tamaño del efecto. Una mejor comprensión de los mediadores a través de los cuales la internación fuera de área especializada influye en la mortalidad, puede ayudar a identificar grupos de pacientes en los que puede realizarse sin efectos perjudiciales de aquellos en quienes debe particularmente evitarse.

Nota: Las imágenes de este capítulo se encuentran disponibles a color en el ebook; disponible en www.hardineros.com

Apéndice: Código

El código usado en este caso de estudio está disponible en el repositorio de GitHub repository que acompaña el libro https://github.com/MIT-LCP/critical-data-book. En este sitio web se encuentra disponible mayor información sobre el código.

Referencias

1. Textor J, Hardt J, Knüppel S (2011) DAGitty: a graphical tool for analyzing causal diagrams. Epidemiology 22 (5): 745.
2. Johnson AEW, Kramer AA, Clifford GD (2013) A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy. Crit Care Med 41 (7): 1711-1718.
3. Spitz D, Spencer D (2015) Chatto-transform.
4. Jupyter Team, “Project Jupyter”.
5. PyData Development Team (2015) Pandas data analysis library.
6. Marra G, Giampiero M, Rosalba R (2011) Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity. Can J Stat 39 (2): 259-279.
7. Angrist JD, Pischke J-S (2008) Mostly harmless econometrics: an empiricist’s companion. Princeton University Press, Princeton.
8. Shalizi CR (2016) Advanced data analysis from an elementary point of view, 18 Jan 2016.
9. Garabedian LF, Chu P, Toh S, Zaslavsky AM, Soumerai SB (2014) Potential bias of instrumental variable analyses for observational comparative effectiveness research. Ann Intern Med 161 (2): 131-138.

Capítulo 20

PREDICCIÓN DE MORTALIDAD EN LA UNIDAD DE CUIDADOS INTENSIVOS BASADA EN LOS RESULTADOS EN MIMIC-II DEL PROYECTO SICULA (SUPER ICU LEARNER ALGORITHM)

Romain Pirracchio

Objetivos de aprendizaje

En este capítulo ilustramos el uso de los datos clínicos de MIMIC II, el algoritmo de predicción no paramétrico, aprendizaje automatizado ensamblado y el algoritmo Super Learner.

20.1 Introducción

La predicción de la mortalidad en pacientes hospitalizados en unidades de cuidados intensivos (UCI) es crucial para evaluar la severidad de enfermedad y entender el valor de tratamientos novedosos, intervenciones y políticas de salud. Se han desarrollado distintos puntajes de severidad con el objetivo de predecir la mortalidad hospitalaria a partir de las características de base de los pacientes, definidas como aquellas mediciones obtenidas dentro de las primeras 24 hs luego del ingreso en la UCI. Los primeros puntajes propuestos, APACHE [1] (Acute Physiology and Chronic Health Evaluation), APACHE II [2] y SAPS [3] (Simplified Acute Physiology Score), se basaban en métodos subjetivos para determinar la medida de importancia de las variables, que era definida por un panel de expertos que elegían y asignaban un peso a las variables de acuerdo con la relevancia percibida para la predicción de la mortalidad. Otros puntajes como SAPS II [4] fueron desarrollados utilizando técnicas de modelado estadístico [4-7]. Al día de hoy, los puntajes SAPS II [4] y APACHE II [2] siguen siendo los más utilizados en la práctica clínica. Sin embargo, desde su primera publicación han sido modificados en varias oportunidades para mejorar su desempeño predictivo [6-11]. Pese a estas modificaciones de SAPS, la predicción de la mortalidad hospitalaria continúa siendo sobreestimada [8, 9, 12-14]. A modo de ejemplo, Poole y colegas [9] compararon el desempeño de SAPS II y SAPS 3 en una cohorte de más de 28.000 ingresos en 10 UCI italianas. Concluyeron que ambos puntajes realizaban predicciones no fidedignas, pero de manera

inesperada el puntaje más reciente SAPS 3 sobreestimaba la mortalidad en mayor medida que la versión previa SAPS II. De manera consistente, Nassar y colegas [8] evaluaron el desempeño de los puntajes APACHE IV, SAPS 3 y el Mortality Probability Model III [MPM (0) –III] en una población ingresada en 3 unidades de cuidados intensivos quirúrgicas de Brasil y encontraron que todos los modelos mostraron pobre calibración, mientras que la discriminación fue buena en todos ellos.

La mayoría de los puntajes de severidad de UCI se basan en un modelo de regresión logística. Dichos modelos imponen rigurosas restricciones en la relación entre las variables explicativas y el riesgo de muerte. Por ejemplo, la regresión logística de términos principales depende de la presunción de que existe una relación lineal y aditiva entre el resultado y sus predictores. Dada la complejidad de los procesos que subyacen a la muerte de los pacientes en UCI, este supuesto puede ser poco realista.

Dado que la verdadera relación entre el riesgo de mortalidad en UCI y las variables explicativas no se conoce, creemos que dicha predicción puede ser mejorada utilizando un algoritmo automatizado no paramétrico para estimar el riesgo de muerte sin requerir ninguna especificación acerca del grado de la relación subyacente. De hecho, los algoritmos no paramétricos ofrecen la gran ventaja de no depender de ningún supuesto acerca de la distribución subyacente, lo que los hace más adecuados para ajustar datos tan complejos. Algunos estudios han evaluado los beneficios de abordajes no paramétricos, principalmente basados en redes neuronales o minería de datos para predecir la mortalidad hospitalaria de pacientes de UCI [15-20]. Estos estudios concluyeron unánimemente que los métodos no paramétricos pueden desempeñarse por lo menos tan bien como la regresión logística estándar en la predicción de mortalidad en UCI.

Recientemente, el *Super Learner* fue desarrollado como una técnica no paramétrica para seleccionar un algoritmo de regresión óptimo entre un conjunto de algoritmos candidatos provistos por el usuario [21]. El *Super Learner* jerarquiza los algoritmos de acuerdo con su desempeño predictivo y luego construye un algoritmo agregado que se obtiene según la combinación ponderada de los algoritmos candidatos. Los resultados teóricos han demostrado que el *Super Learner* no se desempeña peor que la elección óptima de los algoritmos provistos, por lo menos en grandes muestras. Capitaliza la riqueza de la librería de algoritmos a partir de la cual se

construye y generalmente ofrece beneficios en comparación a cualquier algoritmo candidato específico en términos de flexibilidad para ajustar adecuadamente los datos.

El principal objetivo de este estudio fue desarrollar un procedimiento que otorgue un puntaje para pacientes de UCI basado en el *Super Learner* y utilizando datos del estudio Medical Information Mart for Intensive Care II (MIMIC-II) [22-24] y determinar si este resulta en una mejor predicción de mortalidad en relación con los puntajes SAPS II, APACHE II y SOFA. Los resultados completos de este estudio han sido publicados en 2015 en el Lancet Respiratory Medicine [25]. También buscamos desarrollar una implementación web de nuestro puntaje que fuera fácilmente accesible y amigable para el usuario, pese a la complejidad de nuestro abordaje.

20.2 Set de datos y Preprocesamiento

20.2.1 Recolección de datos y características de los pacientes

El estudio MIMIC-II [22-24] incluye todos los pacientes ingresados a una UCI en el Beth Israel Deaconess Medical Center (BIDMC) en Boston, MA, desde 2001. Para el presente estudio, únicamente se incluyeron datos sobre pacientes de UCI de la versión 2.6 (2001-2008) de MIMIC-II. Los pacientes menores a 16 años no fueron incluidos. En pacientes con múltiples ingresos, sólo consideramos el primer ingreso en UCI. Se incluyeron en el estudio un total de 24.508 pacientes.

20.2.2 Inclusión de los pacientes y mediciones

Se reunieron dos categorías de datos: datos clínicos, agregados de los sistemas de información de las UCI y de los archivos hospitalarios, y datos fisiológicos de alta resolución (formas de ondas y series de tiempo derivadas de mediciones fisiológicas) registrados en los monitores de los pacientes. Los datos clínicos fueron obtenidos del Sistema de Información Clínica CareVue (Philips Healthcare, Andover, Massachusetts) utilizado en todas las UCI del estudio y de las Historias Clínicas Electrónicas. Los datos incluyeron mediciones fisiológicas con fecha y hora y verificadas por enfermería (por ejemplo, documentación horaria de la frecuencia cardíaca, presión arterial, presión arterial pulmonar), notas de evolución de enfermeros y terapistas respiratorios, infusión continua intravenosa de medicamentos, balance

hidroelectrolítico, datos demográficos, informes de estudios por imágenes, prescripciones médicas, resúmenes de alta y códigos CIE-9. Se obtuvieron los resultados de los análisis de laboratorio (por ejemplo, análisis sanguíneo completo, gases arteriales, resultados microbiológicos) de toda la estadía hospitalaria de los pacientes, incluyendo aquellos períodos fuera de la UCI. En el presente estudio, nos focalizamos exclusivamente en variables de resultados (especialmente, mortalidad hospitalaria y en UCI) y variables incluidas en los puntajes SAPS II [4] y SOFA [26].

En primer lugar, realizamos un relevamiento de todas las características necesarias para evaluar los diferentes puntajes considerados. Luego se extrajeron datos crudos de la versión 2.6 de la base de datos MIMIC II. Decidimos utilizar únicamente funciones en R (sin ninguna rutina SQL) ya que la mayoría de nuestros investigadores solamente conocen el paquete R. Se revisaron las tablas de los archivos de datos de los pacientes y se extrajeron todas las características. Finalmente, creamos un archivo CSV global que incluyera todos los datos y fuera fácilmente manipulable con R.

Las variables basales y los resultados se resumen en la Tabla 20.1.

Tabla 20.1 Características de base y medidas de resultados

	Población general (n=24.508)	Fallecidos al alta hospitalaria (n=3.002)	Vivos al alta hospitalaria (n=21.506)
Edad	65 [51-77]	74 [59-83]	64 [50-76]
Género (femenino)	13.838 (56.5%)	1607 (53.5%)	12.231 (56.9%)
Primer SAPS	13 [10-17]	18 [14-22]	13 [9-17]
Primer SAPS II	38 [27-51]	53 [43-64]	36 [27-49]
Primer SOFA	5 [2-8]	8 [5-12]	5 [2-8]
Origen Médico Trauma	2453 (10%) 7703 (31.4%) 10,803	240 (8%) 1055 (35.1%) 1583 (52.7%)	2213 (10.3%) 6648 (30.9%) 9220 (42.9%)

Cirugía de urgencia Cirugía programada	(44.1%) 3549 (14.5%)	124 (4.1%)	3425 (15.9%)
Ubicación UCI clínica (MICU) UCI médico quirúrgica (MSICU) Unidad Coronaria (CCU) UCI cardiovascular (CSRU) Unidad de Cirugía Torácica (T-SICU)	7488 (30.6%) 2686 (11%) 5285 (21.6%) 8100 (33.1%) 949 (3.9%)	1265 (42.1%) 347 (11.6%) 633 (21.1%) 664 (22.1%) 93 (3.1%)	6223 (28.9%) 2339 (10.9%) 4652 (21.6%) 7436 (34.6%) 856 (4%)
Frecuencia cardíaca (lpm)	87 [75-100]	92 [78-109]	86 [75-99]
TAM (mmHg)	81 [70-94]	78 [65-94]	82 [71-94]
Frecuencia respiratoria (rpm)	14 [12-20]	18 [14-23]	14 [12-18]
Na (mmol/l)	139 [136-141]	138 [135-141]	139 [136-141]
K (mmol/l)	4.2 [3.8-4.6]	4.2 [3.8-4.8]	4.2 [3.8-4.6]
HCO3 (mmol/l)	26 [22-28]	24 [20-28]	26 [23-28]
Recuento de glóbulos blancos ($10^3/mm^3$)	10.3 [7.5-14.4]	11.6 [7.9-16.9]	10.2 [7.4-14.1]
PAFI	281 [130-447]	174 [90-352]	312 [145-461]
Hematocrito (%)	34.7 [30.4-39]	33.8 [29.8-38]	34.8 [30.5-39.1]
Urea (mmol/l)	20 [14-31]	28 [18-46]	19 [13-29]
Bilirrubina (mg/dl)	0.6 [0.4-1]	0.7 [0.4-1.5]	0.6 [0.4-0.9]
Estadía hospitalaria (días)	8 [4-14]	9 [4-17]	8 [4-14]

Mortalidad en UCI (%)	1978 (8.1%)	1978 (65.9%)	-
Mortalidad hospitalaria (%)	3002 (12.2%)	-	-

Las variables continuas se presentan como mediana [rango intercuartilo]; las variables binarias o categóricas como frecuencias (%)

20.3 Métodos

20.3.1 Algoritmos de predicción

El principal resultado medido fue la mortalidad hospitalaria. Un total de 1978 muertes ocurrieron en UCI (tasa de mortalidad estimada: 8,1%, IC95%: 7,7-8,4) y 1024 muertes adicionales fueron observadas luego del alta de UCI, resultando en una tasa de mortalidad hospitalaria estimada en 12,2% (IC95%: 11,8-12,7).

Los datos registrados dentro de las primeras 24 hs del ingreso en UCI fueron utilizados para calcular dos de los puntajes de gravedad más utilizados, SAPS II [4] y SOFA [26]. La predicción de mortalidad individual para el puntaje SAPS II fue calculada de la siguiente manera definida por sus autores [4]:

$$\log\left[\frac{\text{pr(death)}}{1-\text{pr(death)}}\right] = -7.7631 + 0.0737 * \text{SAPSII} + 0.9971 * \log(1 + \text{SAPSII})$$

Adicionalmente, desarrollamos una nueva versión del score SAPS II ajustando a nuestros datos un modelo de regresión logística de términos principales que utilizaba las mismas variables explicativas que el puntaje original SAPS II [4]: edad, frecuencia cardíaca, presión arterial sistólica, temperatura corporal, escala de Glasgow, ventilación mecánica, PaO_2, FIO_2, diuresis, nitrógeno ureico en sangre (BUN), sodio sérico, potasio, bicarbonato, bilirrubina, recuento de glóbulos blancos, enfermedades crónicas (SIDA, cáncer metastásico, enfermedad maligna hematológica) y el tipo de ingreso (cirugía electiva, médica, cirugía no programada). Se utilizó el mismo procedimiento para construir una nueva versión del puntaje APACHE II [2]. Finalmente, también se calculó el puntaje SOFA [26] para todos los sujetos dado que se utiliza ampliamente en la práctica clínica como proxy para la predicción de resultado. La predicción de la mortalidad basada en el puntaje SOFA se obtuvo haciendo mediante una regresión de la mortalidad

hospitalaria en el puntaje SOFA utilizando una regresión logística de términos principales. Estos dos algoritmos para la predicción de mortalidad fueron comparados con nuestra propuesta basada en el *Super Learner*.

El *Super Learner* se ha propuesto como un método para seleccionar a través de validación cruzada el algoritmo de regresión óptimo entre todas las combinaciones ponderadas de un set dado de algoritmos candidatos, de aquí en adelante referido como *la librería* [21, 27, 28] (Fig. 20.1). Para implementar el *Super Learner* el usuario debe proveer una colección personalizada de varios algoritmos que se ajusten a los datos. Luego, el *Super Learner* estima el riesgo asociado a cada algoritmo del set provisto utilizando validación cruzada. Una ronda de validación cruzada implica particionar la muestra de datos en subsets complementarios, realizando el análisis en uno de los subset (llamado set *de entrenamiento*) y validando el análisis en el otro subset (llamado set *de validación* o set *de prueba*). Para reducir la variabilidad, se realizan múltiples rondas de validación cruzada utilizando diferentes particiones y los resultados de las validaciones se promedian entre las rondas. De esta estimación del riesgo asociada a cada algoritmo candidato, el *Super Learner* construye un algoritmo agregado obtenido como la combinación ponderada optima de los algoritmos candidatos. Los resultados teóricos sugieren que para optimizar el desempeño del algoritmo resultante la librería ingresada debería incluir tantos algoritmos sensibles como fuera posible.

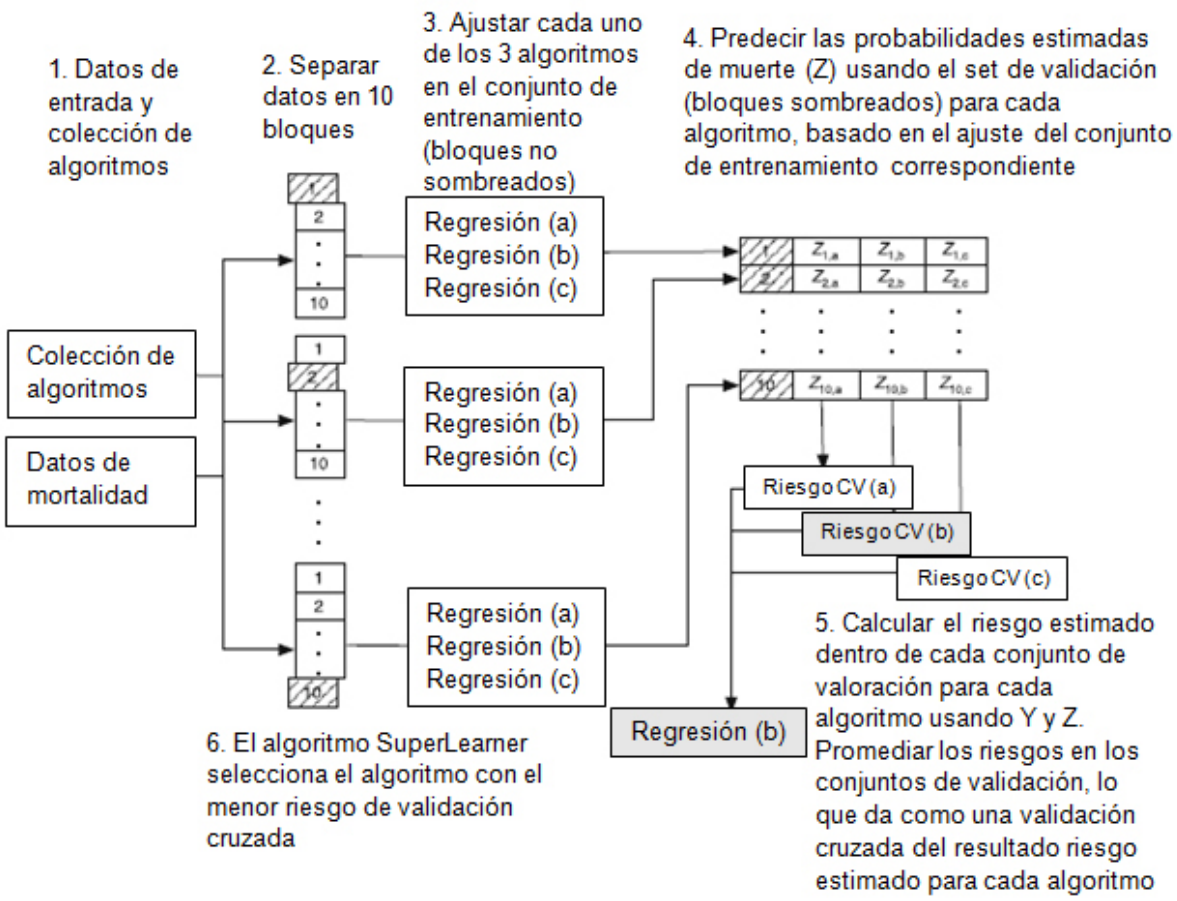

Fig. 20.1 Algoritmo *Super Learner*. De van derLaan, targeted learning 2011 (con permiso) [41].

En este estudio, el tamaño de la librería se limitó a 12 algoritmos (la lista está disponible en el Apéndice) por razones computacionales. Entre estos 12 algoritmos, algunos fueron paramétricos como la regresión logística de métodos afiliados habitualmente utilizada para sistemas de puntajes de UCI y algunos no paramétricos, por ejemplo, métodos que ajustan los datos sin ningún supuesto acerca de la distribución de datos subyacente. En el presente estudio, elegimos la librería de modo que incluyera la mayoría de los algoritmos paramétricos (incluyendo modelos de regresión con varias combinaciones de términos principales y de interacción, así como regresiones splines, y ajustados utilizando máxima probabilidad con o sin penalización) y no paramétricos previamente evaluados en la literatura para la predicción de mortalidad en pacientes críticamente enfermos. La regresión de términos principales es el algoritmo paramétrico que se ha utilizado para construir ambos puntajes SAPS II y APACHE II. Este algoritmo

fue incluido en la librería SL para que el ajuste revisado del puntaje SAPS II basado en los datos actuales también compitiera contra otros algoritmos.

La comparación de los 12 algoritmos requirió una validación cruzada de 10 iteraciones. Los datos primero fueron divididos en 10 bloques mutuamente excluyentes de aproximadamente igual tamaño. Cada algoritmo se ajustó en los 9 bloques correspondientes al set de entrenamiento y luego se usó dicho ajuste para predecir la mortalidad de todos los pacientes en los bloques restantes utilizando un set de validación. Se promediaron los errores al cuadrado entre los resultados predichos y observados, evaluando de esta forma el desempeño de cada algoritmo. Este procedimiento se repitió exactamente 10 veces utilizando un bloque distinto cada vez para el set de validación. Las medidas de desempeño fueron agregadas a lo largo de las 10 iteraciones, obteniendo una estimación del error cuadrático medio por validación cruzada (ECM-VC) para cada algoritmo. Un aspecto crucial de este enfoque es que en cada iteración no hay un solo paciente que aparezca tanto en el set de prueba como en el de validación. De este modo se reduce la posibilidad de que se produzca un sobreajuste, en el que el ajuste de un algoritmo se adapta excesivamente a los datos disponibles a expensas del rendimiento en los datos futuros, ya que es más probable que se produzca un sobreajuste cuando se cruzan los set de entrenamiento y validación.

Los algoritmos candidatos fueron jerarquizados de acuerdo con su ECM-VC y se identificó el algoritmo con menor ECM-VC. Este algoritmo se volvió a ajustar utilizando todos los datos disponibles, llevando a una regla de predicción conocida como el *Discrete Super Learner.* Posteriormente, también se calculó la regla de predicción que consiste en la combinación de ECM-VC con la minimización convexa ponderada de todos los algoritmos candidatos y se reajustó en todos los datos. Esto es a lo que nos referimos como algoritmo combinado *Super Learner* [28].

Los datos utilizados para ajustar nuestro algoritmo predictor incluyeron las 17 variables utilizadas en el score SAPS II: 13 variables fisiológicas (edad, escala de Glasgow, presión arterial sistólica, frecuencia cardíaca, temperatura corporal, PaO_2/FiO_2, excreción urinaria, nitrógeno ureico en sangre (BUN), natremia, kalemia, bicarbonato, bilirrubina, globulos blancos en sangre y nivel de bilirrubina), el tipo de ingreso (cirugía programada, médico, cirugía no programada) y tres variables de enfermedad subyacente (SIDA, cáncer metastásico y cáncer hematológico, derivadas de los códigos

CIE9 al egreso). Se produjeron dos sets de predicciones basadas en el *Super Learner*: el primero basado en las 17 variables tal y como aparecen en el puntaje SAPS II (SL1), y el segundo, con las variables originales sin transformar (SL2)

20.3.2 Métricas de desempeño

Un objetivo clave de este estudio fue comparar el desempeño predictivo de los puntajes SAPS II y SOFA con aquellos basados en el *Super Learner*. Esta comparación se basó en una variedad de medidas de rendimiento predictivo, que se describen a continuación.

1. Se dice que un algoritmo de predicción de mortalidad tiene una discriminación adecuada si tiende a asignar puntajes de severidad más elevados a pacientes que murieron en el hospital en comparación a aquellos que no. Evaluamos la discriminación utilizando validación cruzada para el área bajo la curva ROC (AUROC), con el correspondiente intervalo de confianza del 95% (IC95%). La discriminación puede ilustrarse de manera gráfica utilizando las curvas ROC. Otras herramientas adicionales para evaluar la discriminación incluyen diagramas de caja (boxplots) de las probabilidades de muerte predichas para sobrevivientes y no sobrevivientes y las correspondientes pendientes de discriminación, definidas como la diferencia entre la media de los riesgos previstos en los sobrevivientes y en los no sobrevivientes Todas ellas son proporcionadas más adelante.

2. Se dice que un algoritmo de predicción de mortalidad está calibrado en forma adecuada si las probabilidades de muerte predichas y observadas coinciden de forma adecuada. Evaluamos la calibración utilizando la prueba de calibración de Cox [9,29,20]. Debido a sus numerosos defectos, que incluyen un pobre desempeño en muestras grandes, evitamos utilizar la estadística más convencional de Hosmer-Lemeshow [31,32]. Bajo perfecta calibración, un algoritmo de predicción podrá satisfacer la ecuación de regresión logística: log odds de muerte observada = $\alpha + \beta^*$ log odds de muerte predicha' con $\alpha = 0$.

Para implementar la prueba de calibración de Cox, se realiza una regresión logística para estimar α y β; estas estimaciones sugieren el grado de

desviación de la calibración ideal. La hipótesis nula (α, β) = (0,1) se prueba formalmente utilizando estadísticas U [33].

3. Las medidas resumen de reclasificación, incluyendo el índice continuo de reclasificación neta (cNRI) y el Índice de mejora de la discriminación integrada (IDI), son métricas relativas que han sido ideadas para superar las limitaciones de las medidas habituales de discriminación y calibración [34-36]. El cNRI que compara los puntajes de severidad A y B se define como el doble de la diferencia entre la proporción de no sobrevivientes y de sobrevivientes, respectivamente, considerados más severos según el puntaje A en lugar del puntaje B. El IDI para comparar el puntaje de severidad A con el B es el promedio de la diferencia en el puntaje A entre sobrevivientes y no sobrevivientes menos el promedio de la diferencia en el puntaje B entre sobrevivientes y no sobrevivientes. Valores positivos de cNRI e IDI indican que el puntaje A tiene mayor capacidad discriminativa que el puntaje B, mientras que valores negativos indican lo contrario. Hemos calculado las tablas de reclasificación y las medidas de resumen asociadas para comparar cada propuesta del *Super Learner* con el puntaje original SAPS II y con cada uno de los ajustes revisados de los puntajes SAPS II y APACHE II.

Todos los análisis fueron realizados utilizando el software estadístico R versión 2.15.2 para Mac OS X (The R Foundation for Statistical Computing, Vienna, Austria; paquetes específicos: cvAUC, Super Learner y ROCR). Los códigos relevantes de R son proporcionados en el Apéndice.

20.4 Análisis

20.4.1 Discriminación

La curva ROC para predicción de mortalidad hospitalaria se muestra debajo (Fig. 20.2). El AUROC fue 0,71 (IC95%: 0,70-0,72) para el puntaje SOFA y 0,78 (IC95%: 0,81-0,83) para el puntaje SAPS II. Los dos modelos predictivos *Super Learner* (SL1 y SL2) superaron sustancialmente los puntajes SAPS II y SOFA. El AUROC fue 0,85 (IC95% 0,84-0,85) para SL1 y 0,88 (IC95%: 0,87-0,89) para SL2 revelando una clara ventaja de los algoritmos predictores basados en *Super Learner* por sobre ambos puntajes SOFA y SAPS II.

La discriminación también fue evaluada al comparar las diferencias entre las probabilidades de muerte predichas entre los sobrevivientes y no sobrevivientes utilizando cada algoritmo de predicción. La pendiente de discriminación igualó a 0,09 para el puntaje SOFA, 0,26 para el puntaje SAPS II, 0,21 para SL1 y 0,26 para SL2.

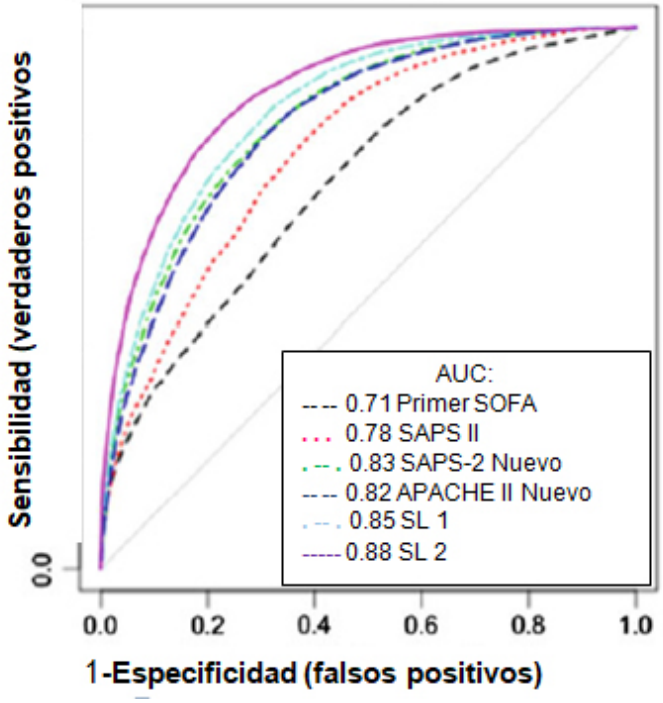

Fig. 20.2 Curvas ROC Super Learner 1: Super Learner con variables categorizadas; Super Learner 2: Super Learner con variables sin transformar.

20.4.2 Calibración

Los gráficos de calibración (Fig. 20.3) indican una falta de ajuste para el puntaje SAPS II. Los valores estimados de α y β fueron –1,51 y 0,72 respectivamente (estadística-U = 0,25; p < 0,0001). Las propiedades de calibración mejoraron notablemente reajustando el puntaje SAPS II: α < 0,0001 y β = 1 (U < 0,0001; p = 1,00). La predicción basada en los puntajes SOFA y APACHE II mostró propiedades de calibración excelentes, como bien lo refleja un α < 0,0001 y β =1 (U < 0,0001; p = 1,00). Para las predicciones basadas en *Super Learner*, pese a que las estadísticas-U fueron significativamente diferentes a cero, los estimados de α y β fueron cercanos a los valores nulos: SL1:0,14 y 1,04 respectivamente (U = 0,0007; p = 0,0001); SL2:0,24 y 1,25 respectivamente (U = 0,006; p < 0,0001).

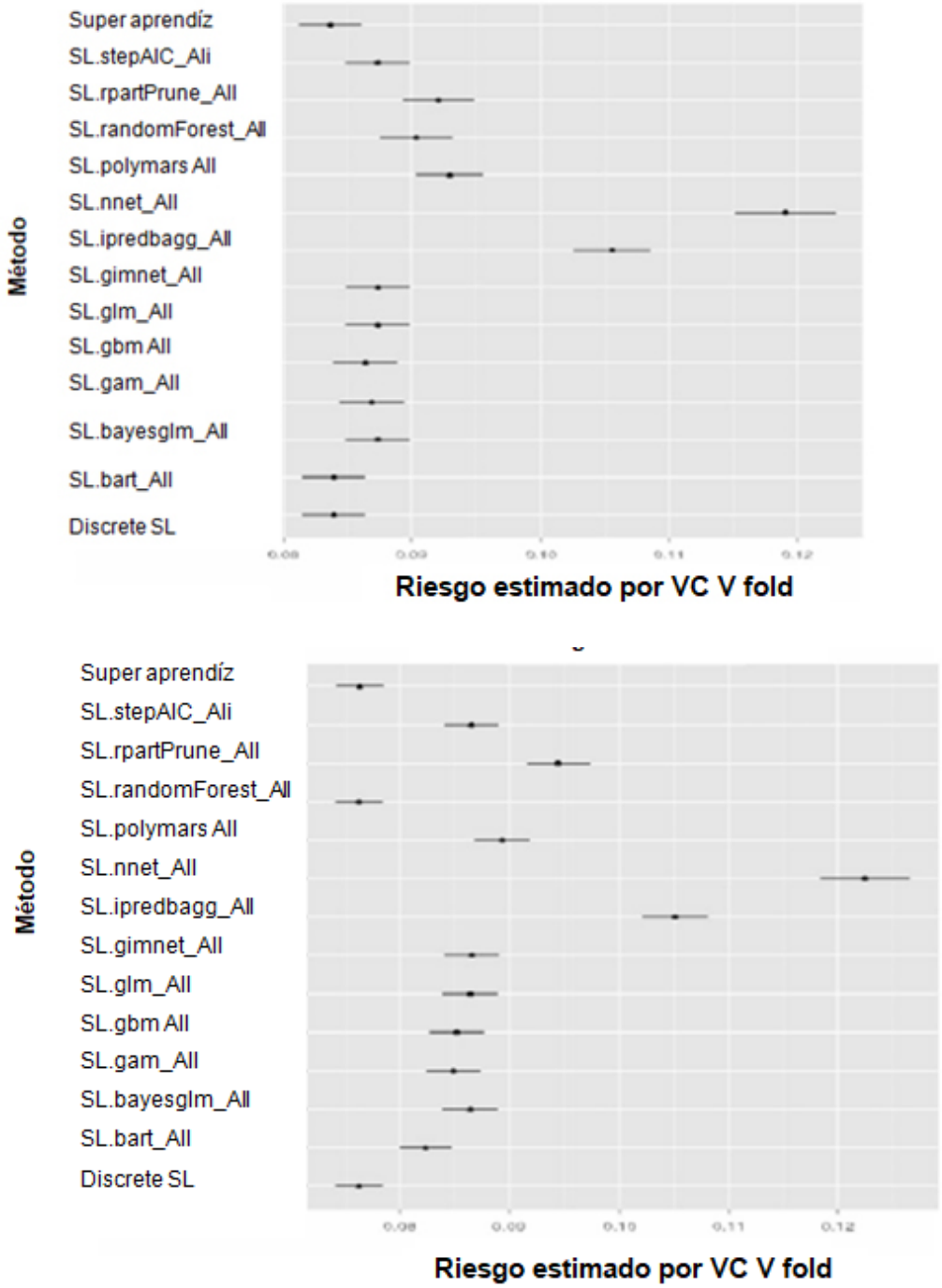

Fig. 20.3 Gráficos de calibración y discriminación para SAPS 2 (panel superior) y SL1 (panel inferior)

20.4.3 La librería Super Learner

El desempeño de los 12 algoritmos candidatos, el Discrete *Super Learner* y las combinaciones de algoritmos *Super Learner* evaluados por EMC-VC y AUROC-VC se ilustran en la Fig. 20.4.

Como se sugirió en la teoría, cuando se utilizaron ya sea variables categóricas (SL1) o variables no transformadas (SL2) el algoritmo combinado *Super Learner* consiguió el mismo desempeño que el mejor de los 12 candidatos, con un EMC-VC promedio de 0,084 (DE = 0,001) y una AUROC promedio de 0,85 (IC95%: 0,84-0,85) para SL1 [mejor algoritmo individual: Modelo de Árbol de Regresión aditivo Bayesiano, con EMC-VC = 0,084 y AUROC = 0,84 (IC95%: 0,84-0,85)]. Para el SL2, el EMC-VC promedio fue de 0,076 (DE = 0,001) y AUROC promedio fue de 0,88 (IC95%: 0,87-0,89) [mejor algoritmo individual: Random Forest, con EMC-VC = 0,076 y AUROC = 0,88 (95%IC: 0,87-0,89)]. En ambos casos (SL1 y SL2), el *Super Learner* superó la regresión de términos principales utilizada para desarrollar los puntajes SAPS II o APACHE II [regresión logística de términos principales: EMC-CV = 0,087 (ES = 0,001) y AUROC = 0,83 (95%IC: 0,82-0,83)].

20.4.4 Tabla de reclasificación

Las tablas de reclasificación que involucran el puntaje SAPS II en sus versiones original y actualizada, el puntaje APACHE II revisado y los puntajes SL1 y SL2 se muestran en la Tabla 20.2. Cuando se compara con la clasificación provista por el SAPS II original, el SAPS II actualizado o el APACHE II revisado, los puntajes basados en *Super Learner* resultaron en un descenso de la gran mayoría de pacientes hacia un estrato de menor riesgo. Este fue el caso especialmente para pacientes con una probabilidad de muerte por encima de 0,5.

Se calculó el cNRI y el IDI considerando cada propuesta de *Super Learner* (puntaje A) como el modelo actualizado y el SAPS II original, el nuevo SAPS II y el nuevo puntaje APACHE II (puntaje B) como el modelo inicial. En este caso, valores positivos de cNRI e IDI indicarían que el puntaje A tiene una mejor capacidad discriminativa que el puntaje B, mientras que valores negativos indican lo contrario. Para SL1, tanto el cNRI (cNRI = 0,088 (IC95%: 0,050,0,126), $p < 0,0001$) como el IDI (IDI = –0,048 (IC95%: –0,055, –0,041), $p < 0,0001$) fueron significativamente diferentes a cero. Para SL2, el cNRI fue significativamente diferente a cero (cNRI = 0,247 (IC95%: 0,209,0,285), $p < 0,0001$), mientras que el IDI fue cercano a cero (IDI = –0,001 (95%IC: –0,010, –0,008), $p = 0,80$). Cuando se compararon con la clasificación provista por el SAPS II actualizado, el cNRI e IDI fueron significativamente diferentes a cero para tanto SL1 como SL2: cNRI = 0,295 (IC95%: 0,257,0,333), $p < 0,0001$ e IDI

= 0,012 (IC95%: 0,008,0,017), p < 0,0001 para SL1; cNRI = 0,528 (IC95%: 0,415,0,565), p < 0,001 e IDI = 0,060 (IC95%: 0,056,0,065), p < 0,0001 para SL2. Cuando se compararon al puntaje APACHE II actualizado, el cNRI e IDI también fueron significativamente diferentes a cero para ambos SL1 y SL2: cNRI = 0,336 (IC95%: 0,298,0,374), p < 0,0001 e IDI = 0,029 (IC95%: 0,023,0,035), p < 0,0001 para SL1; cNRI = 0,561 (IC95%: 0,524,0,598), p < 0,001 e IDI = 0,076 (IC95%: 0,069,0,082) para SL2. Cuando se compararon a los nuevos puntaje SAPS II y APACHE II, ambos *Super Learner* reclasificaron una gran proporción de pacientes, especialmente de los estratos de alta probabilidad predicha a estratos inferiores.

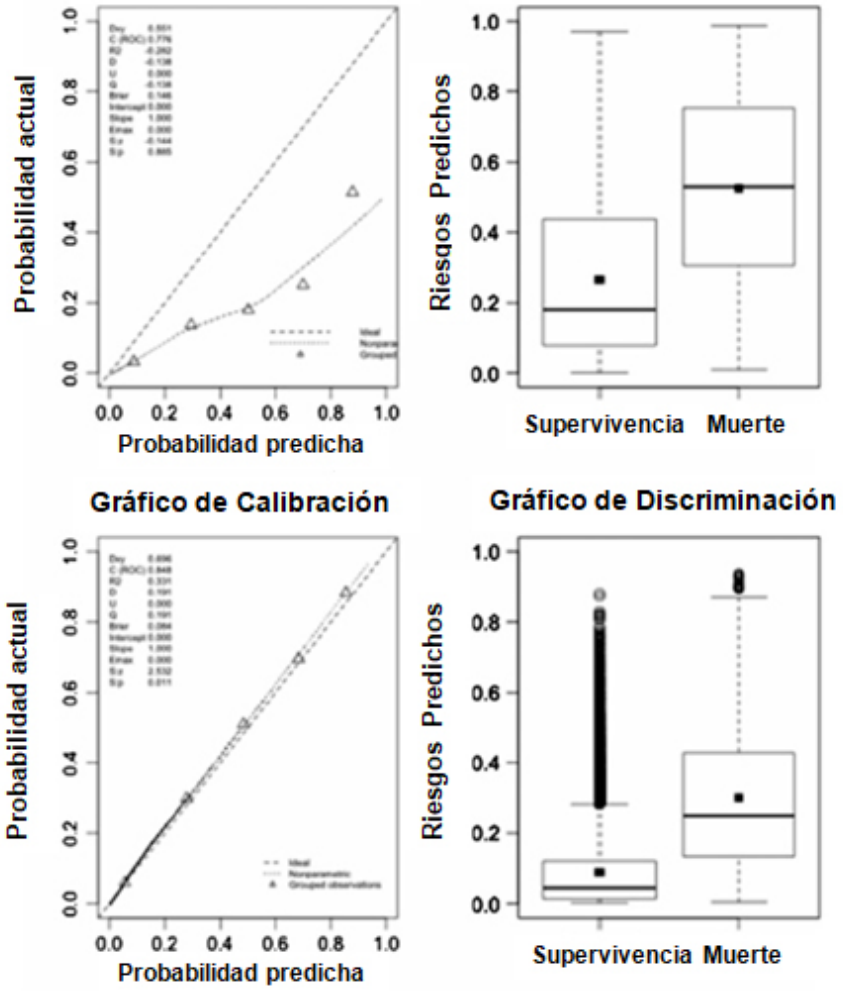

Fig. 20.4 Validación cruzada del error cuadrático medio para el Super Learner y los 12 algoritmos candidatos incluidos en la biblioteca. El panel superior muestra al Super Learner con variables categorizadas (Super Learner 1): error cuadrático medio (ECM) asociado con cada algoritmo candidato (figura superior) – Curvas ROC para cada algoritmo candidato (figura inferior); el panel inferior muestra al Super Learner con variables sin transformar (Super Learner 2): error cuadrático medio (ECM)

asociado con cada algoritmo candidato (figura superior) – Curvas ROC para cada algoritmo candidato (figura inferior).

Tabla 20.2 Tablas de reclasificación

	Modelo actualizado				
	0-0.25	0.25-0.5	0.5-0.75	0.75-1	% Reclasificación
Super Learner I					
Modelo Inicial: SAPS II original					
0-0.25	13.341	134	3	0	1%
0.25-0.5	4529	723	50	0	86%
0.5-0.75	2703	1090	174	2	96%
0.75-1	444	705	473	137	92%
Super Learner II					
Modelo Inicial: SAPS II original					
0-0.25	12.932	490	55	1	4%
0.25-0.5	4062	1087	142	11	79%
0.5-0.75	2531	1165	258	15	93%
0.75-1	485	775	448	51	97%
Super Learner I					
Modelo Inicial: SAPS II Nuevo					
0-0.25	20.104	884	30	2	4%
0.25-0.5	894	1426	238	9	44%
0.5-0.75	18	328	361	62	53%
0.75-1	1	14	71	66	57%
Super Learner II					

Modelo Inicial: SAPS II Nuevo					
0-0.25	19.221	1667	124	8	9%
0.25-0.5	765	1478	318	6	42%
0.5-0.75	24	346	367	32	52%
0.75-1	0	26	94	32	79%
Super Learner I					
Modelo Inicial: Nuevo APACHE II					
0-0.25	19.659	1140	107	6	6%
0.25-0.5	1262	1195	296	34	57%
0.5-0.75	89	298	264	71	63%
0.75-1	7	19	33	28	68%
Super Learner II					
Modelo Inicial: Nuevo APACHE II					
0-0.25	18.930	1764	200	18	9%
0.25-0.5	1028	1395	345	19	50%
0.5-0.75	50	333	309	30	57%
0.75-1	2	25	49	11	87%

20.5 Discusión

Los nuevos puntajes basados en el *Super Learner* mejoran la predicción de mortalidad hospitalaria en esta muestra, tanto para la discriminación como para la calibración, al ser comparado con los puntajes SAPS II y APACHE II. El puntaje de severidad *Super Learner* basado en variables sin transformar, también referido como SL2 ó SICULA, se encuentra disponible online a través de una aplicación web. Un resultado secundario importante es que la base de datos MIMIC II puede servir fácilmente y de forma fiable para desarrollar nuevos puntajes de gravedad para pacientes en UCI.

Nuestros resultados muestran la crucial ventaja del *Super Learner* que puede incluir tantos algoritmos candidatos como los introducidos por los investigadores, incluyendo algoritmos que reflejan el conocimiento científico disponible y la fortaleza de la diversidad de la librería. De hecho, la teoría indica que en grandes muestras el *Super Learner* se desempeña por lo menos tan bien como la (desconocida) elección óptima de la librería de algoritmos candidatos [28]. Esto se ilustra comparando el EMC-VC asociado con cada algoritmo incluido en la librería: SL1 consigue un desempeño similar al de BART, que es el mejor candidato en este caso, mientras que SL2 consigue un desempeño similar al de random forest, que en este caso superó a todos los otros candidatos. Por lo tanto, el *Super Learner* constituye una alternativa más flexible a otros métodos no paramétricos.

Dada la similitud en la calibración de los dos puntajes basados en *Super Learner* (SL1 y SL2), recomendamos utilizar *Super Learner* con variables explicativas sin transformar (SL2) en vistas de su mayor discriminación. Cuando se considera el riesgo de reclasificación, los dos algoritmos predictivos *Super Learner* tuvieron un cNRI similar, pero SL2 obtuvo un mejor IDI. Debe enfatizarse que, cuando se considera el IDI, el SL1 parece desempeñarse peor que el puntaje SAPS II. Sin embargo, el IDI debe ser utilizado cuidadosamente debido a que sufre de los mismos inconvenientes que la AUROC: resume las características de predicción de manera uniforme a lo largo de todos los umbrales de clasificación posibles pese a que muchos de estos nunca se considerarían en la práctica.

20.6 ¿Cuáles son los siguiente pasos?

El SICULA debe compararse con puntajes de severidad más recientes. Sin embargo, se ha reportado que dichos puntajes (por ejemplo, SAPS 3 y APACHE III) presentan los mismos inconvenientes que SAPS II [9,12,38]. Además, estos puntajes siguen siendo los más utilizados en la práctica [39]. Pese al hecho de que MIMIC II engloba datos de múltiples UCIs, la muestra sigue proviniendo de un único hospital y por lo tanto requiere de mayor validación externa. Sin embargo, los pacientes incluidos en la cohorte MIMIC-II parecen representativos de la población total de pacientes de UCI, como se refleja en que la tasa de mortalidad hospitalaria en la cohorte MIMIC II es similar a la reportada para pacientes de UCI durante ese mismo período de tiempo [40]. Por lo tanto, podemos esperar que nuestro puntaje

exhiba, en otras muestras, características de desempeño similares a las reportadas aquí, por lo menos en muestras extraídas de poblaciones de pacientes similares. Una gran representación en nuestra muestra de pacientes de UCO, que a menudo tienen puntajes de severidad menores a los de los pacientes de UCI médica o quirúrgica, puede haber limitado la aplicabilidad de nuestro puntaje en pacientes más críticamente enfermos. Finalmente, una hipótesis clave que justificó este estudio fue que la pobre calibración asociada a los puntajes de severidad actuales deriva del uso de modelos estadísticos que no son lo suficientemente flexibles más que de una selección inapropiada de variables a incluir en el modelo. Por esta razón y por el bien de proveer una adecuada comparación entre nuestro nuevo puntaje con el SAPS II, incluimos las mismas variables explicativas que las utilizadas en SAPS II. Expandir el conjunto de variables explicativas utilizadas podría resultar potencialmente en un puntaje con todavía mejor desempeño predictivo. En el futuro, ampliar el número de variables explicativas probablemente mejorará el desempeño predictivo del puntaje.

20.7 Conclusiones

Gracias a una gran colección de predictores potenciales y a una muestra lo suficientemente grande, la base de datos MIMIC II ofrece una oportunidad única para desarrollar y validar nuevos puntajes de severidad. En esta población, la predicción de mortalidad hospitalaria basada en el *Super Learner* consigue un desempeño significativamente mejorado, tanto en términos de calibración como discriminación, cuando se compararon los puntajes de severidad convencionales. El algoritmo predictivo SICULA es una alternativa promisoria que podría probarse valiosa en la práctica clínica y en la investigación. La validación externa de los resultados de este estudio en diferentes poblaciones (especialmente poblaciones fuera de los EE.UU.), la actualización periódica del algoritmo SICULA y la evaluación del beneficio potencial de incluir variables adicionales en el puntaje, siguen siendo importantes desafíos futuros que serán abordados en la segunda etapa del proyecto SICULA.

Nota: Las imágenes de este capítulo se encuentran disponibles a color en el ebook; disponible en www.hardineros.com

Apéndice: Código

Este caso de estudio utilizó códigos de la librería *Super Learner* implementados en R. Mayor detalle y otros códigos están disponibles en el repositorio GitHub que acompaña este libro: https://github.com/MIT-LCP/critical-data-book. Los siguientes algoritmos se incluyen en la librería de *Super Learner.*

Algoritmos paramétricos

– Regresión Logística: regresión logística estándar, incluyendo sólo términos principales para cada covariable e incluyendo términos de interacción [42] (SL. glm),
– Regresión Stepwise; regresión logística incluyendo procedimiento de selección de variables basado en Criterios de Información de Akaike [43] (SL. stepAIC),
– Modelos Aditivos Generalizados [43] (SL. gam): ,
– Modelos Lineales Generalizados con Máxima Verosimilitud penalizada [44] (SL. glmnet),
– Regresión Adaptativa Multivariada Spline polinomial [44] (SL. polymars),
– Modelo lineal Bayesiano generalizado [45] (SL. bayesglm).

Algoritmos no paramétricos

– Random Forest [46] (SL. randomForest),
– Redes Neuronales [47] (SL. nnet),
– Arboles de Clasificación Bagging [48] (SL. ipredbagg),

– Modelos de Regresión Generalizada Potenciados (Generalized boosted regression model) [49] (SL. gbm),
– Partición Recursiva Podada y Arboles de Regresión (Pruned Recursive Partitioning and Regression Trees) [50] (SL. rpartPrune),
– Arboles de Regresión Aditiva Bayesiana [51] (SL. bart).

Referencias

1. Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE (1981) APACHE-acutephysiology and chronic health evaluation: a physiologically based classification system. CritCare Med 9 (8): 591-597.
2. Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1985) APACHE II: a severity ofdisease classification system. Crit Care Med 13 (10): 818-829.
3. Le Gall JR, Loirat P, Alperovitch A, Glaser P, Granthil C, Mathieu D, Mercier P, Thomas R, Villers D (1984) A simplified acute physiology score for ICU patients. Crit Care Med 12 (11): 975-977.
4. Le Gall JR, Lemeshow S, Saulnier F (1993) A new simplified acute physiology score (SAPSII) based on a European/North American multicenter study. JAMA 270 (24): 2957-2963.
5. Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J (1993) Mortality probability models (MPM II) based on an international cohort of intensive care unit patients. JAMA 270 (20): 2478-2486.
6. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, Sirio CA, Murphy DJ, Lotring T, Damiano A (1991) The APACHE III prognostic system. Riskprediction of hospital mortality for critically ill hospitalized adults. Chest 100 (6): 1619-1636.
7. Le Gall JR, Neumann A, Hemery F, Bleriot JP, Fulgencio JP, Garrigues B, Gouzes C, Lepage E, Moine P, Villers D (2005) Mortality prediction using SAPS II: an update forFrench intensive care units. Crit Care 9 (6): R645-R652.
8. Nassar AP, Jr, Mocelin AO, Nunes ALB, Giannini FP, Brauer L, Andrade FM, Dias CA (2012) Caution when using prognostic models: a prospective comparison of 3 recentprognostic models. J Crit Care 27 (4), 423. e1-423. e7.
9. Poole D, Rossi C, Latronico N, Rossi G, Finazzi S, Bertolini G (2012) Comparison betweenSAPS II and SAPS 3 in predicting hospital mortality in a cohort of 103 Italian ICUs. Is newalways better? Intensive Care Med 38 (8): 1280-1288.
10. Metnitz B, Schaden E, Moreno R, Le Gall J-R, Bauer P, Metnitz PGH (2009) Austrianvalidation and customization of the SAPS 3 admissión score. Intensive Care Med 35 (4): 616-622.
11. Moreno RP, Metnitz PGH, Almeida E, Jordan B, Bauer P, Campos RA, Iapichino G, Edbrooke D, Capuzzo M, Le Gall J-R (2005) SAPS 3-From evaluation of the patient toevaluation of the intensive

care unit. Part 2: development of a prognostic model for hospitalmortality at ICU admissión. Intensive Care Med 31 (10): 1345-1355.

12. Beck DH, Smith GB, Pappachan JV, Millar B (2003) External validation of the SAPS II, APACHE II and APACHE III prognostic models in South England: a multicentre study. Intensive Care Med 29 (2): 249-256.
13. Aegerter P, Boumendil A, Retbi A, Minvielle E, Dervaux B, Guidet B (2005) SAPS IIrevisited. Intensive Care Med 31 (3): 416-423.
14. Ledoux D, Canivet J-L, Preiser J-C, Lefrancq J, Damas P (2008) SAPS 3 admissión score: anexternal validation in a general intensive care population. Intensive Care Med 34 (10): 1873-1877.
15. Dybowski R, Weller P, Chang R, Gant V (1996) Prediction of outcome in critically ill patientsusing artificial neural network synthesised by genetic algorithm. Lancet 347 (9009): 1146-1150.
16. Clermont G, Angus DC, DiRusso SM, Griffin M, Linde-Zwirble WT (2001) Predictinghospital mortality for patients in the intensive care unit: a comparison of artificial neuralnetworks with logistic regressión models. Crit Care Med 29 (2): 291-296.
17. Ribas VJ, López JC, Ruiz-Sanmartin A, Ruiz-Rodríguez JC, Rello J, Wojdel A, Vellido A (2011) Severe sepsis mortality prediction with relevance vector machines. Conf Proc IEEEEng Med Biol Soc 2011:100-103.
18. Kim S, Kim W, Park RW (2011) A comparison of intensive care unit mortality predictionmodels through the use of data mining techniques. Health Inform Res 17 (4): 232-243.
19. Foltran F, Berchialla P, Giunta F, Malacarne P, Merletti F, Gregori D (2010) Using VLADscores to have a look insight ICU performance: towards a modelling of the errors. J Eval ClinPract 16 (5): 968-975.
20. Gortzis LG, Sakellaropoulos F, Ilias I, Stamoulis K, Dimopoulou I (2008) Predicting ICUsurvival: a meta-level approach. BMC Health Serv Res 8:157-164.
21. Dudoit S, Van Der Laan MJ (2003) Asymptotics of cross-validated risk estimation inestimator selection and performance assessment. Statistical Methodology 2 (2): 131-154.
22. Lee J, Scott DJ, Villarroel M, Clifford GD, Saeed M, Mark RG (2011) Open-access MIMIC-IIdatabase for intensive care research. Conf Proc IEEE Eng Med Biol Soc 2011:8315-8318.
23. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L-W, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG (2011) Multiparameter intelligent monitoring in intensive care II: apublic-access intensive care unit database. Crit Care Med 39 (5): 952-960.
24. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101 (23): E215-E220.
25. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ (2015) Mortality prediction in intensive care units with the super ICU learner algorithm (SICULA): apopulation-based study. Lancet Respir Med 3 (1).

26. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, Reinhart CK, Suter PM, Thijs LG (1996) The SOFA (Sepsis-related Organ Failure Assessment) score todescribe organ dysfunction/failure. On behalf of the working group on sepsis-relatedproblems of the European Society of Intensive Care Medicine. Intensive Care Med 22 (7): 707-710.
27. Van Der Laan MJ, Dudoit S (2003) Unified cross-validation methodology for selectionamong estimators and a general cross-validated adaptive epsilon-net estimator: finite sampleoracle inequalities and examples.U.C. Berkeley División of Biostatistics Working PaperSeries, Working Paper, no 130, pp 1-103.
28. van der Laan MJ, Polley EC, Hubbard AE (2007) Super learner. Stat Appl Genet Mol Biol6:25.
29. Cox DR (1958) Two further applications of a model for binary regressión. Biometrika 45 (3/4): 562-565.
30. Harrison DA, Brady AR, Parry GJ, Carpenter JR, Rowan K (2006) Recalibration of riskprediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom. Crit Care Med 34 (5): 1378-1388.
31. Kramer AA, Zimmerman JE (2007) Assessing the calibration of mortality benchmarks incritical care: the Hosmer-Lemeshow test revisited. Crit Care Med 35 (9): 2052-2056.
32. Bertolini G, D'Amico R, Nardi D, Tinazzi A, Apolone G (2000) One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regressión model. J Epidemiol Biostat 5 (4): 251-253.
33. Miller ME, Hui SL, Tierney WM (1991) Validation techniques for logistic regressión models. Stat Med 10 (8): 1213-1226.
34. Cook NR (2007) Use and misuse of the receiver operating characteristic curve in riskprediction. Circulation 115 (7): 928-935.
35. Cook NR (2008) Statistical evaluation of prognostic versus diagnostic models: beyond theROC curve. Clin Chem 54 (1): 17-23.
36. Pencina MJ, D'Agostino RB, Sr, D'Agostino RB, Jr, Vasan RS (2008) Evaluating the addedpredictive ability of a new marker: from area under the ROC curve to reclassification andbeyond. Stat Med 27 (2): 157-172; discussión 207-212, Jan 2008.
37. Greenland S (2008) The need for reorientation toward cost-effective prediction: comments on'Evaluating the added predictive ability of a new marker: From area under the ROC curve toreclassification and beyond' by M.J. Pencina et al., Statistics in Medicine 10.1002/sim.2929. Stat Med 27 (2): 199-206.
38. Sakr Y, Krauss C, Amaral ACKB, Réa-Neto A, Specht M, Reinhart K, Marx G (2008) Comparison of the performance of SAPS II, SAPS 3, APACHE II, and their customizedprognostic models in a surgical intensive care unit. Br J Anaesth 101 (6): 798-803.

39. Rosenberg AL (2002) Recent innovations in intensive care unit risk-prediction models. CurrOpin Crit Care 8 (4): 321-330.
40. Zimmerman JE, Kramer AA, Knaus WA (2013) Changes in hospital mortality for UnitedStates intensive care unit admissions from 1988 to 2012. Crit Care 17 (2): R81.
41. Van der Laan MJ, Rose S (2011) Targeted learning: causal inference for observational andexperimental data. Springer, Berlin.
42. McCullagh P, Nelder JA (1989) Generalized linear models, vol 37. Chapman & Hall/CRC.
43. Venables WN, Ripley BD (2002) Modern applied statistics with S. Springer, Berlin.
44. Friedman JH (1991) Multivariate adaptive regressión splines. Ann Stat 1-67.
45. Gelman A, Jakulin A, Pittau MG, Su YS (2008) A weakly informative default priordistribution for logistic and other regressión models. Ann Appl Stat 1360-1383.
46. Breiman L (2001) Random forests. Mach Learn 45 (1): 5-32.
47. Ripley BD (2008) Pattern recognition and neural networks. Cambridge university press, Cambridge.
48. Breiman L (1996) Bagging predictors. Mach Learn 24 (2): 123-140.
49. Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regressión trees. J AnimEcol 77 (4): 802-813.
50. Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regressión trees. Chapman & Hall, New York.
51. Chipman HA, George EI, McCulloch RE (2010) BART: Bayesian additive regressión trees. Ann Appl Stat 4 (1): 266-298.

Capítulo 21

PREDICCIÓN DE MORTALIDAD EN LA UCI

Joon Lee, Joel A Dubin y David M. Maslove

Objetivos de aprendizaje

Construir y evaluar modelos de predicción de mortalidad:

1. Aprender como extraer variables predictoras de MIMIC-II
2. Aprender como construir una regresión logística, máquinas de vector soporte (MVS) y árboles de decisión para la predicción de mortalidad
3. Aprender cómo utilizar el algoritmo Boosting Adaptativo para mejorar el desempeño de un clasificador débil
4. Aprender cómo crear y evaluar modelos predictivos utilizando validación cruzada

21.1 Introducción

Los pacientes que ingresan a la unidad de cuidados intensivos (UCI) poseen lesiones o enfermedades críticas y presentan un riesgo alto de morir. La mortalidad en la UCI difiere ampliamente dependiendo de la enfermedad subyacente, con tasas tan bajas como 1 en 20 para cirugías electivas así como 1 en 4 para pacientes con patología respiratoria [1]. El riesgo de morir puede estimarse evaluando la severidad de la enfermedad del paciente, determinada por factores fisiológicos, clínicos y demográficos.

En la práctica clínica, la estimación del riesgo de morir puede ser útil en el triaje y asignación de recursos, para determinar los niveles de atención apropiados e, incluso, para discutir con los pacientes y sus familias los resultados esperados. Sin embargo, las estimaciones de mortalidad se basan en estudios de grandes poblaciones heterogéneas por lo que no puede asegurarse la validez de la misma en un paciente en particular.

Esta deficiencia puede ser mitigada con estimaciones de riesgo de mortalidad personalizados, los cuales se discuten bien en [2,3], pero no son objeto de discusión del presente estudio.

Quizás los usos más notables de los predictores de mortalidad en la UCI se encuentran en las áreas de investigación y administración de la salud, que a menudo implican observar cohortes de pacientes en estado crítico.

Tradicionalmente, se acepta la aplicación de los modelos de predicción de mortalidad en estos estudios poblacionales teniendo en cuenta que los modelos predictivos se basan en cohortes. En este contexto, la predicción de mortalidad es usada para comparar el promedio de severidad de enfermedad entre grupos de pacientes críticamente enfermos (por ejemplo entre pacientes en diferentes UCIs, hospitales o sistemas de salud) y entre grupos enrolados en ensayos clínicos. La mortalidad predicha puede compararse con las tasas de mortalidad observadas con fines de evaluación comparativa y evaluación de performance de UCIs y sistemas de salud.

Una serie de puntajes de severidad de enfermedad (PSE) han sido introducidos en la UCI para predecir resultados, incluyendo la muerte. Estos incluyen los puntajes APACHE [4], SAPS (Simplified Acute Physiology Score) [5], MPM (Mortaly Probability Model) [6] y SOFA (Sequential Organ Failure Assessment) [7]. Estos puntajes se desempeñan bien, presentando áreas bajo la curva de Característica Operativa del Receptor (ROC) (AUROCs) típicamente entre 0.8 y 0.9 [5, 6, 8]. En la actualidad, la investigación explora formas de aprovechar la mejoría en la completitud y expresividad de las historias clínicas electrónicas (HCEs) para mejorar la precisión de las predicciones.

En particular, la naturaleza granular de la HCE (es decir, un nutrido conjunto de variables clínicas registradas con alta resolución temporal) puede conducir a la creación de un modelo predictivo personalizado para un paciente determinado al identificar y utilizar datos de pacientes similares.

21.2 Set de Datos de Estudio

Este caso de estudio tiene como objetivo crear modelos predictivos de mortalidad utilizando la primera admisión a la UCI de todos los pacientes adultos, disponible en la base de datos MIMIC-II versión 2.6. En la tabla *icustay_detail*, los pacientes adultos de MIMIC-II pueden ser identificados mediante *icustay_age_group='adult'*, mientras que la primera admisión a UCI para cada paciente puede seleccionarse a través de *subject_icustay_seq=l*. Además se excluyeron todas las estadías en UCI con un valor nulo en *icustay_id* debido a que *icustay_id* se utilizó para encontrar en otras tablas los datos de las estadías en UCI incluidas en el estudio. Un

total de 24581 admisiones en UCI en MIMIC-II cumplieron los criterios de inclusión.

Se extrajeron las siguientes variables administrativas/demográficas para ser usadas como predictores: edad al momento de la admisión en UCI, género, tipo de admisión (electiva, urgencia, emergencia) y el tipo de UCI primaria de admisión. Además, se extrajeron como predictores las primeras mediciones en la UCI de los siguientes signos vitales y exámenes de laboratorio: frecuencia cardíaca, presión arterial sistólica y media (combinando mediciones invasivas y no invasivas), temperatura corporal, SpO2, frecuencia respiratoria, creatinina, potasio, sodio, cloro, bicarbonato, hematocrito, recuento de glóbulos blancos, glucosa, calcio, fósforo y lactato.

A pesar de que las variables extraídas fueron las primeras mediciones de UCI, el tiempo exacto con respecto al momento del ingreso puede variar entre pacientes. Además, este enfoque para la extracción de datos variable por variable no garantiza mediciones concurrentes en cada paciente.

Sin embargo, para la amplia mayoría de las admisiones a la UCI en MIMIC-II las mediciones de estas variables clínicas habituales fueron obtenidas al inicio de la admisión en UCI, o al menos dentro de las primeras 24 hrs.

Se extrajo como resultado a predecir la mortalidad a los 30 días del alta hospitalaria. En MIMIC-II, este resultado binario puede obtenerse comparando la fecha de muerte (encontrada en la tabla *d_patients*) con la fecha de alta del hospital (encontrada en la tabla *icustay_detail*).

Si nuestro enfoque hubiera sido la identificación de la muerte posterior al alta en un período de tiempo mayor, habríamos extraído la fecha de mortalidad para intentar predecir el tiempo de sobrevida.

21.3 Preprocesamiento

Algunas de las variables extraídas requieren un procesamiento adicional antes de que puedan usarse para el modelado predictivo. En MIMIC-II, algunas edades son irrealmente grandes (~ 200 años) ya que se insertaron intencionalmente para enmascarar las edades reales de aquellos pacientes que tenían 90 años o más y aún estaban vivos (según los últimos datos del índice de defunción de la seguridad social), por ser información de salud protegida. Para estos pacientes, se sustituyó la mediana de tales edades

enmascaradas (concretamente 91,4). Además, en relación al tipo de UCI, UCIF (Finard UCI, término específico del Centro Médico Beth Israel Deaconess, de donde los datos de MIMIC-II fueron recolectados) se suplantó por UCIM (UCI médica) ya que sólo existe un pequeño número de admisiones en UCIF en el MIMIC-II y por otro lado UCIF no es otra cosa que una UCIM especial.

En MIMIC-II hay muchos datos faltantes. Aunque existen maneras de utilizar las admisiones en UCI con datos incompletos (por ej. imputación), este estudio de caso excluye los casos con datos incompletos ya que este tema es discutido en profundidad en el capítulo 13 de este libro. Luego de la exclusión de casos con datos incompletos, solo quedaron 9269 admisiones en UCI. De todas formas, esta sigue siendo una muestra suficiente para realizar el presente caso de estudio. Si se requiere un tamaño de muestra mayor, deberían considerarse enfoques como la imputación y / o exclusión de variables con datos faltantes.

Utilizando la configuración de R predeterminada, las variables numéricas normalmente se importan correctamente con un manejo adecuado de los datos faltantes (marcados como NA), pero puede ser necesario tener especial cuidado para importar variables categóricas. Para evitar que el campo vacío se importe como una categoría por sí sola, en este caso de estudio (1) se importaron las variables categóricas como cadena de caracteres o "*strings*" (2), se convirtieron todos los campos vacíos a NA y luego (3) se convirtieron las variables categóricas en factores. Este caso de estudio incluye las siguientes variables categóricas: género, tipo de admisión, tipo de UCI y mortalidad a los 30 días.

21.4 Métodos

Se emplearon los siguientes modelos predictivos: regresión logística (RL), máquinas de vectores soporte (MVS) y árboles de decisión (AD). Estos modelos fueron elegidos debido a su uso generalizado en el aprendizaje automático. Aunque el lector debería consultar los capítulos apropiados en la Parte 2 del libro para obtener más información sobre estos modelos, aquí se proporciona una breve descripción de cada uno.

La RL es un modelo que puede aprender la relación matemática, dentro de un marco restringido, utilizando una función logística entre un set de

covariables (por ej. en este caso de estudio, variables predictoras) y un resultado binario variable (por ej. en este estudio de caso, mortalidad). Una vez que se ha aprendido esta relación, el modelo puede realizar predicciones para un nuevo caso dado. La RL se usa ampliamente en investigación en salud gracias a la facilidad de su interpretación.

La técnica de MVSs es similar a la RL en cuanto a que puede clasificar (o predecir) un caso dado en términos de un resultado, pero esto lo realiza creando un límite de decisión óptimo en el espacio de datos donde las dimensiones son las covariables y donde se grafican todos los puntos de datos disponibles. En otras palabras, los MVSs intentan trazar un límite de decisión que coloque tantos casos negativos (sobrevivientes) como sea posible en un lado del límite y tantos casos positivos (muertos) como sea posible en el otro lado.

Por último, los ADs tienen la estructura de un árbol, que consiste en una jerarquía a través de nodos de decisión. Cada nodo de decisión conduce a dos ramas, dependiendo del valor de la covariable en particular (por ej. edad >65 ó no). Cada caso sigue su camino a través de las distintas ramas hasta llegar a un nodo terminal, el cual se asocia con un resultado en particular. Los algoritmos de aprendizaje de ADs aprenden en forma automática el árbol de decisión óptimo a partir de un set de datos.

También intentamos mejorar el rendimiento predictivo del AD mediante la aplicación de refuerzo adaptativo, ej. AdaBoost [9]. AdaBoost puede mejorar los modelos predictivos débiles de forma efectiva al construir un conjunto de modelos que se centren de forma progresiva en los casos que el modelo anterior predice de manera imprecisa. En otras palabras, AdaBoost nos permitió construir una serie de AD donde los que se construyeron al final fueron expertos en los casos más desafiantes. En AdaBoost, la predicción final es el promedio de las predicciones de los modelos individuales.

Para hacer correr los códigos en R, deben instalarse los siguientes paquetes de R vía *install. packages (): e1071*, *ada*, *rpart* y *ROCR*. Las funciones de entrenamiento para RL, MVS y AD son *glm (), svm (),* and *rpart (),* respectivamente. Para todos los modelos se utilizaron los parámetros predeterminados.

Para entrenamiento y testeo, se utilizó validación cruzada de 10 iteraciones. Bajo dicho esquema, las admisiones en la UCI incluidas en el

caso de estudio se dividieron aleatoriamente en 10 grupos de tamaño similar (los subconjuntos o “folds”). El procedimiento rotó a través de los 10 subconjuntos para entrenar modelos predictivos basados en 9 subconjuntos (datos de entrenamiento) y probados en el subconjunto restante (datos de prueba), hasta que cada subconjunto se haya utilizado como datos de prueba.

El desempeño predictivo se midió utilizando AUROC, que es una medida de desempeño ampliamente utilizada para la clasificación binaria. Para cada modelo predictivo, se calculó el AUROC de cada subconjunto de validación cruzada. En el código de R proporcionado, se llama a la función *comp. auc ()* para calcular el AUROC dado un conjunto de probabilidades predichas de un modelo y los datos de mortalidad reales correspondientes.

21.5 Análisis

Las siguientes fueron las AUROC de los modelos predictivos (mostradas como media [desviación estándar]): RL ---- 0.790 [0.015]; MVS-0,782 [0,014]; AD-0.616 [0.049]; AdaBoost ---- 0.801 [0.013]. Por lo tanto, en términos de AUROC promedio, AdaBoost resultó en el mejor rendimiento, mientras que AD fue claramente el peor modelo predictivo. AD fue solo moderadamente mejor que adivinar al azar (lo que correspondería a una AUROC de 0.5) y, como resultado, puede considerarse un aprendiz débil (“weak learner”). Nótese que AdaBoost pudo mejorar sustancialmente los AD, lo que es consistente con su conocida capacidad de mejorar efectivamente a los aprendices débiles. Debido a la partición aleatoria de datos de validación cruzada, se producirán resultados ligeramente diferentes cada vez que se ejecute el código R proporcionado. El uso de *set. seed ()* en R puede forzar la generación de números aleatorios en la muestra () y hacer que los resultados sean reproducibles, pero esto no se usó en este caso de estudio para obtener una evaluación más robusta de los resultados.

Como comparación, un estudio previo [2] informó AUROCs promedios de 0.658 (intervalo de confianza (IC) 95%: [0.648,0.668]) y 0.633 (IC 95%: [0.624,0.642]) para SAPS I y SOFA, respectivamente, para predecir la mortalidad a 30 días para 17.152 estadías en la UCI de adultos en MIMIC-II, a pesar de que esa cohorte fue algo diferente de la de este caso de estudio. Algunos PSE más avanzados, como APACHE IV, habrían alcanzado un

rendimiento comparable o mejor que los modelos predictivos investigados en este caso de estudio (solo SAPS I y SOFA están disponibles en MIMIC-II), pero hay que tener en cuenta que esos PSE avanzados tienden a utilizar un conjunto de predictores mucho más completo que los utilizados en este caso de estudio.

21.6 Visualización

La figura 21.1 muestra el rendimiento de los modelos predictivos en un gráfico de cajas (*boxplot).* Es visualmente evidente que AdaBoost, RL y MVS presentaron un desempeño similar, mientras que AD produjo no solo el peor desempeño sino también la mayor variabilidad en AUROC, lo que demuestra su sensibilidad a la partición aleatoria de datos en la validación cruzada.

La figura 21.2 es una visualización interesante de los resultados de la predicción, donde cada círculo representa a un paciente y el color del círculo indica el resultado de la predicción (correcta o incorrecta) del paciente. Se agregó una dosis de ruido horizontal ("jitter") a cada punto (esto simplemente significa que se aplicó un pequeño cambio aleatorio al valor x de cada punto) para reducir la superposición de puntos.

Se muestran los resultados de predicción de solo una de las diez iteraciones de validación cruzada, con un umbral de 0.5 (seleccionado arbitrariamente; el lector puede estar interesado en estudiar cómo este umbral afecta esta cifra) aplicado a los riesgos de mortalidad estimados de los modelos predictivos (utilizando la función *th. pred ()* en R).

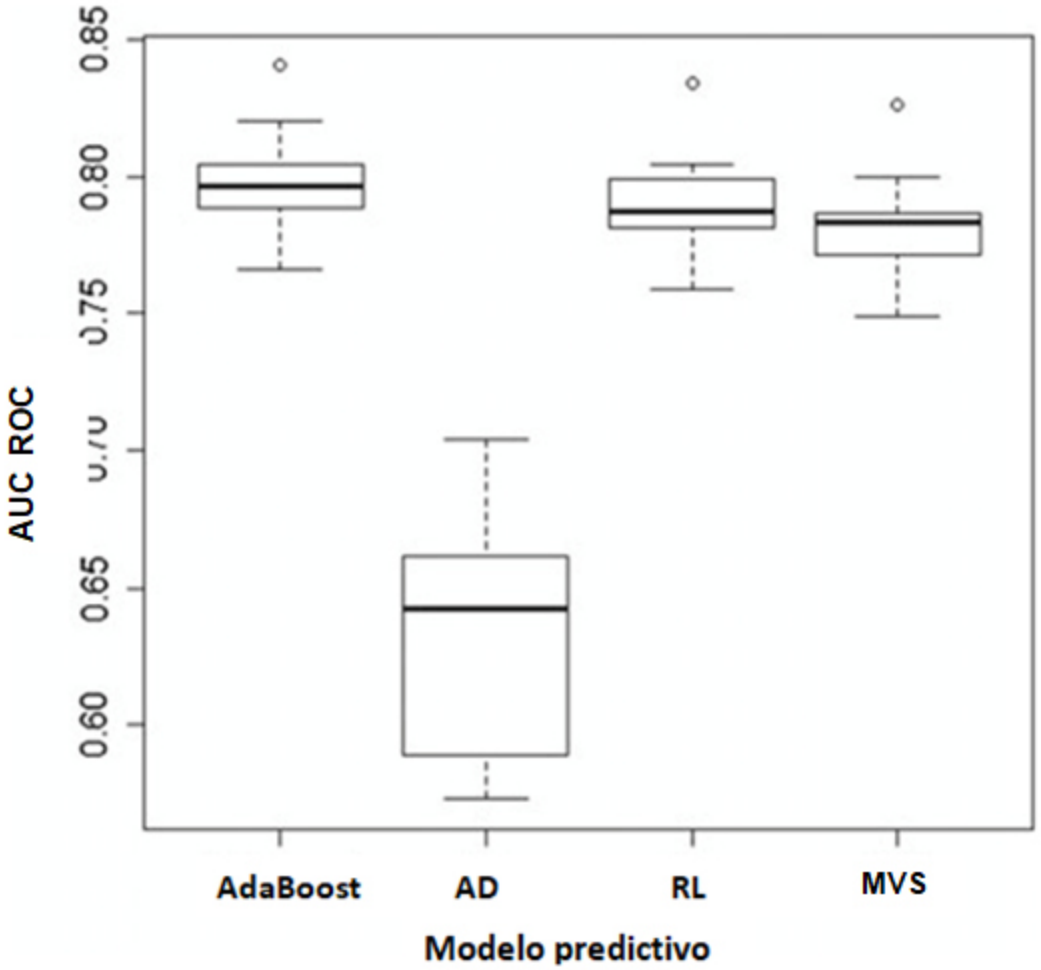

Fig 21.1 Un diagrama de cajas y bigotes que muestra el desempeño en la predicción de mortalidad de distintos modelos predictivos de la validación cruzada de 10 iteraciones. AD= Árbol de decisión, RL= Regresión logística, MVS= máquinas de vectores de soporte.

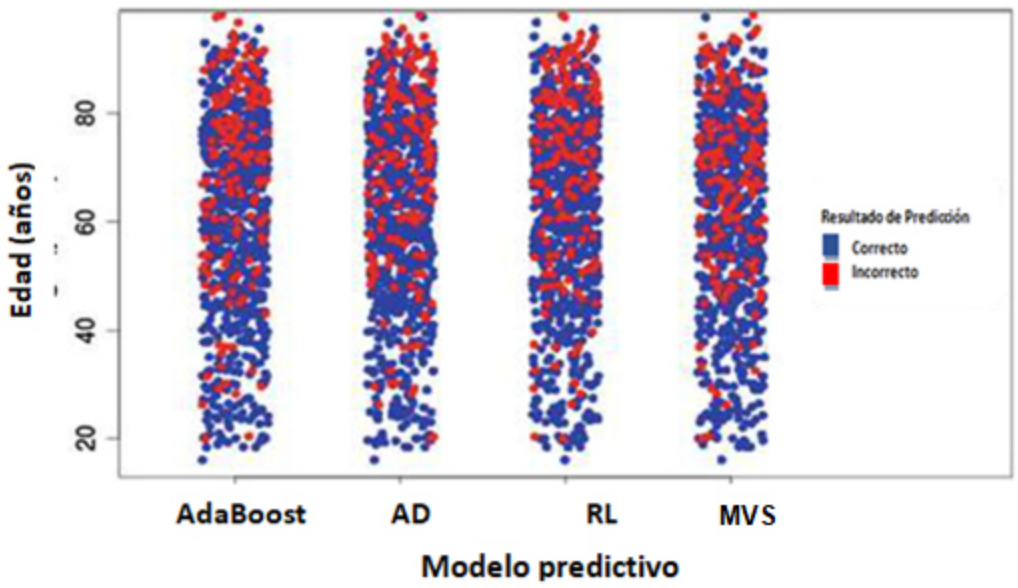

Fig. 21.2 Resultados de la predicción para pacientes individuales como una función de la edad, estratificado por el modelo de predicción. Se grafican aquí solamente los resultados de uno de las 10 iteraciones de validación cruzada.
AD= Árbol de decisión, RL= Regresión logística, MVS= máquinas de vectores de soporte

La figura 21.2 muestra los resultados de la predicción en función de la edad, pero la variable en el eje y se puede cambiar fácilmente a otra variable de interés (por ejemplo, frecuencia cardíaca, creatinina).

Una observación que está clara en la Fig. 21.2 pero no en la Fig. 21.1 es que la precisión predictiva es mayor para pacientes más jóvenes (por ejemplo, <40 años) que para pacientes mayores, en todos los modelos predictivos. Esto se debe probablemente al hecho de que la tasa de mortalidad es mucho más baja entre los pacientes más jóvenes que en los pacientes mayores, y los modelos predictivos pueden lograr mayor precisión si están sesgados hacia predecir riesgos de mortalidad bajos (sin embargo, esto conduciría a una baja sensibilidad). Por lo tanto, es importante tener en cuenta que aunque la Fig. 21.2 transmite una sensación de precisión general, no revela sensibilidad, especificidad, valor predictivo positivo ni valor predictivo negativo.

21.7 Conclusiones

Utilizando datos clínicos y demográficos de la base de datos MIMIC II, este estudio de caso utilizó algoritmos de aprendizaje automático para

clasificar a los pacientes como vivos o muertos a los 30 días después del alta hospitalaria. Los resultados fueron comparables a los obtenidos por los puntajes de severidad de enfermedad que actualmente se encuentran más en uso. Sin embargo, a diferencia de estos puntajes, los algoritmos de aprendizaje utilizados no tenían acceso a diagnósticos y procedimientos específicos, lo que puede agregar un poder predictivo considerable. Sin embargo, una ventaja de usar solo datos clínicos y demográficos es que están disponibles de manera más rutinaria y, como resultado, los modelos predictivos basados en ellos se pueden usar más ampliamente. Además, nuestros algoritmos se aplicaron a una población indiferenciada de pacientes críticos, en lugar de adaptarse a grupos específicos como post cirugía cardiovascular (es decir, pacientes de la unidad de recuperación de cirugía cardíaca), lo que también ha demostrado mejorar el desempeño predictivo [3]. El éxito de la predicción visto en este caso de estudio probablemente refleja el poder de los algoritmos de aprendizaje utilizados, así como la utilidad tanto del tamaño como de la granularidad de la base de datos estudiada.

Una perspectiva útil que aprovecha la naturaleza dinámica de los datos de la HCE es el potencial para actualizar los datos de entrenamiento y los modelos de predicción a medida que se vuelven disponibles datos clínicos más recientes.

Teóricamente, esto conduciría a sistemas de puntuación igualmente dinámicos que generan predicciones más precisas al reflejar las prácticas actuales. Se hace evidente una negociación entre el uso de los datos más actuales, que es probable que sean los más representativos, y la inclusión de datos más antiguos, que pueden ser menos relevantes pero que proporcionan un mayor poder estadístico.

21.8 Próximos pasos

Aunque las AUROC cercanas a 0,8 representan un buen rendimiento, el hecho de que la RL, los MVS y AdaBoost produjeran un rendimiento similar puede implicar que el rendimiento podría estar limitado por las variables predictoras en lugar de la selección del modelo. Futuros estudios podrían investigar más a fondo la selección de predictores o las diferentes representaciones de las mismas variables (por ejemplo, patrones

temporales en lugar de mediciones en un punto de tiempo específico; consulte el capítulo Selección de hiperparámetros de la Parte 3).

Dado que fueron utilizados los parámetros predeterminados de R para RL, MVS, AD y AdaBosst, un próximo paso que pareciera razonable es investigar cómo se afecta la performance de la predicción al modificar los parámetros. Consulte la ayuda de R o la documentación del paquete R correspondiente para obtener más información sobre los parámetros del modelo.

Para mejorar el rendimiento predictivo, hemos considerado previamente un enfoque de predicción de mortalidad personalizado en el que sólo se utilizan los datos de pacientes que son similares a un paciente índice (para quienes se debe realizar la predicción) para entrenar modelos predictivos personalizados [2]. Utilizando una métrica particular de similitud de pacientes basada en la coseno-similitud y RL, el AUROC máxima que informó este estudio fue de 0,83. A la luz de este prometedor resultado, se invita al lector a seguir enfoques personalizados similares con nuevas métricas de similitud de pacientes.

Los métodos bayesianos [10] ofrecen otro paradigma de predicción que puede valer la pena investigar. Los métodos bayesianos logran un equilibrio entre la experiencia en el tema (para la predicción de mortalidad en la UCI, esto correspondería a la experiencia clínica con respecto al riesgo de mortalidad) y la evidencia empírica de los datos clínicos. Dado que los modelos de aprendizaje automático discutidos en este capítulo fueron puramente empíricos, la adición explícita de experiencia clínica a través del paradigma bayesiano puede mejorar potencialmente el rendimiento predictivo.

Además del AUROC, hay otras formas de evaluar el rendimiento predictivo, incluida la puntuación de Brier escalada. Ver [11] para más información. Una vez que se aplica un umbral al riesgo de mortalidad previsto, también se pueden calcular medidas de rendimiento más convencionales, como precisión, sensibilidad, especificidad, etc. Dado que cada medida de rendimiento tiene ventajas y desventajas (por ejemplo, si bien el AUROC proporciona una evaluación más completa que la simple precisión, se vuelve sesgada para los conjuntos de datos asimétricos [12]), puede ser mejor calcular una variedad de medidas para una evaluación integral del rendimiento predictivo.

Por último, la calidad de los datos a menudo se pasa por alto, pero desempeña un papel importante en la determinación de qué rendimiento predictivo es posible con un conjunto determinado de datos. Este es un problema particularmente crítico con los datos retrospectivos de la HCE, cuyo registro puede haber tenido controles mínimos de calidad. La implementación de controles de calidad de datos más rigurosos (p. ej., valores atípicos, factibilidad fisiológica) antes del entrenamiento del modelo predictivo es un punto importante para futuras etapas.

21.9 Conexiones

Si bien este capítulo se centró en la predicción de la mortalidad, la extracción de datos y las técnicas analíticas discutidas aquí son ampliamente aplicables a la predicción de otros resultados discretos (por ejemplo, reingreso hospitalario) y continuos (por ejemplo, duración de la estadía) del paciente. Además, los matices relacionados con MIMIC-II, como el manejo de edades cercanas a los 200 años y el tipo de servicio FICU, son cuestiones importantes para cualquier estudio en MIMIC-II.

Los modelos (RL, AD, MVS) y técnicas (validación cruzada, AdaBoost, AUROC) de aprendizaje automático se usan ampliamente en una variedad de aplicaciones de predicción, detección y minería de datos, no solo en medicina sino también más allá de ella. Además, dado que R es uno de los lenguajes de programación más populares en ciencia de datos, se vuelve invaluable ser capaz de manipular los datos de la HCE y aplicar el aprendizaje automático en R.

Nota: Las imágenes de este capítulo se encuentran disponibles a color en el ebook; disponible en www.hardineros.com

Apéndice: Códigos

El código utilizado en este caso práctico está disponible en el repositorio de GitHub que acompaña este libro: https://github. corn/MIT-LCP/critical-data-book. En este sitio web se encuentra disponible más información sobre el código. El lector puede reproducir el presente caso de estudio ejecutando los siguientes códigos SQL y R:

query. sql: se utiliza para extraer datos de la base de datos MIMIC II.

analysis. R: se utiliza para realizar el procesamiento de datos.

Referencias

1. Kuzniewicz MW, Vasilevskis EE, Lane R, Dean ML, Trivedi NG, Rennie DJ, Clay T, Kotler PL, Dudley RA (2008) Variation in ICU risk-adjusted mortality: impact of methods of assessment and potential confounders. Chest 133 (6): 1319-1327.
2. Lee J, Maslove DM, Dubin JA (2015) Personalized mortality prediction driven by electronic medical data and a patient similarity metric. PLoS ONE 10 (5): e0127428.
3. Lee J, Maslove DM (2015) Customization of a severity of illness score using local electronic medical record data.J. Intensive Care Med, 0885066615585951.
4. Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1985) APACHE II: a severity of disease classification system. Crit Care Med 13 (10): 818-829.
5. Legall JR, Lemeshow S, Saulnier F (1993) A new simplified acute physiology score (SAPS-II) based on a european north-american multicenter study. Jama-J Am Med Assoc 270:2957-2963.
6. Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J (1993) Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. JAMA 270 (20): 2478-2486.
7. Vincent J, Moreno R, Takala J, Willatts S, De Mendonca A, Bruining H, Reinhart C, Suter P, Thijs L (1996) The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. Intensive Care Med 22 (7): 707-710.
8. Gursel G, Demirtas S (2006) Value of APACHE II, SOFA and CPIS scores in predicting prognosis in patients with ventilator-associated pneumonia. Respiration. 73 (4): 503-508.
9. Freund Y, Schapire R (1995) A desicion-theoretic generalization of on-line learning and an application to boosting. Comput Learn Theory 55 (1): 119-139.

10. Gelman A, Carlin JB, Stern HS, Rubin DB (2014) Bayesian data analysis, vol 2. Taylor &Francis, UK.
11. Wu YC, Lee WC (2014) Alternative performance measures for prediction models. PLoS One9 (3).
12. Davis J, Goadrich M (2006) The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning-ICML'06, pp 233–240.

Capítulo 22

TÉCNICA DE FUSIÓN DE DATOS PARA ALERTA TEMPRANA DE DETERIORO CLÍNICO

Peter H. Charlton, Marco Pimentel y Sharukh Lokhandwala

Objetivos de Aprendizaje

Diseñar y evaluar algoritmos para puntajes de alerta temprana (EWS, por sus siglas en inglés Early Warning Scores) que fusionan los signos vitales con parámetros fisiológicos adicionales comúnmente disponibles en las historias clínicas electrónicas (HCE) hospitalarias.

1. Extraer variables fisiológicas, demográficas y bioquímicas de la base de datos MIMIC II.
2. Extraer resultados de pacientes de la base de datos MIMIC II.
3. Preparar los datos de la HCE para su análisis en Matlab®.
4. Diseñar algoritmos de fusión de datos en Matlab ®.
5. Comparar el desempeño de los algoritmos de fusión de datos.

22.1 Introducción

Los pacientes hospitalizados con enfermedades agudas corren el riesgo de sufrir complicaciones clínicas como infección, insuficiencia cardíaca congestiva y paro cardíaco [1]. La detección temprana y el manejo de tales complicaciones pueden mejorar los resultados del paciente y reducir la utilización de recursos de salud [2,3]. Actualmente, los puntajes de alerta temprana (EWS) se utilizan para ayudar en la identificación de pacientes con deterioro. Los EWS se diseñaron para usarse junto a la cama del paciente: se pueden calcular manualmente y los datos requeridos (signos vitales) se pueden medir fácilmente en el lugar. Ahora que las HCE se están generalizando en la atención hospitalaria, existe la posibilidad de desarrollar EWSs mejorados utilizando algoritmos más complejos calculados por computadora e incorporando datos fisiológicos adicionales de la HCE.

La mayoría de los métodos para la detección de deterioro clínico se basan en el supuesto de que los cambios en la fisiología se manifiestan en las etapas tempranas del deterioro. Esta suposición está bien documentada por

Schein y col. en 1990, quienes encontraron que el 84% de los pacientes "tenía observaciones documentadas de deterioro clínico o nuevos problemas" en las ocho horas previas al paro cardíaco [4]. Esto fue respaldado por un estudio realizado por Franklin y col. [5]

También se han observado anormalidades fisiológicas previas a otros eventos, como por ejemplo ingreso no planificado a la Unidad de Cuidados Intensivos (UCI) [6] y muerte prevenible [7]. La evidencia de deterioro se puede observar 8-12 hs antes de los eventos principales [8, 9].

Se propuso que la incidencia de deterioro clínico podría reducirse al reconocer y responder a los primeros cambios en la fisiología [10-12]. Posteriormente, los EWSs se desarrollaron para permitir el reconocimiento oportuno de pacientes con riesgo de deterioro. Los EWSs son puntajes calculados a partir de un conjunto de parámetros fisiológicos medidos de forma rutinaria y frecuente, conocidos como signos vitales. Cuanto mayor sea el puntaje, más anormal será la fisiología del paciente y mayor será el riesgo de deterioro futuro. Los EWS se usan ampliamente en salas de hospitalización de pacientes agudos [13]. Los EWSs actuales se correlacionan con puntos finales importantes centrados en el paciente, como los niveles de intervención [14], la mortalidad hospitalaria [14, 15] y la duraciónn de la internación [15], y se ha demostrado que son un mejor predictor de paro cardíaco que los parámetros individuales [16]. Sin embargo, su desempeño puede mejorarse ya que la mayoría de los EWSs usan fórmulas simples que pueden calcularse en forma manual al lado de la cama, y solo incluyen un conjunto limitado de signos vitales como '*inputs*' [17]. Ahora que las HCE se utilizan ampliamente en la atención hospitalaria, existe la oportunidad de utilizar algoritmos más complejos y automatizados y una gama más amplia de *inputs*. En consecuencia, se han propuesto en la literatura algoritmos que mejoran el desempeño mediante el uso de técnicas de fusión de datos para combinar signos vitales con otros parámetros como datos de laboratorio y demográficos [18, 19].

El resto de este capítulo está diseñado para brindar al lector las herramientas necesarias para desarrollar y evaluar algoritmos de fusión de datos para la predicción de deterioro clínico.

22.2 Set de datos de estudio

Los datos se extrajeron de la base de datos MIMIC II (v. 2.26) [21], que está disponible en forma pública en PhysioNet [22]. Se eligió esta base de datos porque contiene datos de miles de pacientes en estado crítico registrados en forma rutinaria en la HCE, los cuales poseen un alto riesgo de deterioro. La extracción de datos se realizó utilizando las tres consultas SQL *cohort_labs. sql*, *cohort_vitals. sql* y *cohort_selection. sql*. Para facilitar el análisis, se extrajeron los datos de 500 pacientes únicamente. Sólo se extrajeron datos de adultos, ya que los datos pediátricos tienen rangos fisiológicos normales diferentes. Los parámetros extraídos de la base de datos, que figuran en la Tabla 22.1, fueron elegidos en línea con los utilizados previamente en la literatura [18,19].

Tradicionalmente, el desempeño de los EWSs se ha evaluado utilizando tres medidas de resultado con las que se han evaluado los sistemas de respuesta rápida: mortalidad, paro cardiorrespiratorio (PCR) y tasas de ingreso en la UCI [20]. Sin embargo, los PCR son difíciles de identificar de manera confiable en el conjunto de datos MIMIC II, y el conjunto de datos sólo contiene datos de pacientes que ya se encuentran en la UCI. Por lo tanto, se eligió la mortalidad, como la medida de resultado para este caso de estudio ya que se puede extraer de manera confiable y fácil del conjunto de datos.

Tabla 22.1 Parámetros de la HCE extraídos de la base de datos MIMIC II como *input* para los algoritmos de fusión de datos.

Parámetros Bioquímicos	**Signos Vitales**
Albúmina	Frecuencia Respiratoria
Anión GAP	Frecuencia Cardíaca
pCo2 Arterial	Presión Arterial – Sistólica y diastólica
pH Arterial	Temperatura
Transaminasa Glutámico-Oxalacética (TGO)	Saturación de oxígeno
Bicarbonato	Nivel de Consciencia
Nitrógeno Ureico en Sangre (BUN)	**Datos Demográficos**

Calcio	Edad
Creatinina	Género
Glucosa	
Hemoglobina	
Plaquetas	
Potasio	
Sodio	
Bilirrubina Total	
Recuento de Glóbulos Blancos (GB)	

22.3 Preprocesamiento

El análisis de los datos se realizó en Matlab®. El primer paso de preprocesamiento fue importar los archivos CSV generados por la consulta SQL en Matlab® (usando *LoadData. m*). El propósito de este paso fue crear:

1. Una matriz de diseño de variables predictoras (los parámetros enumerados en la Tabla 22.1): esta matriz MxN contenía valores para cada uno de los N parámetros en cada uno de los M puntos temporales. Esto se realizó utilizando la metodología en [19]: los puntos de tiempo se calcularon como los tiempos finales de períodos sucesivos de cuatro horas que abarcan la estadía en la UCI de cada paciente; los valores de los parámetros en los puntos temporales se establecieron a partir del último valor medido durante ese período de tiempo.
2. Una matriz de respuesta Mx3 de las tres variables dependientes fácilmente adquiridas, a saber: variables binarias de muerte en UCI y muerte en UCI dentro de las próximas 24 h, y una variable continua de tiempo hasta la muerte en UCI.

Los pasos y análisis de preprocesamiento restantes se realizaron utilizando solo datos de estas matrices.

Se requirió un preprocesamiento adicional para preparar los datos para el análisis (*PreProcessing.m*). En primer lugar, se observó que los valores de temperatura exhibían una distribución bimodal centrada en 37.1 y 98.8, lo que indica que algunos se habían medido en grados Celsius y otros en Fahrenheit. Los medidos en Fahrenheit se convirtieron a Celsius. En segundo lugar, el conjunto de datos contenía valores de presión arterial (PA) adquiridos de forma invasiva y no invasiva. Se retuvieron las mediciones invasivas ya que habían sido adquiridas con mayor frecuencia. Las mediciones no invasivas se reemplazaron con valores invasivos subrogados corrigiendo los sesgos observados entre las dos técnicas de medición cuando ambas se habían utilizado en los mismos períodos de cuatro horas (las diferencias medias entre las mediciones invasivas y no invasivas fueron 2,7 y 6 mmHg para PA sistólica, diastólica y media respectivamente). Finalmente, el conjunto de datos contenía valores faltantes donde los parámetros no se habían medido dentro de los períodos particulares de cuatro horas. Estos datos faltantes tuvieron que ser imputados ya que la técnica de análisis a utilizar, la regresión logística, requiere un conjunto de datos completo. Para hacerlo, seguimos el enfoque propuesto anteriormente de imputar el último valor medido, a menos que no se hubiera medido ningún valor, en cuyo caso se imputó el valor medio de la población [19]. Tenga en cuenta que este enfoque podría aplicarse a un conjunto de datos en tiempo real.

22.4 Métodos

Se crearon nuevos algoritmos de fusión de datos utilizando `CreateDataFusionAlgs.m`. Se utilizaron modelos lineales generalizados para fusionar variables continuas y binarias para proporcionar un resultado indicativo del riesgo de deterioro del paciente. Se utilizó un conjunto de datos de entrenamiento, que contenía el 50% de los datos, para crear los algoritmos.

Se utilizó la regresión logística para estimar la probabilidad de que cada una de las variables de respuesta binaria “muerte en la UCI” y “muerte en la UCI dentro de las 24 h” fuera verdadera. La regresión logística difiere de la regresión lineal ordinaria en que limita a que el resultado se encuentre entre 0 y 1, haciéndola adecuada para la estimación de la probabilidad de que una variable de respuesta sea verdadera. La regresión logística proporciona una estimación de:

$$y = \ln\left[\frac{p(x)}{1 - p(x)}\right]$$

donde *p* (x) es la probabilidad de que la variable de respuesta sea verdadera y *x* es un vector de variables predictoras. Observe que *p* (x) está restringido a estar entre 0 y 1 para todos los valores reales de *y*.

Cuando se utiliza la regresión logística, se debe decidir cómo modelar las relaciones entre las n variables predictoras contenidas dentro de *x*, y el resultado, *y*. El método más simple es suponer que *y* está linealmente relacionado con las variables predictoras como:

$$y = \alpha + \sum_{i=1}^{n} \beta_i x_i$$

donde α es el intercepto (ordenada al origen) y β un vector de coeficientes.

Para variables como la presión arterial diastólica, es razonable la suposición de una relación lineal porque cambian constantemente en una dirección particular durante un episodio de deterioro clínico. Sin embargo, otras variables como el nivel de sodio podrían cambiar en cualquier dirección lejos de la normalidad. Para estas variables, una relación no lineal es más apropiada, como la cuadrática

$$y = \alpha + \sum_{i=1}^{n} \beta_i x_i + \sum_{i=1}^{n} \gamma_i x_i^2$$

donde *y* es un vector de coeficientes para los cuadrados de las variables predictoras. Tenga en cuenta que esta relación 'puramente cuadrática' no contiene términos de interacción como $x_i x_j$. La importancia de la elección de la relación entre las variables predictoras y la estimación se demuestra en la figura 22.1.

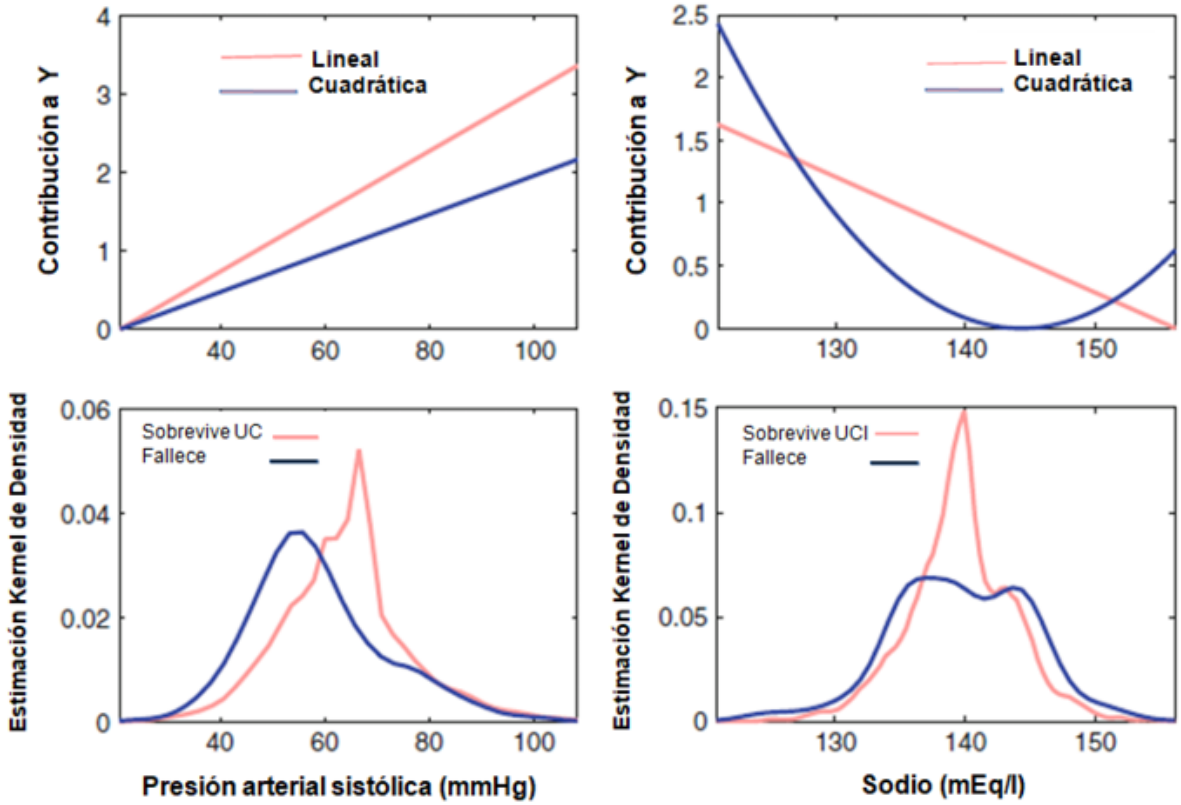

Fig 22.1 Una comparación de las contribuciones de las variables *input* al resultado del algoritmo, *Y*, bajo el supuesto de una relación lineal o no lineal entre las variables *input* e *Y*. La elección de la relación tuvo poco impacto en la contribución de la presión arterial diastólica (arriba, a la izquierda), ya que tendió a reducirse en aquellos pacientes que murieron (abajo, a la izquierda). Sin embargo, una relación cuadrática proporcionó una contribución muy diferente para el nivel de sodio (arriba, a la derecha), ya que los niveles de sodio de aquellos pacientes que murieron exhibieron una distribución bimodal indicando un aumento o una disminución fuera del rango normal (debajo, a la derecha).

En este caso de estudio se crearon algoritmos separados usando relaciones lineales y cuadráticas. En primer lugar, solo se incluyeron los parámetros que se utilizan en los EWS (signos vitales). En segundo lugar, se incluyeron todos los parámetros extraídos de la HCE.

En tercer lugar, se utilizó la regresión escalonada (*stepwise)* para evitar incluir términos que no aumentaran el desempeño del modelo. Esto consistió en construir un modelo mediante la inclusión de términos hasta que ningún término adicional aumentara el desempeño del modelo, y luego eliminar los términos cuya eliminación no disminuyera significativamente su desempeño.

22.5 Análisis

Los algoritmos de EWS deben desencadenar una respuesta clínica efectiva para impactar en los resultados del paciente. Por lo general, se exige una respuesta particular cuando el resultado del algoritmo se eleva por encima de un valor umbral. La respuesta puede incluir una revisión clínica por parte del personal de la sala o un equipo centralizado de respuesta rápida. El siguiente análisis se basa en el supuesto de que los algoritmos se utilizarían para ordenar respuestas como esta.

El desempeño de cada algoritmo se analizó utilizando el último 50% de los datos, el conjunto de datos de validación. En todos los puntos de tiempo de 4 h, se usó el modelo para estimar la probabilidad de que un paciente muriera durante su estadía en la UCI. La figura 22.2 muestra gráficos ejemplo de los resultados de cuatro pacientes a lo largo de sus estadías en la UCI. A lo largo del análisis, cada punto de tiempo se clasificó como positivo o negativo, lo que indica que el modelo predijo que el paciente murió posteriormente en la UCI o sobrevivió hasta el alta de la UCI. Por lo tanto, se identifica un verdadero positivo en un momento determinado cuando el modelo predice correctamente la muerte de un paciente que falleció en la UCI, mientras que se identifica un falso positivo cuando el modelo predice incorrectamente la muerte de un paciente que sobrevivió al alta de la UCI. Los verdaderos negativos y falsos negativos se identificaron de manera similar.

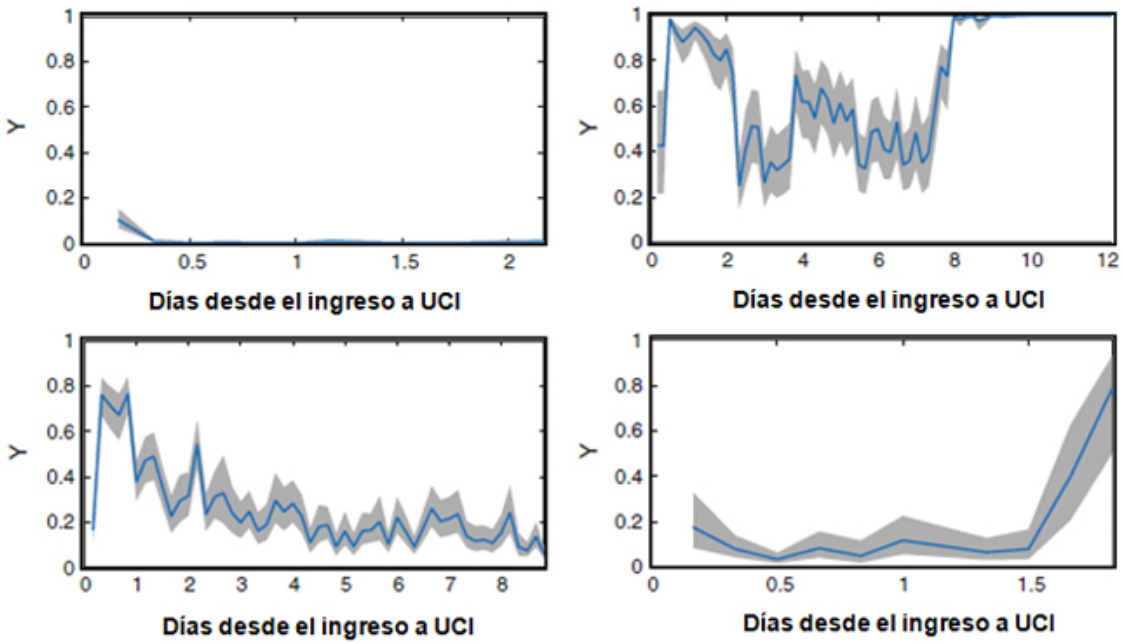

Fig 22.2 Gráficos ejemplo del resultado (Y) de los resultados de los algoritmos durante el tiempo de estadía de los pacientes en UCI. Los gráficos de la izquierda muestran a los pacientes que sobrevivieron durante su estadía en UCI mientras que los gráficos de la derecha muestran aquellos que murieron. Los gráficos superiores muestran ejemplos en los cuales el algoritmo funcionó bien, mientras que los gráficos inferiores muestran ejemplos en los que no.

La Tabla 22.2 muestra el desempeño de cada algoritmo evaluado usando el área bajo la curva de características operativas del receptor (ROC) (AUROC).

Tabla 22.2 El desempeño de los algoritmos de fusión de datos para la predicción de muertes en la UCI, mostrados como áreas bajo la curva ROC (AUROC) y las máximas sensibilidades donde los algoritmos eran forzados a satisfacer los requerimientos clínicos de VPP ≥ 0.33 y una tasa de alerta ≤ 17%.

Relación entre la variable predictora y el resultado	Variables predictoras candidatas	Número de variables predictoras incluidas	AUROC	Sensibilidades Máximas (%)	
				VPP≥ 0.33	Tasa de alerta ≤ 17%
Lineal	Signos vitales únicamente	6	0.757	14.4	42.5
Lineal	Todas	25	0.800	46.6	49.7
Lineal	Inclusión *stepwise* de todas	23	0.800	45.8	48.9
Cuadrática Pura	Signos vitales únicamente	6	0.774	13.2	41.4
Cuadrática Pura	Todas	25	0.799	55.5	53.9
Cuadrática Pura	Inclusión *stepwise* de todas	21	0.810	59.3	56.3

El algoritmo con el AUROC más alta, de 0.810, utilizó la inclusión de parámetros *stepwise* y la relación cuadrática. Las curvas ROC para este algoritmo y el algoritmo correspondiente utilizando sólo signos vitales se muestran en la figura 22.3. Los algoritmos que usaron todos los parámetros disponibles como *inputs* tuvieron AUROC más altas que aquellos que usaron sólo signos vitales, lo que demuestra el beneficio de fusionar los signos vitales con parámetros adicionales. En la mayoría de los casos, el uso de una relación cuadrática resultó en un AUROC más alta. Además, la selección

escalonada (stepwise) de parámetros redujo el número de parámetros requeridos, al tiempo que mantuvo o mejoró el AUROC.

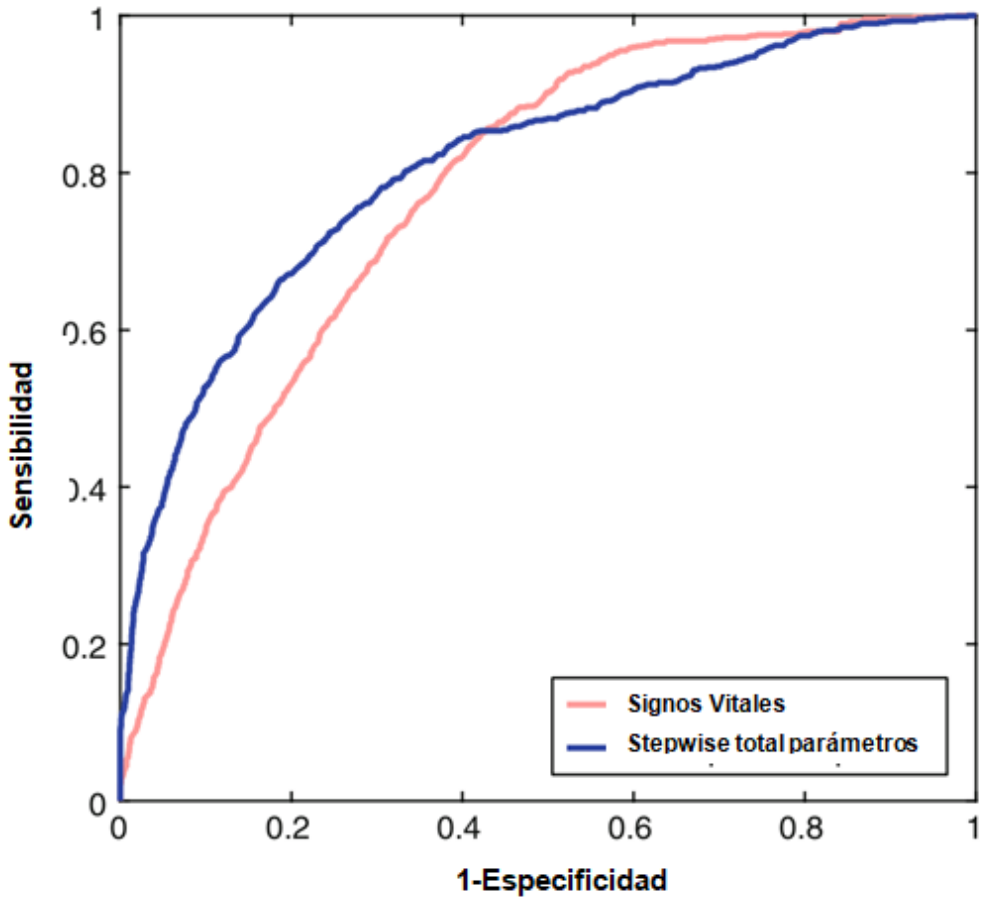

Fig 22.3 Curvas ROC que muestran el desempeño de los mejores algoritmos utilizando la inclusión escalonada *(stepwise)* de todos los parámetros y los signos vitales únicamente. Estos algoritmos asumen una relación cuadrática entre las variables predictoras y el resultado.

Se han sugerido otras métricas para la comparación de algoritmos, incluída la sensibilidad, el valor predictivo positivo (VPP) y la tasa de alerta [23]. Sin embargo, estas son más difíciles de usar ya que cada métrica varía según el valor umbral. Un método útil para comparar algoritmos usando estas métricas es comparar sus sensibilidades cuando se usa un umbral que proporciona un desempeño del algoritmo adecuado para los requisitos clínicos.

En el caso de los algoritmos de EWS, los requisitos clínicos clave son que el VPP alcance o supere un nivel mínimo aceptable, y la tasa de alerta esté por debajo de un nivel máximo aceptable. En ausencia de valores basados en la

evidencia, para fines de demostración utilizamos un VPP mínimamente aceptable de 0,33, lo que indica que una de cada tres alertas es un verdadero positivo, y una tasa de alerta máxima aceptable del 17%, lo que indica que una de cada seis observaciones se convierte en una alerta.

La Tabla 22.2 muestra las sensibilidades proporcionadas por cada algoritmo cuando se ve obligado a satisfacer estos requisitos clínicos. Los VPP y las tasas de alerta en todos los umbrales se muestran en la Fig. 22.4 para los algoritmos de mejor rendimiento utilizando sólo signos vitales y utilizando la inclusión *stepwise* de todos los parámetros.

Las sensibilidades más altas se lograron al utilizar la inclusión stepwise de todos los parámetros, con una relación puramente cuadrática. El beneficio de usar parámetros adicionales más allá de los signos vitales se muestra claramente en las sensibilidades de los algoritmos con un VPP mínimo aceptable, que fue del 13.2% cuando se usaron solo los signos vitales, y del 59.3% cuando se usó la inclusión stepwise de todos los parámetros.

En [19] se utilizaron visualizaciones adicionales para demostrar el efecto de elegir diferentes umbrales. En primer lugar, la variable dependiente del tiempo antes de la muerte en la UCI se usó para examinar cómo cambió el resultado con el tiempo antes de la muerte, como se muestra en la figura 22.5. Esto muestra que un umbral más bajo da como resultado una advertencia más avanzada de deterioro clínico. En segundo lugar, se presentó la proporción de pacientes que alcanzaron cada resultado durante su estadía, como se muestra en la figura 22.6. Esto sugiere que un umbral más bajo genera más alertas falsas y menos alertas verdaderas.

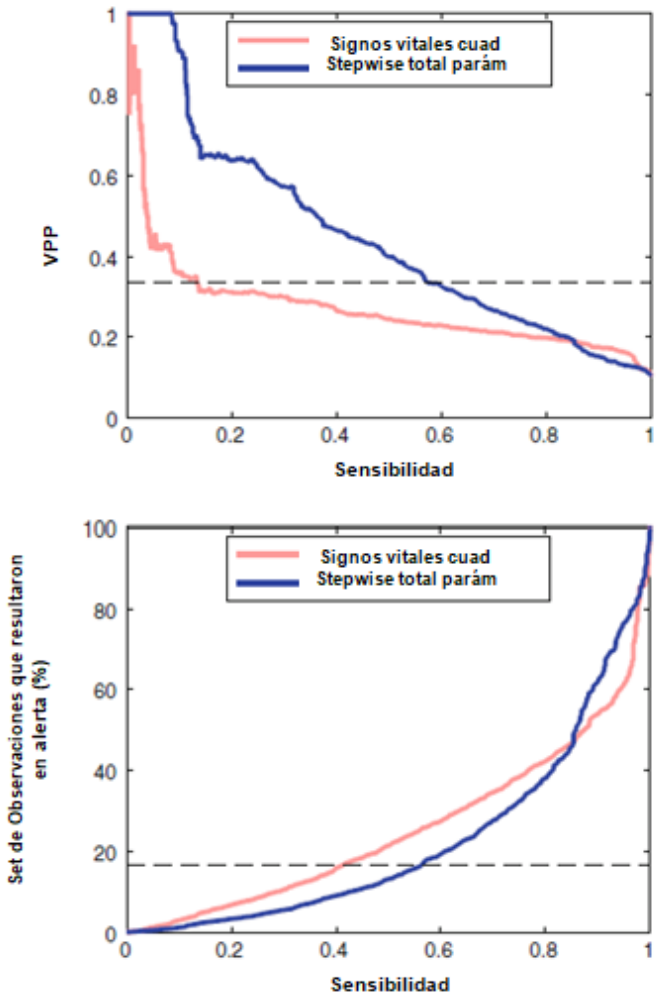

Fig 22.4 Comparación de los VPPs y las tasas de alerta para algoritmos que utilizan signos vitales únicamente y aquellos que usan todos los parámetros. Los requerimientos clínicos ejemplo VPP ≥ 0.33 y tasa de alerta ≤ 17% se muestran en la línea punteada. El algoritmo cuadrático usando signos vitales únicamente posee una sensibilidad del 13.2%, mucho más baja que el algoritmo equivalente que utiliza la incorporación *stepwise* de todos los parámetros cuya sensibilidad es 59.3% cuando alcanza el criterio de VPP. De forma similar, cuando se utiliza el criterio de tasa de alerta la sensibilidad del algoritmo que utiliza sólo signos vitales es 41.4% también más bajo que el algoritmo que utiliza la inclusión stepwise de variables (56.3%).

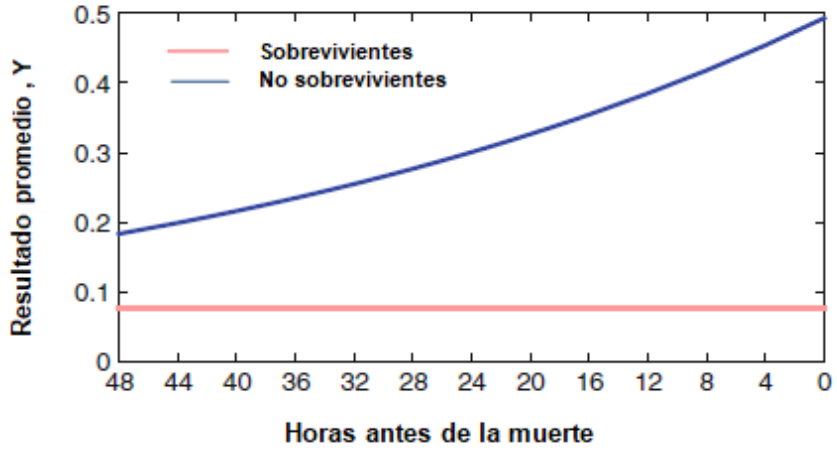

Fig 22.5 Promedio de los resultados del algoritmo durante las 48 hs. previas a la muerte en la UCI (luego de suavización exponencial). Una opción más baja en el umbral de alerta resulta en una advertencia en un nivel más avanzado de deterioro.

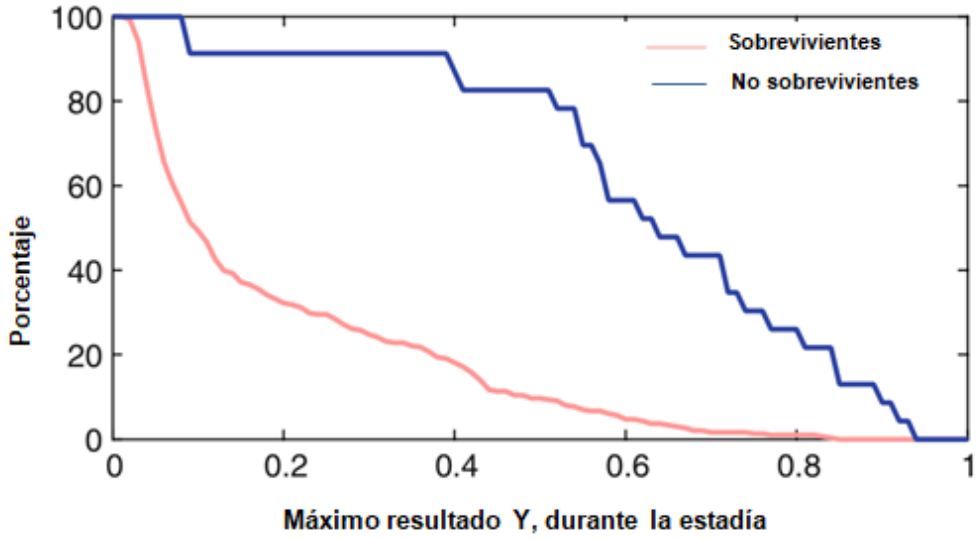

Fig 22.6 La proporción de sobrevivientes y no sobrevivientes que alcanzaron cada valor de resultado del algoritmo durante su estadía en UCI. Una opción más baja en el umbral de alerta resulta en más alertas falsas y menores alertas reales.

22.6 Discusión

La introducción de las HCEs ha brindado la oportunidad de mejorar los algoritmos clínicos utilizados para identificar deterioros clínicos. Los algoritmos de fusión de datos descriptos en este capítulo estiman la probabilidad de que un paciente muera durante su estadía en la UCI cada 4 hs.

En este estudio, la inclusión de parámetros fisiológicos adicionales, más allá de los signos vitales, dio como resultado mejoras en el desempeño del algoritmo cuando se evaluó utilizando el AUROC, como se observó anteriormente [18,19] y cuando se evaluó utilizando las sensibilidades mínimas correspondientes a los requisitos clínicos.

Este caso de estudio ha demostrado los pasos fundamentales necesarios para diseñar y evaluar algoritmos de fusión de datos para la predicción de deterioros clínicos. Durante el preprocesamiento, se extrajeron los datos requeridos de los archivos de datos sin procesar y se procesaron en matrices listas para el análisis. Fue importante realizar este paso de forma separada al análisis para reducir el tiempo requerido en el diseño del algoritmo. Durante este paso identificamos deficiencias en el conjunto de datos. Desafortunadamente, no hay una forma sistemática de garantizar que se hayan identificado todas las deficiencias. Recomendamos que primero se inspeccionen las distribuciones de cada variable para identificar discrepancias obvias, como las diferentes unidades utilizadas para la temperatura en este conjunto de datos. En segundo lugar, es útil trazar los datos sin procesar a lo largo del tiempo para identificar cualquier cambio en la práctica que pueda haber ocurrido durante la adquisición de los mismos. En tercer lugar, a menudo es valioso buscar la orientación de un médico o un curador de la base de datos en la institución, o un investigador que haya trabajado con el conjunto de datos anteriormente.

Los resultados presentados aquí no pueden generalizarse a la población general de pacientes de todo un hospital por dos razones. En primer lugar, el conjunto de datos consta de datos de pacientes en estado crítico, mientras que los EWS están diseñados principalmente para identificar deterioro en pacientes con enfermedades agudas. Dado que los procesos de enfermedad de los pacientes críticos son más avanzados y tienen intervenciones clínicas adicionales, como ventilación mecánica y soporte de órganos, tanto la fisiología basal como los cambios fisiológicos que acompañan los deterioros clínicos pueden diferir en esta población en comparación con los pacientes agudamente enfermos. En segundo lugar, la muerte en la UCI se utilizó como variable dependiente en este estudio. La muerte es la última etapa posible de deterioro y, por lo tanto, un algoritmo que predice la muerte puede no predecir el inicio de deterioros clínicos lo suficientemente temprano como para ser de utilidad clínica en pacientes agudamente enfermos.

La elección de métodos estadísticos para evaluar el desempeño de los EWS es tema de debate [23]. El AUROC se ha utilizado a menudo para cuantificar el desempeño de los algoritmos de EWS, como en [17]. Esta estadística se calcula a partir de las sensibilidades y especificidades de un algoritmo en un rango de valores umbral. Sin embargo, recientemente se ha sugerido que el AUROC es engañosa debido a la baja prevalencia de deterioros [23]. En [23] se propusieron medidas estadísticas alternativas para dar cuenta de los requisitos clínicos de los algoritmos de EWS. Las medidas estadísticas deben evaluar en primer lugar los beneficios y los costos del uso de EWS. El beneficio es que los EWS pueden actuar como una red de seguridad para capturar pacientes con deterioro, no diagnosticados en las evaluaciones clínicas de rutina.

Esto requiere una alta sensibilidad (la proporción de evaluaciones EWS de pacientes con deterioro que disparan una alerta). El costo de los EWS es el tiempo utilizado para responder a las alertas falsas. Este costo es relativamente pequeño, ya que la evaluación clínica adicional desencadenada por una alerta toma solo un corto período de tiempo. Esto significa que no es de gran importancia una alta especificidad (la proporción de pruebas negativas que son verdaderas negativas). En segundo lugar, es importante asegurarse de que el valor predictivo positivo (la proporción de alertas que son verdaderas) sea lo suficientemente alto como para evitar que el personal sufra desensibilización a las alertas, lo que puede dar como resultado respuestas menos efectivas a los pacientes que se identifican correctamente como con deterioro [24]. En tercer lugar, la tasa de alerta debe ser manejable para evitar la utilización excesiva de recursos. En este caso de estudio presentamos el AUROC y las sensibilidades máximas cuando los algoritmos se restringieron a un VPP mínimamente aceptable y una tasa de alerta máximamente aceptable [23].

22.7 Conclusiones

Este estudio de caso ha demostrado la utilidad potencial de las técnicas de fusión de datos para predecir el deterioro clínico. Actualmente, la identificación de los deterioros se logra utilizando EWS que toman signos vitales como *inputs*. El desempeño de los algoritmos de fusión de datos evaluados en este estudio se mejoró al aumentar el conjunto de *inputs* para

incluir parámetros fisiológicos que se encuentran habitualmente disponibles en las HCE, pero que no se miden al pie de la cama del paciente.

Se han mostrado las técnicas fundamentales para el diseño y evaluación de algoritmos de fusión de datos. Se utilizaron algoritmos de regresión logística para predecir una variable de respuesta binaria: muerte en la UCI.

Se mostró el uso de relaciones lineales y cuadráticas entre las variables predictoras y de respuesta, así como el uso de la inclusión escalonada *stepwise* de variables. Se presentó una gama de medidas estadísticas para la evaluación de algoritmos, que ilustran los beneficios del uso de medidas estadísticas alternativas al AUROC.

Los resultados no deben interpretarse como representativos de los resultados que podrían esperarse cuando los EWS se usan en salas de internación general, ya que el conjunto de datos del estudio correspondió a pacientes críticos y se utilizó como variable dependiente la muerte en la UCI. Sin embargo, las técnicas utilizadas para diseñar y evaluar algoritmos se pueden aplicar fácilmente a una amplia gama de pacientes, proporcionando la base para un trabajo en el futuro.

22.8 Próximos pasos

Se han identificado dos áreas en particular para futuras investigaciones. En primer lugar, puede repetirse el estudio utilizando datos de pacientes agudamente enfermos en lugar de pacientes en estado crítico, utilizando una variable dependiente distinta de la muerte. Esto facilitará el diseño de algoritmos generalizables a la población del hospital. En segundo lugar, pueden explorarse una serie de funciones adicionales para modelar la relación entre las variables predictoras y el resultado.

Las funciones más complejas que las funciones lineales o puramente cuadráticas, como las polinómicas de alto orden o las funciones logísticas, pueden mejorar el desempeño. Además, sería prudente investigar el efecto de la inclusión de términos de interacción para tener en cuenta las relaciones entre las variables predictoras.

22.9 Predicción personalizada del deterioro clínico

Los algoritmos presentados aquí tienen un alcance limitado por los parámetros usados como *inputs*. Habitualmente, brindan una descripción detallada del estado fisiológico del paciente a partir de los signos vitales y los

valores bioquímicos, que constituyen 23 de los 25 *inputs*. Sin embargo, estos parámetros muestran diferencias mínimas entre pacientes individuales según su estado al ingreso al hospital. Por el contrario, los médicos utilizan información adicional presente al ingreso hospitalario durante la internación de un paciente, para contextualizar las evaluaciones fisiológicas.

Para ilustrar esto, considere la respuesta de los algoritmos a dos hombres ficticios de 65 años, los pacientes A y B. El paciente A tiene antecedentes de hipertensión y una presión arterial sistólica (PAS) alta antes del ingreso hospitalario de 147 mmHg. El paciente B ha llevado una vida activa, tiene una dieta saludable y tiene una PAS relativamente baja antes del ingreso de 114 mmHg. Durante su estadía en el hospital, la PAS de ambos pacientes se mide en 114 mmHg. Los algoritmos no pueden distinguir si esto es representativo del paciente A durante un deterioro significativo, como las primeras etapas de la hipotensión que precede al shock séptico, o si es representativo del estado habitual del paciente B en ausencia de cualquier deterioro. Si los algoritmos utilizaran una gama más amplia de *inputs* indicativos del estado del paciente antes del ingreso, como la presencia o ausencia de comorbilidades (afecciones médicas existentes), incluída la hipertensión, podrían ser capaces de diferenciar entre los pacientes A y B en esta situación.

Esto ilustra el beneficio potencial de incorporar *inputs* adicionales indicadores de comorbilidades. Se puede obtener un beneficio aún mayor personalizando los algoritmos de EWS de acuerdo con el estado fisiológico antes de la admisión. Los algoritmos de EWS personalizados no solo estratificarían a los pacientes utilizando *inputs* adicionales para contextualizar la fisiología, sino que también personalizarían los coeficientes de regresión de acuerdo con el estado fisiológico del paciente medido previamente en un momento de relativa salud.

licencia Creative Commons de la obra y la acción respectiva no está permitida por el reglamento, los usuarios deberán obtener permiso del titular de la licencia para duplicar, adaptar o reproducir el material.

Nota: Las imágenes de este capítulo se encuentran disponibles a color en el ebook; disponible en www.hardineros.com

Apéndice: Códigos

El código utilizado en este estudio de caso está disponible en el repositorio de GitHub que acompaña a este libro: https://github.com/MIT-LCP/critical-data-book. En este sitio web se encuentra disponible más información sobre el código. Los siguientes scripts clave se utilizaron para extraer datos de la base de datos MIMIC II:

- `cohort_selection.sql`: se usó para identificar una cohorte de pacientes de la cual serían extraídos los datos.
- `cohort_labs.sql`: se usó para extraer los resultados de la pruebas de laboratorio.
- `cohort_vitals.sql`: se usó para extraer los signos vitales

Los datos se extrajeron en formato CSV. El análisis posterior se realizó en Matlab® utilizando RunFusionAnalysis.rn. Contiene el siguiente script:

- SetupUniversalParams: usado para establecer los parámetros universales (en este caso, rutas de archivos) que son utilizados para cargar y guardar archivos durante el análisis. Estos parámetros deben adaptarse cuando se usan en el código. Luego llamaba a los siguientes scripts:
- `LoadData.m`: usado para cargar datos CSV en Matlab® para análisis.
- `PreProcessing.m`: ejecuta el preprocesamiento para preparar los datos para análisis.
- `CreateDataFusionAlgs.m`: crea algoritmos de fusión de datos usando datos de entrenamiento.
- `AnalysePerformances.m`: analiza el desempeño de los algoritmos de fusión de datos usando datos de validación.

Referencias

1. Silber JH et al (1995) Evaluation of the complication rate as a measure of quality of care incoronary artery bypass graft surgery. JAMA 274 (4): 317-323.
2. Khan NA et al (2006) Association of postoperative complications with hospital costs and length of stay in a tertiary care center. J Gen Intem Med 21 (2): 177-180.
3. Lagoe RJ et al (2011) Inpatient hospital complications and lengths of stay: a short report. BMC Res Notes 4 (1): 135.
4. Schein RM et al (1990) Clinical antecedents to in-hospital cardiopulmonary arrest. Chest 98 (6): 1388-1392.
5. Franklin C et al (1994) Developing strategies to prevent inhospital cardiac arrest: analyzing responses of physicians and nurses in the hours before the event. Crit Care Med 22 (2): 244-247.
6. Buist MD et al (1999) Recognising clinical instability in hospital patients before cardiac arrest or unplanned admissión to intensive care. A pilot study in a tertiary-care hospital. Med J Aust 171 (1): 22-25.
7. Hillman KM et al (2001) Antecedents to hospital deaths. Intem Med J 31 (6): 343-348.
8. Hillman KM et al (2002) Duration of life-threatening antecedents prior to intensive care admissión. Intensive Care Med 28 (11): 1629-1634.
9. Whittington J et al (2007) U sing an automated risk assessment report to identify patients at risk for clinical deterioration. Jt Comm J Qual Patient Saf 33 (9): 569-574.
10. Smith AF et al (1998) Can sorne in-hospital cardio-respiratory arrests be prevented? A prospective survey. Resuscitation 37 (3): 133-137.
11. Patient Safety Observatory (2007) Safer care for the acutely ill patient: learning from serious incidents. National Patient Safety Agency, London.
12. Whittington J et al (2007) Using an automated risk assessment report to identify patients at risk for clinical deterioration. Jt Comm J Qual Patient Saf 33 (9): 569-574.
13. Royal College of Physicians (2012) National early warning score (NEWS): standardising the assessment of acute-illness severity in the NHS", Report of a working party. RCP, London.
14. Goldhill DR et al (2005) A physiologically-based early warning score for ward patients: the association between score and outcome. Anaesthesia 60 (6): 547-553.
15. Paterson R et al (2006) Prediction of in-hospital mortality and length of stay using an early warning scoring system: clinical audit. Clin. Med. 6 (3): 281-284.
16. Churpek MM et al (2012) Predicting cardiac arrest on the wards: a nested case-control study. Chest 141 (5): 1170-1176.
17. Smith GB et al (2013) The ability of the national early warning score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admissión, and death. Resuscitation 84 (4): 465-4 70.

18. Alvarez CA et al. (2013) Predicting out of intensive care unit cardiopulmonary arrest or death using electronic medical record data. BMC Med lnform Decis Mak 13 (28).
19. Churpek MM et al (2014) Multicenter development and validation of a risk stratification tool for ward patients. Am J Respir Crit Care Med 190:649-655.
20. Maharaj Retal (2015) Rapid response systems: a systematic review and meta-analysis. Crit Care 19 (1): 254.
21. Saeed M et al (2011) Multiparameter intelligent monitoring in intensive care 11: a public-access intensive care unit database. Crit Care Med 39 (5): 952-960.
22. Goldberger AL et al (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101 (23): E215-E220.
23. Romero-Brufau S, Huddleston JM, Escobar GJ, Liebow M (2015) Why the C-statistic is not informative to evaluate early warning seores and what metrics to use. Crit Care 19 (1): 285.
24. Cvach M (2012) Monitor alarm fatigue: an integrative review. Biomed Instrum Technol 46 (4): 268-277.

Capítulo 23

EFECTIVIDAD COMPARATIVA: ANÁLISIS POR PUNTAJE DE PROPENSIÓN

Kenneth P. Chen y Ari Moskowitz

Objetivos de aprendizaje

Comprender los beneficios y desventajas del uso del análisis por puntaje de propensión para el modelado estadístico y la inferencia causal en la investigación basada en el análisis de la historia clínica electrónica (HCE).

Este caso de estudio incorpora conceptos que deberían mejorar el conocimiento de lo siguiente:

- Conocer los diferentes enfoques para la estimación de los puntajes de propensión: paramétrico, no paramétrico y por aprendizaje automático; comprender los pro y contras de cada uno.
- Aprender las diferentes formas de usar los puntajes de propensión para el ajuste por condiciones pre-tratamiento y evaluar el balance de las condiciones pre-tratamiento entre diferentes grupos de tratamiento.
- Valorar los conceptos que subyacen al análisis por puntaje de propensión de las HCEs, incluyendo estratificación, pareamiento y ponderación por probabilidad inversa (incluyendo ponderación directa, ponderación estabilizada y regresión ponderada doblemente robusta).

23.1 Motivos para utilizar Análisis por Puntaje de Propensión

Cuando llevamos a cabo investigaciones usando la historia clínica electrónica (HCE) u otras fuentes de *big data*, accedemos a una gran cantidad de covariables [1]. Estas covariables incluyen datos de pacientes, como ser: demográficos, parámetros clínicos (por ej. signos vitales y datos del examen físico), parámetros de laboratorio, medicamentos, condiciones premórbidas, etc. Todas estas covariables podrían actuar como factores de confusión al considerar la asociación entre una exposición y un resultado. Podemos usar modelos estadísticos para ajustar el efecto de confusión de estas covariables y establecer una asociación entre la exposición y el resultado de interés [2, 3]. El análisis por puntaje de propensión es particularmente útil cuando nos toca trabajar con un gran número de

covariables [1]. El resto de este capítulo asume que el lector cuenta con una comprensión básica de la estadística y del modelado de regresión (especialmente regresión logística).

Ajustar tantas covariables como sea posible sienta las bases para una inferencia causal convincente al reducir los sesgos latentes generados por variantes latentes [4]. Sin embargo, esto incrementa la dimensionalidad [5]. Aunque las HCEs complejas a menudo cuentan con muestras de tamaño suficiente como para permitir estudios de altas dimensiones, la reducción de la dimensionalidad sigue siendo útil por las siguientes razones: (i) para simplificar el modelo final y facilitar la interpretación, (ii) permitir la realización de análisis de sensibilidad para explorar términos de orden superior o términos de interacción para aquellas covariables que podrían estar correlacionadas o interactuando con el resultado, y (iii) dependiendo de la pregunta de investigación, el tamaño de la cohorte del estudio podría seguir siendo pequeño pesar de haber sido tomado de una gran base de datos, y entonces la reducción de la dimensionalidad se vuelve crucial para que un modelo sea válido.

23.2 Precauciones al usar Análisis por Puntaje de Propensión

Aunque el análisis por puntaje de propensión tiene las ventajas mencionadas previamente, es importante comprender la teoría de esta técnica para apreciar sus limitaciones. Un puntaje de propensión es una “probabilidad estimada” de que un sujeto sea asignado al grupo de tratamiento o al grupo control en base a sus características o condiciones pre-tratamiento. Es un sustituto, de todas las covariables usadas para estimarlo. No es difícil imaginar que utilizar un único puntaje de propensión para representar todas las características de un sujeto podría introducir sesgo [6]. Por eso, la implementación de puntajes de propensión en un modelo de análisis estadístico tiene que tener en cuenta la pregunta de investigación, el set de datos y las covariables incluidas en el mismo. Además, los resultados siempre deberán validarse con análisis de sensibilidad [7].

23.3 Diferentes Enfoques para la Estimación de los Puntajes de Propensión

En un ensayo clínico controlado aleatorizado (ECA) puede determinarse una relación causal entre la exposición (tratamiento) y el resultado si la aleatorización se llevó a cabo de manera correcta, por ej. si no existen diferencias en las condiciones pre-tratamiento entre los dos grupos. Sin embargo, en estudios retrospectivos casi siempre existe una diferencia en las condiciones pre-tratamiento entre los dos grupos. Para demostrar eficacia comparativa, la inferencia causal a través del modelado estadístico puede realizarse de varias maneras [8, 9]. Para el análisis por puntaje de propensión [3, 10] las condiciones pre-tratamiento pueden usarse como predictores en la determinación de la probabilidad de asignación de un sujeto al grupo de tratamiento o al grupo control. En otras palabras, la probabilidad de caer en uno u otro grupo es una función de las condiciones pre-tratamiento. Hay varias maneras de generar esta función. La más básica es la regresión.

Cuando se usa la regresión para estimar los puntajes de propensión, el resultado de la ecuación de regresión es la asignación: grupo tratamiento o grupo control, por ej. un resultado binario, y las variables en la ecuación de regresión pueden ser una combinación de variables numéricas y nominales. Se trata de una regresión logística multivariable que puede realizarse fácilmente a través de la mayor parte de los paquetes estadísticos libres o comerciales. Si hay más de un grupo de tratamiento (e.g., tratamiento A, tratamiento B y grupo control) [11], entonces el puntaje de propensión puede estimarse usando el método de regresión logística multinomial.

El modelo de regresión convencional es paramétrico. Por eso, el puntaje de propensión estimado estará sujeto a las limitaciones inherentes de los modelos paramétricos, por ej. errores en las especificaciones del modelo [12]. Es posible usar modelos no paramétricos para estimar el puntaje de propensión [13], como los árboles de regresión, enfoques particionados y distribuciones de Kernel. De todas maneras, estos métodos están menos establecidos y es probable que requieran el uso de algoritmos de aprendizaje automático (Machine Learning) [14]. Aunque los métodos no paramétricos a menudo requieren algoritmos de aprendizaje automático, las técnicas de aprendizaje automático pueden aplicarse para métodos paramétricos y no paramétricos. Por ejemplo, algunos estudios pueden usar un algoritmo genético para seleccionar las variables y la especificación del modelo de una regresión logística convencional para estimar el puntaje de propensión [15].

23.4 Usando el Puntaje de Propensión para Ajustar por Condiciones previas al tratamiento

El objetivo del análisis por puntaje de propensión es crear grupos de tratamiento y control que sean indistinguibles entre sí en términos de las estadísticas de las condiciones pre-tratamiento (por ej. promedio y desvío standard de variables numéricas, distribución de variables nominales). En otras palabras, se crean grupos control y tratamiento de manera tal que imitan el resultado de la asignación post aleatorización en un ECA, de manera tal de favorecer la inferencia causal. El análisis por puntaje de propensión es una de las herramientas para alcanzar esta meta [8, 9, 16].

Por ejemplo, consideren un sujeto que recibió la droga en estudio o el tratamiento (grupo de tratamiento) y un sujeto que recibió placebo o el tratamiento standard (grupo control). Si tienen condiciones pre-tratamiento similares, entonces sus chances (probabilidades) de ser asignados al grupo tratamiento son las mismas. Por lo tanto es comparable a la de dos sujetos idénticos que son aleatoriamente asignados a los grupos de tratamiento o control. Cuando encontramos dos sujetos que tienen similares puntajes de propensión y uno recibió tratamiento y el otro placebo, entonces podemos parearlos o hacerlos coincidir en la cohorte final de estudio antes de revisar el efecto del tratamiento (variable resultado). Este proceso se llama "pareamiento por puntaje de propensión". Al hacerlo, tendremos distribuciones similares de los puntajes de propensión (o de las distribuciones de las condiciones pre-tratamiento) entre los grupos de tratamiento y control.

Si el modelo usado para estimar el puntaje de propensión está bien especificado [17, 18], esperaremos que los puntajes de propensión sean representativos de las condiciones pre-tratamiento de los sujetos. De todas maneras, esto no siempre será así, por lo que siempre deberemos revisar las estadísticas de los grupos luego del ajuste por puntaje de propensión. Dado que el objetivo final es eliminar la diferencia en las condiciones pre-tratamiento entre grupos, existen otros métodos como la ponderación por puntaje de propensión para lograr este resultado. Se desarrollaron algoritmos de aprendizaje automático más sofisticados que se enfocan en el balance en las variables pre-tratamiento entre los dos grupos durante el proceso de estimación del puntaje de propensión para garantizar un modelo válido en la simulación de un resultado similar al de un ECA [19].

En la investigación basada en la HCE tenemos acceso a un gran número de covariables pre-tratamiento que podemos extraer de la base de datos y usar en el modelo de puntaje de propensión. Aunque no podemos usar un número indefinido de variables para simular un ECA real (que tiene en cuenta todas las variables no observadas) podemos ganar mayor confianza en nuestra conclusión incluyendo más variables [20, 21]. El análisis por puntaje de propensión es una herramienta robusta para simplificar el modelo final al mismo tiempo que permite incluir una gran cantidad de condiciones pre-tratamiento. La figura 23.1 resume la discusión previa de la aplicación de un modelo de puntaje de propensión.

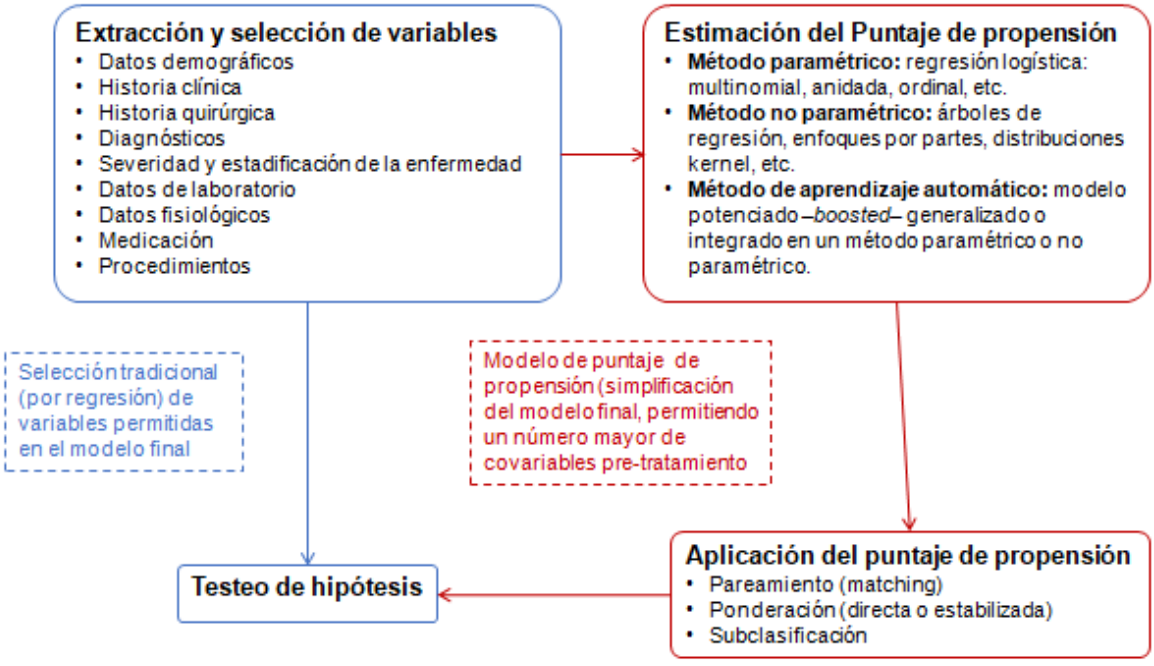

Fig. 23.1 Integración del análisis por puntaje de propensión en el diseño estadístico

Presentamos a continuación un caso de estudio que usó la base de datos MIMIC II (v. 2.26) [22,23] y se enfoca en la aplicación de los puntajes de propensión en la fase analítica. El estudio fue una cohorte retrospectiva de pacientes internados en la unidad de cuidados intensivos (UCI) que fueron tratados con al menos un regulador de la frecuencia cardíaca (FC): metroprolol, amiodarona o diltiazem. El análisis por puntaje de propensión se llevó a cabo usando las siguientes covariables: datos demográficos, signos vitales, laboratorios metabólicos básicos, condiciones médicas previas,

puntajes de severidad, tipos de admisión y tipo de UCI. Los resultados que se midieron fueron: (i) si el control de la FC se logró con uno o con múltiples agentes (resultado binario); y, en aquellos sujetos que lograron el resultado, (ii) el momento en el cual lo hicieron (resultado como variable continua).

23.5 Preprocesamiento de datos

Para identificar aquellos sujetos con fibrilación auricular y respuesta ventricular rápida (FA con RVR) en la base de datos usamos una combinación de datos estructurados y no estructurados. Específicamente, los datos estructurados incluyeron códigos CIE-9 (el código correspondiente a "fibrilación auricular" e 427.31) y datos de administración de medicación. Los datos no estructurados que usamos incluyeron la curva del trazado del electrocardiograma (ECG), la medición seriada de la FC, resúmenes de egreso y notas de enfermería. Desafortunadamente, sólo una fracción pequeña de los pacientes en la base de datos tenía datos de la curva de ECG (2000 de 32000), por lo cual no pudimos hacer uso completo del análisis de las curvas.

Los pacientes que tenían FA con RVR registrado en sus resúmenes de egreso fueron identificados buscando palabras claves equivalentes en los resumenes de egreso, excluyendo la sección de antecedentes. Una vez que estos pacientes fueron identificados utilizamos los campos de registro de la FC y de la administración de medicación para encontrar el subconjunto de pacientes que tuvieron una FC superior a 110 latidos por minuto (lpm) por más de 15 minutos y que recibieron al menos una dosis de las drogas de interés para el control de la FC (metoprolol, dialtiazem o amiodarona). Los datos se extrajeron usando la versión SQL de Oracle® y se procesaron posteriormente con Python® para la búsqueda de texto en los resúmenes de egreso, y Matlab® para el procesamiento y representación gráfica de los datos seriados de FC y el establecimiento de una relación temporal entre la respuesta ventricular rápida y la administración de la medicación.

Los datos seriados de FC estaban disponibles para casi todos los pacientes de la base de datos. De todas maneras, a diferencia del registro continuo de la onda de ECG, la FC sólo se registra cada 5,10 ó 15 minutos y de forma inconsistente. Para homogeneizar los datos y facilitar su procesamiento y representación gráfica, interpolamos los datos de FC cada 5 minutos: durante la estadía en la UCI, si un dato crudo de FC no estaba disponible para

algún período de 5 minutos, se interpoló un valor usando los dos puntos adyacentes como referencia. Debido a que el muestreo de FC es poco frecuente para esta entidad, un punto de FC por encima de 110 lpm podría corresponder a un episodio de FC rápida de 5 minutos de duración. Arbitrariamente, elegimos definir una duración de 15 minutos como un episodio significativo de FC rápida para garantizar que el algoritmo (descripto más adelante) lograra extraer más información para determinar si el episodio de taquicardia reflejaba FA con RVR u otra forma de ritmo rápido (como taquicardia sinusal). Esto no significa que un paciente tenga que tener 15 minutos de FA con RVR antes de que el médico decida iniciar el tratamiento en la práctica clínica. En realidad, es una medida para reducir el ruido inherente de los ritmos cardíacos rápidos aislados. Uno puede experimentar usando diferentes valores de puntos de corte y luego revisar los resultados para determinar el umbral más apropiado.

Luego de identificar la aparición de un episodio de FC rápida de 15 minutos o más de duración, determinamos si el paciente había recibido alguno de los agentes farmacológicos de interés para su control dentro de las 2 horas pre o post episodio. Usamos una ventana de 2 horas porque los datos de medicación y los de FC corresponden a dos entidades diferentes, y los registros de fecha y hora de cada evento podrían no estar perfectamente sincronizados. Además, los registros de fecha y hora asociados con la medicación pueden estar sujetos a errores de registro por parte de las personas que ingresaron los datos. Esta ventana se determinó arbitrariamente; una ventana más pequeña podría haber incrementado la especificidad pero reducido la sensibilidad para detectar la cohorte de interés, y *vice versa* para una ventana más grande.

Un criterio importante para determinar la efectividad de un agente farmacológico en el control de la FA con RVR es el tiempo hasta la finalización del episodio arrítmico. Como esta información no se encuentra explícitamente en la base de datos, uno tiene que definir cuándo una FC está "controlada" y luego correr el algoritmo para encontrar el período de tiempo entre el inicio y la resolución de la RVR. La vida media del metoprolol y diltiazem endovenosos es aproximadamente de 4 horas para ambos; entonces, definimos la resolución de la RVR como la persistencia de una FC menor a 110 lpm por 4 horas. Aunque no hay consenso para esta definición, en la medida que la misma sea consistente dentro de cada sujeto o sub-

cohorte, pueden realizarse comparaciones. Nuestro algoritmo encuentra todo registro de FC menor a 110 lpm luego del episodio de RVR por FA previamente identificado (episodios de FC rápida de una duración de 15 minutos o más que fueron tratados con al menos uno de los agentes farmacológicos de interés) y testea si los datos de FC en las siguientes 4 horas son menores a 110 lpm por al menos el 90% del tiempo. De esta manera puede calcularse el período de tiempo entre el inicio y la resolución.

Las covariables se extrajeron usando SQL. Estas incluyeron información demográfica, signos vitales, laboratorios metabólicos básicos, condiciones médicas previas, puntajes de severidad de enfermedad, tipos de admisión y tipos de UCI. También revisamos la medicación ambulatoria de los pacientes y la historia previa de FA. Estos datos fueron extraídos de las secciones "medicación ambulatoria" y "antecedentes" de los resúmenes de egreso usando técnicas de procesamiento de lenguaje natural para la búsqueda en segmentos específicos de los resúmenes. La figura 23.2 es un ejemplo que nuestro grupo usó para la discusión del modelo analítico.

Aunque identificamos 1876 pacientes tratados por FA con RVR, sólo 320 habían recibido diltiazem como primera droga para el control de la FC. Usar análisis de regresión convencional habría resultado en un sobre-ajuste por el pequeño tamaño muestral, y eliminar variables habría introducido sesgos. Se usó el análisis por puntaje de propensión para reducir la dimensionalidad. El primer escalón es estimar el puntaje de propensión (probabilidad de ser asignado a un grupo de tratamiento dadas las covariables pre-tratamiento). Como fuera mencionado, hay diferentes maneras de estimar los puntajes de propensión, incluyendo métodos paramétricos como la regresión logística multinomial y no paramétricos, como los árboles de predicción. Pueden implementarse técnicas de aprendizaje automático para entrenar el modelo del puntaje de propensión para optimizar la predicción. Una vez estimado el puntaje de propensión, puede usarse tanto como una variable más en el modelo regresión para parear los sujetos con similares puntajes de propensión entre grupos de tratamiento o para calcular las ponderaciones inversas de probabilidad. Cuando estimamos los puntajes de propensión, más allá de optimizar el modelo para predecir mejor la posible asignación de tratamiento dadas las variables pre-tratamiento, un concepto más actual es estimar los puntajes de propensión para eliminar de manera balanceada las covariables pre-tratamiento luego de parear o ponderar. Cuando se usa

ponderación por puntaje de propensión, uno puede elegir usar ponderaciones directas o estabilizadas. La ponderación directa es más susceptible a la presencia de *outliers* con combinaciones singulares de covariables pre-tratamiento y duplicará el tamaño de la cohorte cuando haya dos grupos de tratamiento o la triplicará cuando haya tres. Por otro lado, la ponderación estabilizada es menos susceptible a outliers, y no incrementa el tamaño de la cohorte más allá del número de grupos de tratamiento.

Fig. 23.2. Discusiones grupales sobre el modelo analítico. Las flechas verdes representan el modelo final, y las rojas representan el modelo usado para el análisis de sensibilidad.

Para este estudio elegimos un algoritmo de aprendizaje automático (un modelo potenciado generalizado) para construir un árbol de regresión para la estimación de los puntajes de propensión (un método no paramétrico). La razón por la cual no elegimos un método paramétrico fue la misma por la cual no usamos un análisis de regresión convencional, como fuera explicado previamente. El modelo combina iterativamente una serie de árboles de regresión simples hasta que las métricas predeterminadas para la evaluación del desbalanceo de covariables pre-tratamiento (sesgo estandarizado o test de Kolmogorov-Smirnov) alcanzan un valor mínimo.

Las ponderaciones extremas se eliminaron usando ponderación estabilizada. La ponderación estabilizada se implementó en la regresión ponderada final para el testeo de hipótesis. Dependiendo de la naturaleza de la variable de resultado, se usó una regresión logística ponderada para

resultados binarios y regresión lineal ponderada para resultados reportados como variables continuas. Varias covariables con un poder predictivo superior (de asignación de tratamiento) se incluyeron en el modelo ponderado final.

23.6 Análisis del estudio

El uso más común del análisis por puntaje de propensión ha sido comparar dos grupos de tratamiento, por ej, grupo tratamiento vs. grupo control. También se usa comúnmente para la estratificación (usando el puntaje de propensión como una covariable en un modelo de regresión) y el pareamiento por puntaje de propensión (a través de la creación de grupos de control y de tratamiento con atributos pre-tratamiento similares y por lo tanto, simulando ensayos aleatorizados). De todas maneras, la estratificación sólo puede establecer una relación de asociación y el pareamiento por puntaje de propensión tiene como objetivo sobre todo servir como estrategia para reducir la dimensionalidad. Si bien el pareamiento por puntaje de propensión permite avanzar hacia la inferencia causal, cuando hay que parear tres o más grupos de tratamiento, requiere calcular dos o más distancias dimensionales para cada grupo de sujetos pareados, lo que puede ser matemáticamente desafiante y carece de teoría que lo respalde. Por eso, para nuestro puntaje de propensión elegimos árboles de regresión generados por aprendizaje automático y usamos un modelo de regresión ponderado por un puntaje de propensión para el efecto del resultado. El enfoque no paramétrico evitó las limitaciones y los sesgos introducidos por la especificación del modelo al usar métodos paramétricos. Luego de generar la ponderación del puntaje de propensión, realizamos una regresión ponderada. Esto nos permitió la exploración de términos de interacción y el ajuste de variables con efectos más significativos en los resultados que no podrían ser totalmente eliminados por el uso aislado de un puntaje de propensión.

Para validar nuestro modelo, realizamos una serie de análisis de sensibilidad usando pareamiento por puntaje de propensión de a pares, observándose efectos similares de diferentes grupos de tratamiento en los resultados

23.7 Resultados del estudio

En este estudio retrospectivo de un único centro, el metoprolol intravenoso fue el agente de control de la FC más comúnmente usado para el tratamiento de la FA con RVR en pacientes internados en la unidad de cuidados intensivos. Usando un enfoque novedoso basado en el pareamiento por puntaje de propensión, la efectividad del metoprolol se comparó con otros agentes farmacológicos comúnmente usados para esta indicación: diltiazem y amiodarona. En relación al resultado principal de fallo terapéutico (definido como la necesidad de cambio de tratamiento o agregado de un segundo agente), el metoprolol tuvo la tasa de fallo global más baja. Aquellos pacientes que recibieron diltiazem (odds ratio OR 1.55, intervalo de confianza IC 1.05-2.3, p = 0.027) o amiodarona (OR 1.50, IC 1.1-2.0, p = 0.006) como agente farmacológico primario tuvieron una probabilidad superior de recibir un agente adicional antes de la finalización del episodio de RVR. En un análisis secundario de pacientes que recibieron sólo una droga durante su episodio de RVR, aquellos que fueron tratados con diltiazem tuvieron tiempos significativamente más prolongados hasta la resolución del episodio de RVR. De la misma manera, los pacientes que recibieron sólo diltiazem fueron menos proclives a lograr el control de la RVR a las 4 horas que aquellos que sólo recibieron metoprolol (OR 0.59, IC 0.40-0.86, p=0.007).

Estos resultados sugieren que los pacientes críticamente enfermos con FA con RVR tienen menos probabilidad de requerir un segundo agente y más probablemente controlen el cuadro dentro de las 4 horas si reciben metoprolol como agente de control de la RVR comparándolo con diltiazem o amiodarona. Este efecto parece más pronunciado cuando se compara metoprolol con diltiazem.

23.8 Conclusiones

Si bien está ampliamente aceptado que la FA con RVR en la UCI se asocia con peores resultados globales, no hay consenso claro con respecto al manejo farmacológico óptimo y la práctica es altamente variable entre médicos. A través del uso de un modelo de pareamiento por puntaje de propensión de tres vías comparamos los agentes farmacológicos más comúnmente usados para este fenómeno y encontramos evidencia de que iniciar el tratamiento con metoprolol puede conllevar menos fracasos terapéuticos y una resolución más rápida del episodio de RVR.

La teoría de los puntajes de propensión se implementa más comúnmente en estudios que comparan dos grupos de tratamiento. La estimación de puntajes de propensión en estudios con múltiples grupos de tratamiento y su implementación en inferencia causal pueden ser especialmente desafiantes, tanto desde el punto de vista matemático como estadístico. En este capítulo brindamos un ejemplo de análisis con puntaje de propensión en un caso de múltiples grupos de tratamiento usando un algoritmo de aprendizaje automático. Los conceptos explorados en este capítulo pueden implementarse fácilmente en estudios de dos grupos de tratamiento. También proporcionamos un ejemplo de análisis por puntaje de propensión en dos grupos de tratamiento con el análisis de sensibilidad de nuestro estudio, a través de la realización de comparaciones de a pares entre grupos de tratamiento diferentes. El análisis de puntaje de propensión puede ser una manera robusta de lograr inferencia causal y reducir la dimensionalidad en estudios que usan HCEs.

23.9 Próximos pasos

La estrategia de análisis de datos usada en este proyecto puede ser particularmente útil para contestar una serie de preguntas de investigación en el ámbito de la UCI. Los médicos intensivistas frecuentemente tienen que seleccionar entre una amplia gama de intervenciones o de agentes farmacológicos. A diferencia del pareamiento tradicional por puntaje de propensión, donde sólo dos grupos se comparan, este modelo permite la comparación simultánea de tres grupos independientes. Entre los ejemplos donde este enfoque de análisis puede ser útil se incluye la comparación de la efectividad de diferentes vasopresores en el tratamiento del shock o diferentes agentes sedantes para pacientes con SDRA en ventilación mecánica invasiva.

Dado el grado de equivalencia clínica con respecto al tratamiento de la FA con RVR en la UCI, los resultados previos resultan de utilidad para guiar a los médicos que se enfrentan con este complejo problema clínico. Aún así, algunas preguntas subsisten. No está claro, por ejemplo, si dosis más altas de diltiazem podrían haber sido más efectivas y, por lo tanto, reducir la tasa de fallo terapéutico. Nosotros no revisamos las dosis usadas en este estudio. Tampoco exploramos la vía de administración oral versus intravenosa versus

combinada. La fibrilación auricular durante la enfermedad crítica es un fenómeno común cuyo manejo requiere mayor investigación.

Apéndice: Código

El código usado en este capítulo se encuentra disponible en el repositorio GitHub de este libro: https://github.com/MIT-LCP/critical-data-book. Información adicional sobre el código está disponible en ese sitio web. Los siguientes scripts fueron usados para la investigación comentada en este capítulo:

- `database_query.sql`: utilizado para extraer datos de la base de datos MIMIC II.
- `data_extraction.m`: utilizado para extraer variables para el análisis.
- `propensity_puntaje_analysis.r`: utilizado para el análisis de puntaje de propensión.
- `propensity_puntaje_matching.r`: utilizado para el pareamiento por puntaje de propensión.

Referencias

1. Patorno E et al (2014) Studies with many covariates and few outcomes: selecting covariates and implementing propensity-puntaje-based confounding adjustments. Epidemiology 25 (2): 268-278.
2. Fitzmaurice G (2006) Confounding: propensity puntaje adjustment. Nutrition 22 (11-12): 1214-1216.
3. Austin PC (2011) An introduction to propensity puntaje methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 46 (3): 399-424.
4. Li L et al (2011) Propensity puntaje-based sensitivity analysis method for uncontrolled confounding. Am J Epidemiol 174 (3): 345-353.

5. Toh S, Garcia Rodriguez LA, Hernan MA (2011) Confounding adjustment via a semi-automated high-dimensional propensity puntaje algorithm: an application to electronic medical records. Pharmacoepidemiol Drug Saf 20 (8): 849-857.
6. Guertin JR et al (2015) Propensity puntaje matching does not always remove confounding within an economic evaluation based on a non-randomized study. Value Health 18 (7): A338.
7. Girman CJ et al (2014) Assessing the impact of propensity puntaje estimation and implementation on covariate balance and confounding control within and across important subgroups in comparative effectiveness research. Med Care 52 (3): 280-287 8. Glass TA et al (2013) Causal inference in public health. Annu Rev Public Health 34:61-75.
9. Cousens S et al (2011) Alternatives to randomisation in the evaluation of public-health interventions: statistical analysis and causal inference. J Epidemiol Community Health 65 (7): 576-581.
10. Brookhart MA et al (2013) Propensity puntaje methods for confounding control in nonexperimental research. Circ Cardiovasc Qual Outcomes 6 (5): 604-611.
11. Feng P et al (2012) Generalized propensity puntaje for estimating the average treatment effect of multiple treatments. Stat Med 31 (7): 681-697.
12. Rosthoj S, Keiding N (2004) Explained variation and predictive accuracy in general parametric statistical models: the role of model misspecification. Lifetime Data Anal 10 (4): 461-472.
13. Ertefaie A, Asgharian M, Stephens D (2014) Propensity puntaje estimation in the presence of length-biased sampling: a nonparametric adjustment approach. Stat 3 (1): 83-94.
14. Yoo C, Ramirez L, Liuzzi J (2014) Big data analysis using modern statistical and machine learning methods in medicine. Int Neurourol J 18 (2): 50-57.
15. Hsu DJ et al (2015) The association between indwelling arterial catheters and mortality in hemodynamically stable patients with respiratory failure: a propensity puntaje analysis. Chest 148 (6): 1470-1476.
16. Hernan MA (2012) Beyond exchangeability: the other conditions for causal inference in medical research. Stat Methods Med Res 21 (1): 3-5
17. Austin PC, Stuart EA (2014) The performance of inverse probability of treatment weighting and full matching on the propensity puntaje in the presence of model misspecification when estimating the effect of treatment on survival outcomes. Stat Methods Med Res.
18. Pirracchio R, Petersen ML, van der Laan M (2015) Improving propensity puntaje estimators' robustness to model misspecification using super learner. Am J Epidemiol 181 (2): 108-119.
19. Lee BK, Lessler J, Stuart EA (2010) Improving propensity puntaje weighting using machine learning. Stat Med 29 (3): 337-346.
20. Brookhart MA et al (2006) Variable selection for propensity puntaje models. Am J Epidemiol 163 (12): 1149-1156.

21. Zhu Y et al (2015) Variable selection for propensity puntaje estimation via balancing covariates. Epidemiology 26 (2): e14-e15.
22. Saeed M et al (2011) Multiparameter intelligent monitoring in intensive care II: a public-access intensive care unit database. Crit Care Med 39 (5): 952-960.
23. Goldberger AL et al (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101 (23): E215-E220.

Capítulo 24

MODELOS DE MARKOV Y ANÁLISIS DE COSTO EFECTIVIDAD: APLICACIONES EN LA INVESTIGACIÓN MÉDICA

Matthieu Komorowski y Jesse Raffa

Objetivos de aprendizaje

Conocer cómo pueden usarse los modelos de Markov para analizar las decisiones médicas y desarrollar análisis de costo efectividad.

Este caso de estudio introduce conceptos que deberían mejorar la comprensión de los siguientes temas:

1. Modelos de Markov y su uso en la investigación médica
2. Conceptos básicos de economía sanitaria
3. Replicar los resultados de un estudio ramdomizado controlado prospectivo utilizando una cadena de Markov y simulaciones de Monte Carlo y
4. Relacionar años de vida ajustados por calidad (AVAC) y los costos de las intervenciones para cada estado de la cadena de Markov, para llevar a cabo un análisis simple de costo efectividad.

24.1 Introducción

Los modelos de Markov fueron teorizados inicialmente en los comienzos del siglo 20 por el matemático ruso Andrey Markov [1]. Son procesos estocásticos que se someten a transiciones de un estado a otro. A través de los años se han utilizado en múltiples aplicaciones, especialmente para modelar procesos y tomar decisiones, en el campo de la física, teoría de colas, finanzas, ciencias sociales, estadísticas y por supuesto también en medicina.

Los modelos de Markov son útiles para modelar escenarios y problemas que involucran decisiones estocásticas secuenciales en el tiempo. Representar estos escenarios con árboles de decisión sería confuso o incluso intrincado, si fuera posible, y requeriría una simplificación mayor de los supuestos [2].

Los modelos de Markov pueden ser analizados por un conjunto de herramientas que incluyen el álgebra linear (fuerza bruta), simulaciones de

cohorte, simulaciones de Monte Carlo y, para los Procesos de Decisión de Markov, programación dinámica y aprendizaje por refuerzo (*reinforcement learning*) [3, 4].

Una propiedad fundamental de todos los modelos de Markov es su **falta de memoria.** Satisfacen una **Propiedad de Markov de primer orden** si la probabilidad de mover un nuevo estado a s_{t+1} sólo depende del estado actual s_t y no de otro estado previo, siendo t el estado actual. Dicho de otra forma, dado el estado actual, el estado futuro y el estado pasado son independientes. Formalmente, un proceso estocástico tiene la Propiedad de Markov de primer orden si la distribución de la probabilidad condicional de los estados futuros del proceso (condicional tanto de los valores presentes como pasados) depende solamente del estado actual:

$$P\,(S_{t+1} \mid S_1, S_2, \ldots, S_t) = P\,(S_{t+1} \mid S_t)$$

Este capítulo brindará una breve introducción a los modelos de Markov más comunes y se esbozarán algunas aplicaciones potenciales en la investigación médica y en la economía sanitaria. La última sección discutirá un ejemplo práctico inspirado en la literatura médica, en el cual se usará una cadena Markov para realizar el análisis de costo efectividad de una intervención médica particular. En general los resultados crudos de un estudio no pueden brindar la información necesaria para implementar un análisis de costo efectividad, lo que demuestra el valor de formular el problema como una Cadena de Markov.

24.2 Formalización de los Modelos de Markov comunes

Los cuatro modelos de Markov más comunes se muestran en la tabla 24.1. Pueden clasificarse en 2 categorías dependiendo de si su estado secuencial completo es observable [5]. Además, en los Procesos de Decisión de Markov, las transiciones de un estado a otro se encuentran bajo el mando de un sistema de control llamado el agente, que selecciona acciones que pueden llevar a un estado subsecuente particular. En contraste, en las cadenas de Markov y en los modelos ocultos de Markov, la transición entre los estados es autónoma. Todos los modelos de Markov son finitos

(discretos) o continuos, dependiendo de la definición de su espacio de estado.

Tabla 24.1 Clasificación de Modelos de Markov

	Sistema completamente observable	Sistemas parcialmente observables
Sistema autónomo	Cadena de Markov (MC por sus siglas del inglés Markov Chains)	Modelo Oculto de Markov o HMM (por sus siglas en inglés, Hidden Markov Model)
Sistema que contiene un control de proceso	Proceso de Decisión de Markov (MDP por sus siglas del inglés Markov Decisión Process)	Proceso de Decisión de Markov Parcialmente observable (POMDP, del inglés Partially Observable Markov Decisión Process)

24.2.1 La Cadena de Markov

La Cadena de Markov discreta en el tiempo, definida por la tupla {*S, T*} es el modelo de Markov más simple, donde *S* es un set de estados finitos y *T* es una matriz de probabilidad de transición de estados,

$$T(s'; s) = P(s_{t+1}=s' \mid s_t=s)$$

Una cadena de Markov puede ser **ergódica**, si es posible ir de un estado a cualquier otro en una cantidad finita de movimientos. La figura 24.1 muestra un ejemplo simple de Cadena de Markov.

En la matriz de transición, las entradas en cada columna se encuentran entre 0 y 1 (inclusive) y su suma es 1. Este tipo de vectores se denominan **vectores de probabilidad.** La tabla 24.2 muestra la matriz que corresponde a la figura 24.1**.** Un estado se dice **absorbente** si es imposible dejarlo (por ej. la muerte)

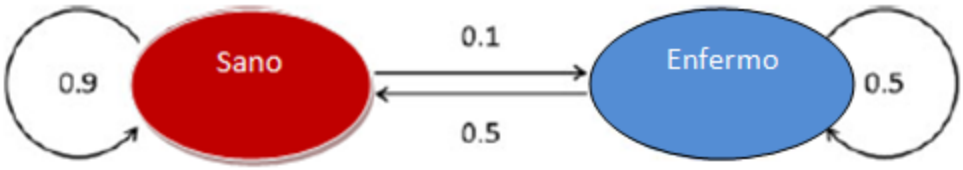

Figura 24.1 Ejemplo de cadena de Markov definida por un set *S* de estados finitos (Sano y Enfermo) y una matriz de transición que contiene las probabilidades de moverse de un estado *s* al siguiente estado s' en cada iteración

Tabla 24.2 Ejemplo de una matriz de transición que corresponde a la figura 24.1

		Nuevo estado s		Total
		Sano	Enfermo	
Estado Inicial s	Sano	0.9	0.1	1
	Enfermo	0.5	0.5	1

24.2.2 Explorando las cadenas Markov con Simulaciones Monte Carlo

Las simulaciones de Monte Carlo (MC) son útiles para explorar y comprender fenómenos y sistemas modelados bajo un modelo de Markov. Las simulaciones MC generan variables pseudoaleatorias en una computadora para aproximar cantidades difíciles de estimar. Se usa ampliamente en muchos campos y aplicaciones [6]. Nuestro enfoque está en la simulación de MC de una cadena de Markov y es simple una vez que se han definido una matriz de probabilidad de transición T (S^1, S) y el tiempo final t*. Asumiremos que al tiempo índice (t=0), el estado es conocido y lo llamaremos S_0. En t=1 simularemos una variable aleatoria categórica utilizando la fila S_0 de la matriz de transición de probabilidad T (S^1, S). Repetiremos esto t=1,2, …; t*-1; t* para simular *una instancia simulada* de la cadena de Markov que estamos estudiando. Una instancia simulada solamente nos dice acerca de la posible secuencia de transiciones dentro de las tantas para esta cadena de Markov, y necesitamos repetir esto muchas (N) veces, registrando la secuencia de estados para cada una de las instancias simuladas. Repetir este proceso varias veces, nos permite estimar

valores como: la probabilidad en t=5, que la cadena está en estado 1; la proporción promedio de tiempo pasado en estado 1 dentro de los primeros 10 puntos de tiempo; o la duración media de la cadena consecutiva más larga en el estado 1 en los primeros puntos de tiempo t*.

Utilizando el ejemplo mostrado en la Fig. 24.1 estimaremos la probabilidad de que un individuo se encuentre sano o enfermo en 5 días, sabiendo que hoy está sano. Los métodos de MC simularán un gran número de muestras (ej. 10000) iniciando en S_0= sano y siguiendo la matriz de transición *T (S*^{1}*, S*) para 5 pasos, seleccionando secuencialmente transiciones a $S^{1,}$ de acuerdo a su probabilidad. La variable resultado (el valor del estado final) se registra para cada muestra y concluimos analizando las características de la distribución de esta variable resultado (Tabla 24.3). La distribución del estado final al día + 5 para 10000 instancias simuladas es representada en la Fig. 24.2.

La Tabla 24.4 informa algunas características de la muestra para estado "sano" en el día 5 para 100 y 10000 instancias simuladas, que ilustran por qué es importante simular un gran número de muestras.

Tabla 24.3 Ejemplo de predicción de salud mediante el uso de simulación de Montecarlo.

	Instancia 1	Instancia 2	…	Instancia 10000
Hoy	Sano	Sano		Sano
Día+1	Sano	Sano		Sano
Día+2	Sano	Enfermo		Sano
Día+3	Sano	Enfermo		Enfermo
Día+4	Sano	Enfermo		Sano
Día+5	Sano	Enfermo		Sano

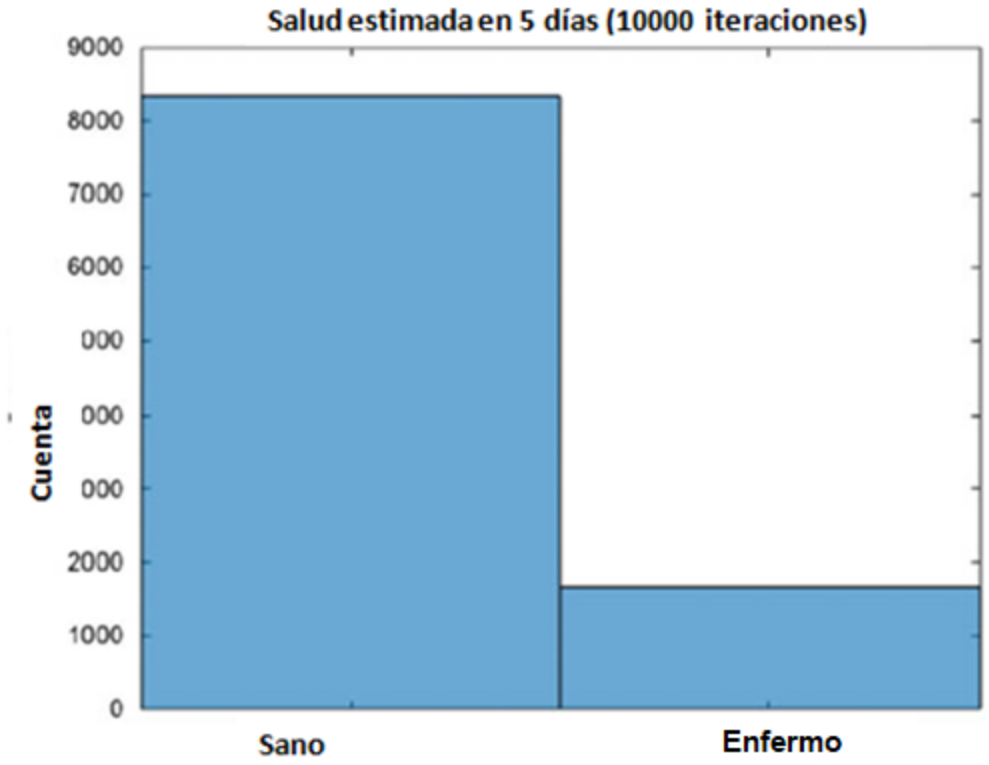

Figura 24.2 Distribución de la salud en día 5 para 10000 instancias

Tabla 24.4 Características de la muestra para 100 y 10000 instancias simuladas

	100 instancias simuladas	10000 instancias simuladas
Media	0,81	0.83
Desvío estándar	0.39	0.37
Intervalo de confianza 95% para la media	0.73-0.89	0.83-0.84

Aumentando el número de instancias simuladas, aumentamos drásticamente nuestra certeza de que la media verdadera de la muestra se encuentra dentro de un margen muy estrecho (0.83-0.84 en este ejemplo). La media verdadera calculada analíticamente es 0.838 que es muy cercano al estimado generado por la simulación de MC.

24.2.3 Proceso de decisión en Markov y Modelos Ocultos de Markov (Hidden Markov Models)

Los procesos de decisión de Markov (MDPs) brindan un escenario para ejecutar métodos de aprendizaje por refuerzo. Los MDPs son una extensión de las cadenas de Markov, que incluyen un proceso de control. Los MDPs son una técnica poderosa y apropiada para modelar decisiones médicas [3]. Los MDPs tienen mayor utilidad en problemas que involucran **decisiones complejas, estocásticas y dinámicas como decisiones de tratamientos médicos**, para las que pueden encontrar soluciones óptimas [3]. Los médicos siempre necesitaran realizar juicios subjetivos acerca de estrategias de tratamiento, pero los modelos de decisiones matemáticos pueden brindar un acercamiento a la naturaleza de las elecciones óptimas y guiar decisiones de tratamiento. En los modelos ocultos de Markov (HMMs) el estado espacio es sólo parcialmente observable [7]. Está formado por dos procesos estocásticos dependientes (Fig 24.3). El primero es una cadena de Markov clásica, cuyos estados no son observables en forma directa externamente, por lo tanto "ocultos". El segundo proceso estocástico genera emisiones observables, condicionales al proceso oculto. Se han desarrollado metodologías para decodificar los estados ocultos de los datos observados y tienen aplicaciones en una multiplicidad de áreas [7].

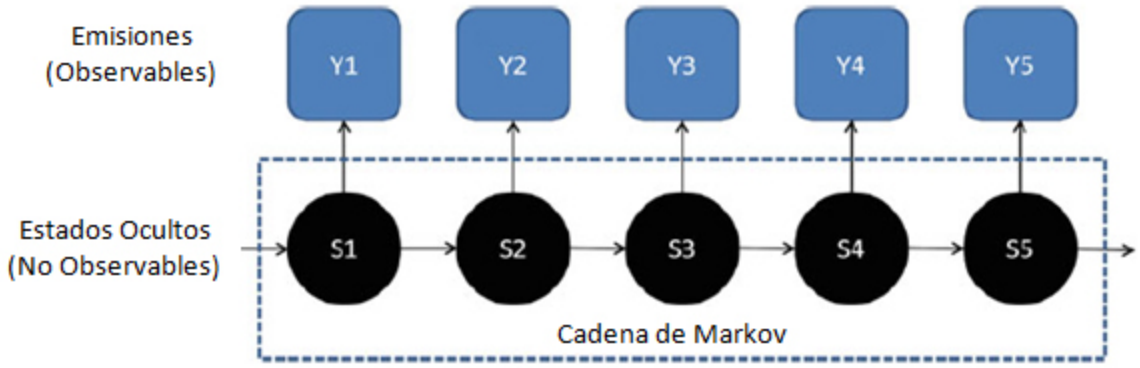

Figura 24.3 Ejemplo de Modelo Oculto de Markov (HMM)

24.2.4 Aplicaciones Médicas de Modelos de Markov

Los procesos de decisión de Markov (MDPs) han sido reconocidos por los autores como un enfoque poderoso y apropiado para modelar secuencias de decisiones médicas [3]. Los modelos controlados de Markov pueden ser resueltos por algoritmos como la programación dinámica o el aprendizaje por refuerzo, que pretenden identificar o aproximarse a la política óptima

(conjunto de reglas que maximizan la suma esperada de recompensas ofrecidas). En la literatura médica los modelos de Markov han explorado problemas muy diversos como el momento adecuado para el trasplante hepático [8], tratamiento para HIV [9], cáncer de mama [10], hepatitis C [11], terapia con estatinas [12] o manejo del alta hospitalaria [5, 13]. Los modelos de Markov pueden ser usados para describir varios estados de salud en una población de interés y para detectar los efectos de distintas políticas o elecciones terapéuticas. Por ejemplo Scott y col han usado un HMM para clasificar pacientes en 7 estados de salud que corresponden a los efectos adversos de 2 medicaciones psicotrópicas [14]. Las transiciones fueron analizadas para especificar qué droga estaba asociada con menos efectos adversos. Recientemente, se propuso una cadena de Markov para modelar la progresión de la retinopatía diabética, usando 5 estados predefinidos, desde retinopatía leve a ceguera [15]. Los MDPs también se han usado en aplicaciones de imágenes médicas. Alterovitz ha usado grandes MDPs (800000 estados) para planificar el movimiento en la dirección de la aguja guiada por imagen [16]. Aparte de esas aplicaciones médicas, los modelos de Markov se usan ampliamente en investigación de economía de la salud, que es el foco de la siguiente sección de este capítulo.

24.3 Fundamentos de Economía de la Salud

24.3.1 Los Objetivos de la Economía de la salud: Maximizando la costo-efectividad

Esta sección brindará al lector una introducción mínima de economía sanitaria seguida de un ejemplo práctico. La economía sanitaria tiene como objetivo maximizar el "valor del dinero" en los cuidados de la salud, optimizando no solamente la efectividad clínica sino también la costo-efectividad de las intervenciones médicas. Como fue explicado por Morris: "*Obtener 'valor por dinero' implica tanto un deseo de alcanzar un objetivo predeterminado a menor costo o un deseo de maximizar [sic] el beneficio de la población de pacientes atendidos con una cantidad limitada de recursos"* [17].

En la economía de la salud pueden distinguirse dos enfoques principales: análisis de costo minimización y de costo efectividad (ACE). En ambos casos

el propósito es idéntico: identificar la opción de tratamiento más efectiva. La costo-minimización se ocupa del caso simple en que distintas opciones de tratamiento disponibles tienen la misma efectividad pero distintos costos. Casi en forma lógica, la costo-minimización favorecerá la opción más económica. El ACE representa el escenario más probable y se usa más ampliamente. En el ACE se comparan distintas opciones con costos diferentes y distinta efectividad. El análisis calculará el costo relativo de una mejora en salud y las métricas para informar en forma adecuada a los decisores.

24.3.2 Definiciones

Midiendo los resultados: sobrevida, Calidad de vida (QoL, del inglés Quality of Life), Años de vida ajustados por calidad (AVAC)

Los resultados son evaluados en términos de mayor supervivencia ("*suma de años de vida*") y mejoramiento de la calidad de vida (QoL) ("*suma de vida a los años*") [17]

A pesar de que algunas veces el concepto de años de vida ajustados por calidad (AVAC) es criticado, continua siendo de central importancia en los análisis de costo-utilidad [18]. Los AVAC aplican un peso que refleja la QoL expresada por el paciente. Un AVAC equivale a un año en estado de perfecta salud. Perfecta salud es equivalente a 1 mientras que muerte es equivalente a 0. Los AVAC son estimados por varios métodos, incluyendo escalas y cuestionarios completados por los pacientes o examinadores externos [19]. Como ejemplo, el cuestionario EuroQoL EQ 5D evalúa la salud en 5 dimensiones: movilidad, autocuidado, actividades habituales, dolor/disconfort y ansiedad/depresión.

Tasa de costo-efectividad (CER, del inglés, cost– effectiveness ratio)

La tasa de costo-efectividad (CER) informará a los decisores del costo de una intervención en relación a los beneficios en la salud que genera la intervención. Por ejemplo una intervención que cuesta $20,000 por paciente y aporta 5 AVAC (5 años de perfecta salud) tiene un CER de $20,000/5=$4000 por AVAC. Esta medida permite una comparación directa de costo efectividad entre intervenciones.

Tasa de Costo-efectividad incremental (ICER, del inglés incremental cost-effectiveness ratio)

El ICER es una medida usada habitualmente en las publicaciones de economía sanitaria y permite comparar dos intervenciones diferentes en términos de "costo de la efectividad ganada". Se construye dividiendo la diferencia en costo de dos intervenciones por la diferencia de su efectividad [20]. Como ejemplo si el tratamiento A cuesta $5000 por paciente y brinda 2 AVAC y el tratamiento B cuesta $8000 mientras que provee 3 AVAC, el ICER del tratamiento B será

$$\frac{(\$8000 - \$5000)}{3 - 2} = \$3000$$

Dicho de otra manera costará $3000 más ganar 1 AVAC con el tratamiento B para una condición médica particular. El ICER puede informar al decisor acerca de la necesidad de adoptar o financiar una nueva intervención médica. En forma esquemática, si el ICER de una nueva intervención se encuentra debajo de un cierto límite, significa que se pueden adquirir beneficios en salud con un nivel aceptable de gasto.

El plano de costo-efectividad

El plano de costo-efectividad es una herramienta importante usada en ACE (Fig. 24.4). Su objetivo es ilustrar en forma clara diferencias en costos y efectos entre distintas estrategias, sea que comprendan intervenciones médicas, tratamientos o una combinación de ambas.

El plano de la CE consiste en un diagrama de 4 cuadrantes, donde el eje de las X representa el nivel incremental de la efectividad de un resultado y el eje de las Y representa el costo adicional total de implementar este resultado. Por ejemplo, mientras más se mueva hacia la derecha en el eje de las X, más efectivo será el resultado. En el cuadrante superior derecho, un tratamiento puede recibir financiamiento si su ICER se encuentra debajo del umbral aceptable máximo.

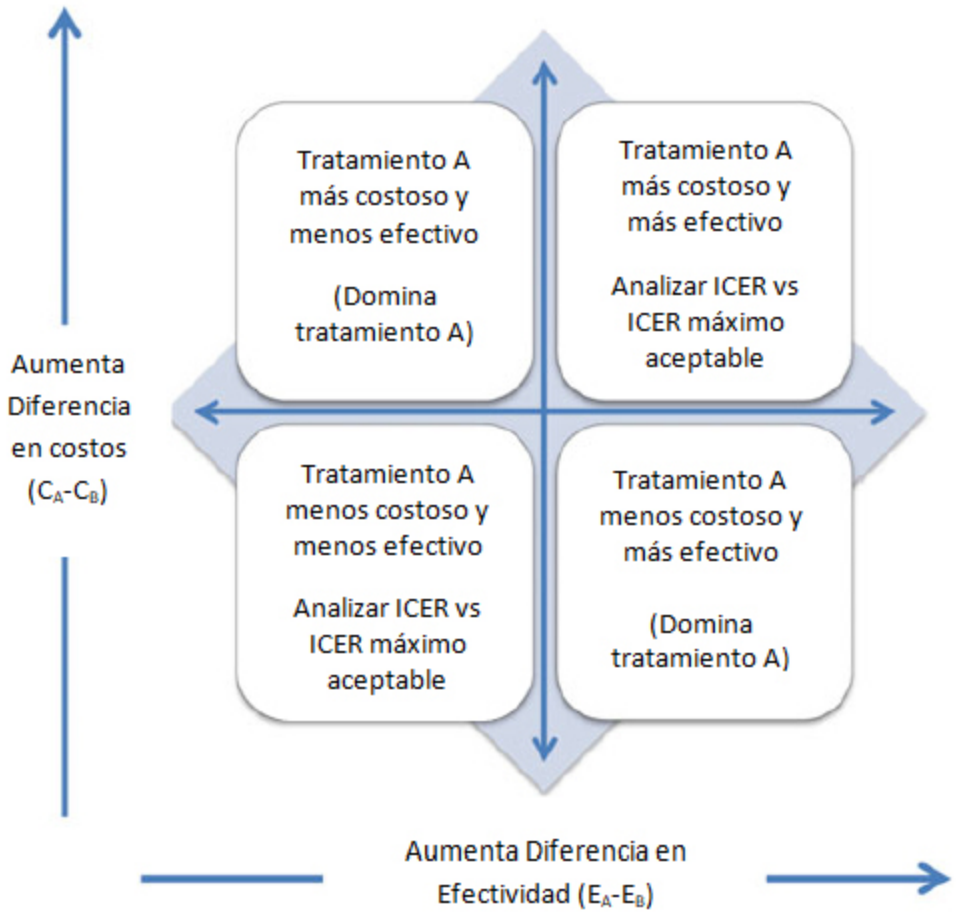

Figura 24.4 El plano de costo-efectividad, comparando el tratamiento A con el tratamiento B

24.4 Caso de estudio: simulaciones de Monte Carlo de una cadena de Markov para evaluar la interrupción diaria de sedación en cuidados intensivos, por análisis de costo-efectividad

Este ejemplo está inspirado en la publicación de Girard y col [21] y nos permitirá ilustrar como construir y examinar una cadena de Markov simple para representar una intervención médica, como relacionar AVAC y costo de intervenciones con cada estado de la cadena de Markov con la finalidad de realizar un análisis de costo efectividad. En este estudio prospectivo aleatorizado controlado, los autores evaluaron el impacto de la suspensión de sedación diaria en terapia intensiva con varios resultados como el número de días libres de ventilador, delirio y mortalidad a los 28 días. En la terapia intensiva, los pacientes generalmente reciben ventilación mecánica

en contexto de severa alteración de la conciencia, luego de procedimientos quirúrgicos severos y cuando presentan falla respiratoria severa.

En forma terapéutica, los pacientes son sedados para mejorar su confort. Aun así, una evidencia creciente de literatura ha identificado los riesgos de la sedación continua en la UCI dado que está asociada con aumento de la mortalidad, delirio, duración de la ventilación mecánica y estadía en la UCI y en el hospital [22]. Para lograr el equilibrio adecuado entre mantener la sedación y la ventilación mecánica el tiempo que el paciente lo necesite, pero también lograr la extubación lo antes posible, Girard y col propusieron despertar activamente a los pacientes en forma diaria para evaluar su preparación para salir del ventilador. Los principales resultados se muestran en la Tabla 24.5.

En este caso de estudio ejemplo, intentaremos aproximar esos resultados utilizando un cadena de Markov simple de 3 estados analizada por simulación de MC. Como ejercicio extenderemos el estudio a un ACE. Este tutorial brindará a los lectores las herramientas necesarias para implementar en otros contextos los métodos de simulación de MC con cadenas de Markov, y estudios de costo efectividad simples.

La mayoría de los resultados del estudio pueden alcanzarse usando una cadena de Markov muy cruda de 3 estados (Fig 24.5) con los siguientes estados de espacio {intubado, extubado, muerto}. En este modelo simplístico, son posibles solamente 7 transiciones y el estado 'muerte' es absorbente.

Tabla 24.5 Resultados principales del estudio original

	Grupo Intervención	**Grupo Control**
Días libres de ventilador (media)	14.7	11.6
Días libres de ventilador (mediana)	20.0	8.1
Pacientes extubados exitosamente a los 28 días (%)	≈93	≈88
Mortalidad a los 28 días (%)	29	35

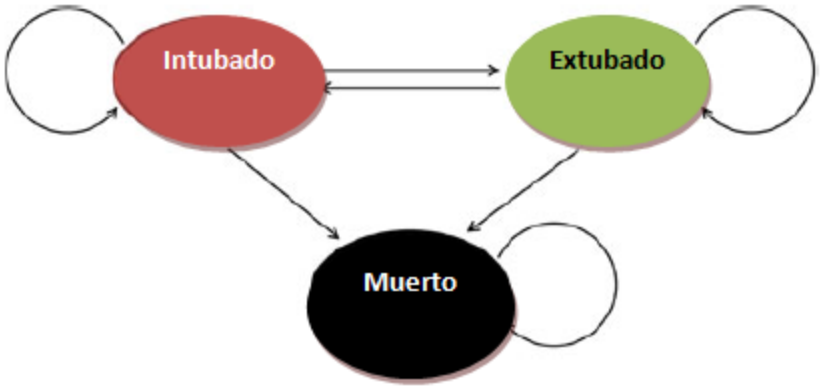

Figura 24.5 La cadena de Markov de 3 estados usada en este ejemplo

Dos matrices diferentes de transición pueden construirse por prueba y error correspondientes a las ramas de intervención y control del estudio (Tabla 24.6). Corresponden a las probabilidades diarias de pasar de un estado a otro. Los valores iniciales fueron seleccionados usando unos pocos supuestos simples: el estado "muerte" es absorbente, la probabilidad de permanecer intubado o extubado es mayor que la probabilidad de cambiar de estado, el riesgo de morir

Tabla 24.6 Matrices de transición usadas en el caso de estudio

Grupo Intervención		**Próximo Estado S'**		
		I	**E**	**M**
Estado Inicial S	**I**	0.862	0.12	0.018
	E	0.0088	0.982	0.0092
	M	0	0	1
Grupo Control		**Próximo Estado S'**		
		I	**E**	**M**
Estado Inicial S	**I**	0.878	0.1	0.022
	E	0.01	0.978	0.012
	M	0	0	1

I: intubado, E: extubado, M: muerto

Podemos ver que nuestra simulación sigue en forma muy cercana lo que teóricamente se sabe es verdadero. Para realizar un ACE, se debe asignar a cada estado un valor de AVACs y costo. Para este ejemplo, asumamos también los valores de AVAC y costos diarios que se muestran en la Tabla 24.7.

La Tabla 24.8 muestra los resultados de las primeras iteraciones para el grupo control cuando se inicia con 100 pacientes intubados (Función `IED_transition.m`). En cada paso de tiempo, el número de pacientes aún intubados corresponde a los pacientes que permanecieron intubados menos los pacientes extubados (probabilidad diaria de 10%) y aquellos que murieron (probabilidad de 2.2%), más los pacientes extubados que han debido reintubarse (probabilidad 1%). Luego de 28 días, la mortalidad acumulada alcanza 35.6% y la tasa de pacientes extubados entre los pacientes aún vivos es 88.8% por lo que coincide en forma cercana con los resultados del estudio original. En cada paso de tiempo, se calcula la suma de los AVAC y los costos para todos los pacientes así como sus valores acumulados. El número de AVAC inicialmente aumenta a medida que más pacientes son extubados y luego disminuye como consecuencia del número de pacientes que mueren.

Tabla 24.7 Definición de AVAC y costo diario para cada estado

Estado	**I**	**E**	**M**
AVAC	0.5	1	0
Costo diario ($)	2000	1000	0

I: intubado, E: extubado, M: muerto

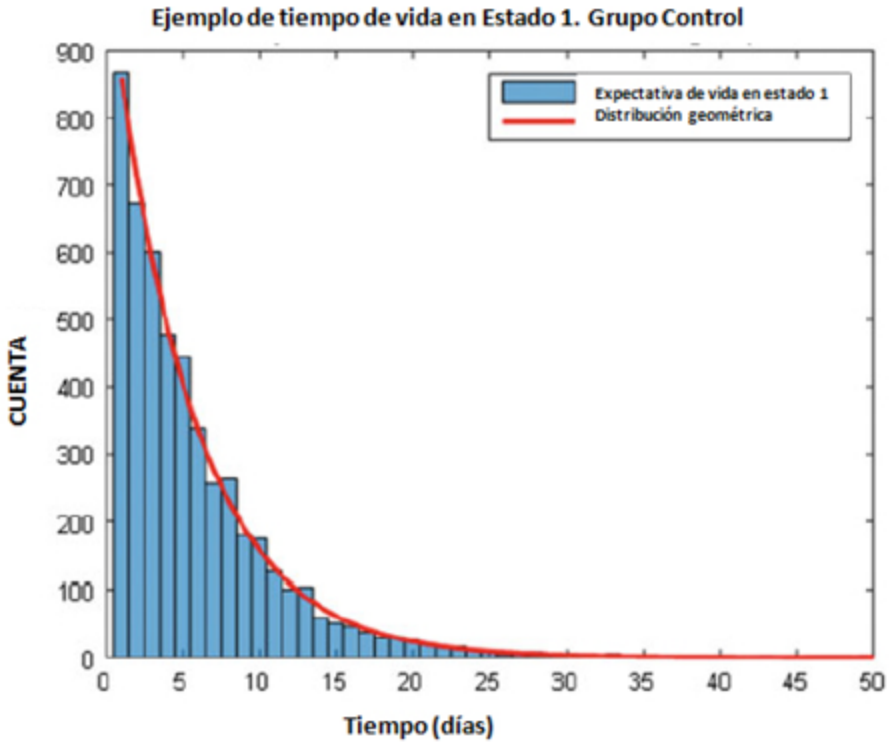

Figura 24.6 Ejemplo de la esperanza de vida en el estado "I" en el grupo de control, con distribución geométrica ajustada. El gráfico de barras representa la distribución del tiempo pasado en el estado "intubado" de la cadena de Markov, antes de pasar a otro estado, para 5000 muestras

Tabla 24.8 Número de pacientes en cada estado, AVAC y análisis de costos, durante 28 iteraciones (grupo control)

Día	I	E	M	Extubados/ Vivos	AVACs	AVACs acumulados	Costo Diario (K$)	Costo acumulado (K$)
0	100.00	0.00	0.00	0.00	50.00	50.00	200.00	200
1	87.80	10.00	2.20	0.10	53.90	103.90	185.60	386
2	77.19	18.56	4.25	0.19	57.15	161.05	172.94	559
3	67.96	25.87	6.17	0.28	59.85	220.90	161.78	720
4	59.92	32.10	7.98	0.35	62.06	282.96	151.95	872
5	52.94	37.38	9.68	0.41	63.85	346.81	143.25	1016
...	...	...	...	..	...	...	...	...
28	7.19	57.21	35.60	0.89	60.80	1863.84	71.59	3184

I: intubado, E: extubado, M: muerto

La siguiente figura representa la tasa de número de pacientes extubados en relación al número de pacientes vivos, en el tiempo y para ambas estrategias (Fig. 24.7). Puede compararse con la figura original en el artículo fuente.

Por medio de la simulación puede calcularse la distribución del número promedio de días libres de ventilador y sus características para ambas estrategias (`function MCMC-solver.m`). La siguiente Tabla 24.9 muestra ejemplos de estados de pacientes calculados usando la matriz de transición del grupo control. La distribución de los días libres de ventilador en nuestras 10000 muestras se grafica en la Fig. 24.8. El promedio y la mediana del número días libres de ventilador para ambos grupos se muestra en la Tabla 24.10.

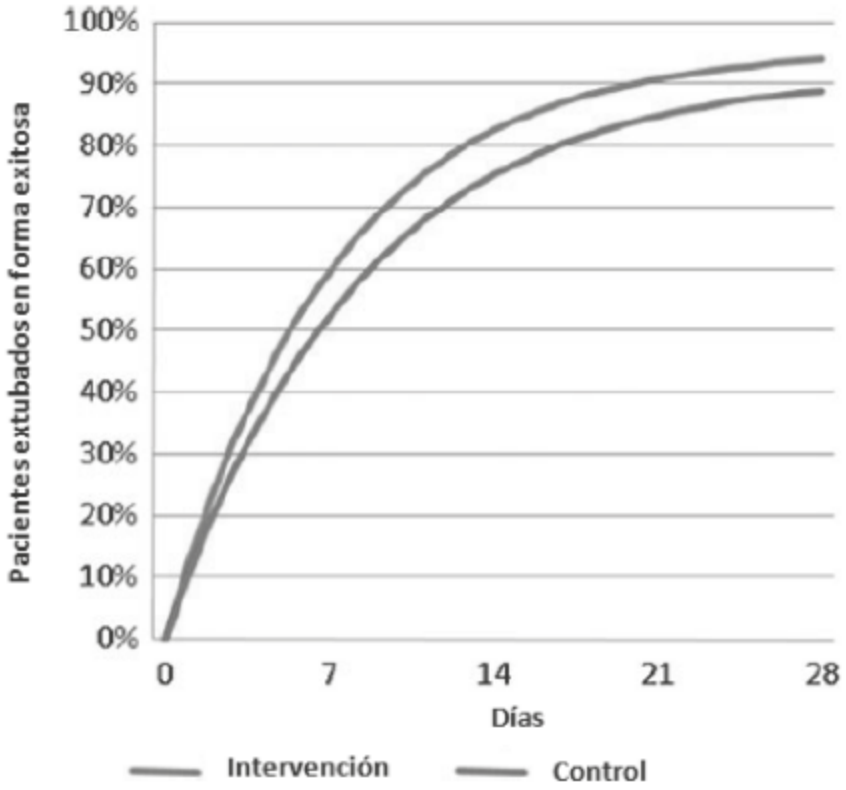

Figura 24.7 Resultado primario del estudio modelado usando una cadena de Markov

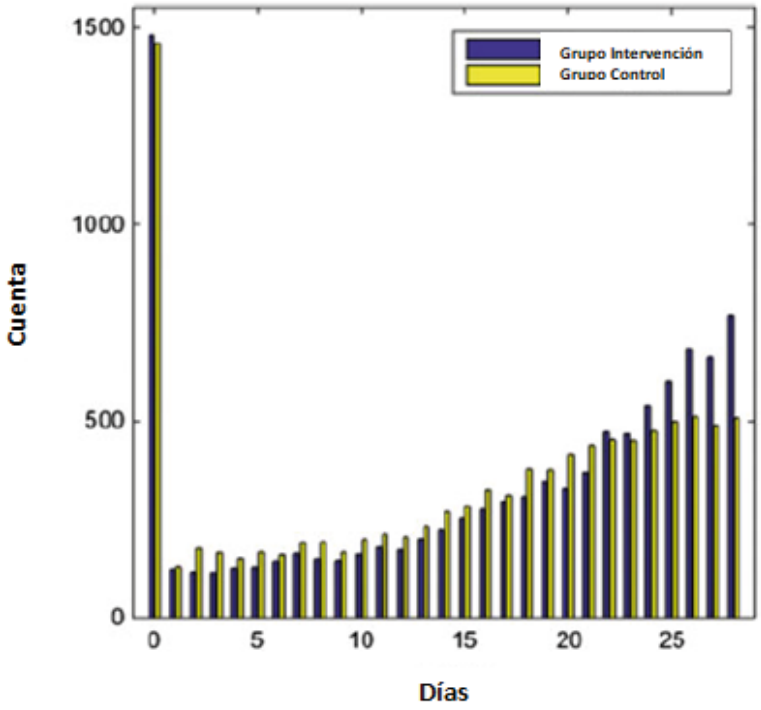

Fig 24.8 Días libres de Asistencia respiratoria mecánica para 10000 muestras para el grupo intervención y el grupo control

Tabla 24.9 Calculando el número de días libres de ventilador por Monte Carlo (10000 instancias simuladas)

Día	Instancia 1	Instancia 2	Instancia 3	....	Instancia 1000
0	I	I	I		I
1	I	I	I		I
2	I	I	I		I
3	I	I	I		I
4	I	I	I		I
5	I	I	I		I
6	I	I	I		I
7	I	I	I		E
8	E	E	I		E
9	E	E	I		E
10	I	E	I		E
...	...	...	...		...
28	M	M	M		E
Total de Días libres de ventilador	7	3	0		22

La tasa de Costo efectividad a los 28 días de ambas estrategias puede calcularse dividiendo el costo acumulado final por el AVAC acumulado (Tabla 24.11).

La intervención es más costosa, pero también se asocia con beneficios para la salud (significativamente más AVACs). Pertenece al cuadrante superior derecho del plano de CE, donde el ICER es usado para determinar costo-efectividad de una intervención.

Tabla 24.10 Promedio y mediana del número de días libres de ventilador para ambos grupos

Numero de días libres de ventilador	**Grupo Intervención**	**Grupo Control**
Promedio	17.1	15.9
Mediana	20	18

Tabla 24.11 Tasa de costo efectividad para ambos grupos

	Grupo Intervención	**Grupo Control**
Costo acumulado (K$)	3213	3184
AVACs acumulados	2029	1864
Tasa de costo efectividad acumulada ($ por AVAC)	1583	1708

El ICER de esta intervención se muestra debajo

$$ICER = \frac{(3,213,000 - 3,184,000)}{(2029 - 1864)} = 177.3$$

De acuerdo con este análisis sin ajustar, las suspensiones de sedación parecen ser una estrategia muy costo-efectiva, costando solo $177 más por cada AVAC adicional, en relación a la estrategia control. Reduciendo el valor AVAC del estado Extubado de 1 a 0.6 aumenta significativamente el ICER a $1918 por AVAC ganado, demostrando el gran impacto que la definición de nuestros estados de salud tiene en los resultados del ACE. Del mismo modo, aumentar el costo diario del estado E de $1000 a $1900 (ahora solo levemente más barato que el estado Intubado) lleva a un ICER mucho más costoso de $2041 por AVAC ganado. Algunas intervenciones médicas pueden o no ser financiadas dependiendo de los supuestos del modelo.

24.5 Validación del modelo y análisis de sensibilidad para el análisis de costo efectividad

Un componente importante de cualquier ACE es evaluar si el modelo es apropiado para el fenómeno que se está examinando, lo cual es el propósito de la validación del modelo y de los análisis de sensibilidad. En la sección previa, modelamos la suspensión de la sedación diaria como una cadena de Markov con una matriz de transición de probabilidad y costos. Los desvíos del Modelo pueden ser de 2 tipos. Primero el uso de una cadena de Markov puede ser inapropiado para describir cómo los sujetos cambian de los estados de intubación, extubación y muerte. Se presume que este proceso sigue una cadena de Markov de primer orden. Teniendo un número suficiente de datos clínicos reales podemos testear y ver si esta asunción es razonable. Por ejemplo, dadas las matrices de transición anteriores, podemos calcular los valores vía simulación de MC y compararlos con los valores reportados en los datos reales. Por ejemplo, los autores informan una tasa de mortalidad a los 28 días de 29% y 35% en los grupos intervención y control, respectivamente. A partir de nuestro estudio de simulación, estimamos que estas cantidades eran 27% y 35% lo cual es razonablemente cercano. Es posible realizar un test de bondad de ajuste

(goodness-of-fit) para evaluar mejor si alguna de las diferencias observadas proporcionan alguna evidencia de que el modelo pudiera ser inespecífico. Este proceso también puede repetirse con otras cantidades, por ejemplo, el promedio de días libres de ventilador. Aparte de validar el modelo de Markov usado para simular los estados y las transiciones para el sistema de interés también es importante realizar un análisis de sensibilidad sobre los supuestos y los parámetros usados en la simulación. Realizar estos pasos permite ver cuán sensibles son los resultados a cambios leves de los valores de los parámetros. Puede ser difícil elegir qué parámetros usar en los análisis de sensibilidad, pero son buenas prácticas el encontrar otros parámetros informados en estudios similares (por ej. las matrices de transición de probabilidades). Para la estimación de costos, uno pudiera desear considerar costos reportados en otros países o incorporar parámetros económicos como la inflación. Si usar estos otros escenarios afecta drásticamente las conclusiones del estudio de simulación no necesariamente significa que el estudio ha sido un fracaso, sino que hay límites para generalizar los resultados del estudio de simulación. Si determinados parámetros causan grandes fluctuaciones, puede ser necesario investigar más a fondo para determinar su causa. Aparte de cambiar los parámetros, se puede tratar de alterar el modelo de manera significativa, por ejemplo usando un modelo de Markov de orden superior o un modelo semi Markov en lugar de un supuesto simple de primer orden, pero estos son tópicos avanzados que exceden el objetivo de este capítulo.

Los conceptos teóricos introducidos en la primera sección de este capítulo se aplicaron a un ejemplo concreto proveniente de la literatura médica. Demostramos como los estados clínicos y las probabilidades de transición podrían ser definidas ad hoc, y como la distribución estacionaria de la cadena podría estimarse usando métodos de Monte Carlo. La metodología delineada en este capítulo permitirá al lector expandir los resultados de otros estudios observacionales a ACE, pero existen innumerables aplicaciones de Modelos Markov, en particular en el dominio de los sistemas de soporte de decisión.

24.6 Conclusión

Los modelos de Markov han sido utilizados ampliamente en la literatura médica y ofrecen un marco atractivo para modelar soporte para la toma de decisiones médicas, con posibles aplicaciones de gran alcance en los sistemas de soporte de decisión y análisis de la economía de la salud. Representan modelos matemáticos relativamente simples, fáciles de comprender por aquellos sin experiencia en ciencia de datos o estadística. Se debe prestar cuidadosa atención a la verificación de un supuesto fundamental que es la propiedad de Markov, sin la cual no debería avanzarse en ningún otro análisis.

24.7 Próximos pasos

Se espera que este tutorial proporcione las herramientas básicas para comprender o realizar un ACE y cadenas de Markov para modelar el efecto de las intervenciones médicas. Para mayor información en economía de la salud, dirigimos al lector a referencias externas como el trabajo de Morris y colaboradores [17]. La guía referente al uso de modelos Markov más avanzados como MDPs y HMMs excede los objetivos de este libro, pero existen muchas fuentes disponibles como la excelente obra de Sutton y Barto disponible en forma gratuita on line [4].

Nota: Las imágenes de este capítulo se encuentran disponibles a color en el ebook; disponible en www.hardineros.com

Apéndice: Código

El código utilizado en este caso de estudio se encuentra disponible en el repositorio de GitHub que acompaña este libro https://github.com/MIT-LCP/critical-data-book.

En el sitio web se encuentra disponible mayor información sobre el código.

Se proveen las siguientes funciones:

- `health_forecast.m`: esta función calcula 100 simulaciones Monte-Carlo de un pronóstico de 5 días y muestra los resultados
- `IED_transition.m`: esta función calcula y muestra la proporción de pacientes en cada estado (Intubado, Extubado, o Muerto), siguiendo la matriz de transición en el grupo de intervención.
- `MCMC_solver.m`: esta función calcula 10,000 simulaciones de Monte Carlo, tanto para el grupo control como para el grupo intervención, y calcula la distribución de los días libres de ventilador.

Referencias

1. Basharin GP, Langville AN, Naumov VA (2004) The life and work of A.A. Markov. Linear Algebra Appl 386:3-26.
2. Sonnenberg FA, Beck JR (1993) Markov models in medical decisión making: a practical guide. Med Decis Mak Int J Soc Med Decis Mak 13 (4): 322-338.
3. Schaefer AJ, Bailey MD, Shechter SM, Roberts MS (2005) Modeling medical treatment using Markov decisión processes. In: Brandeau ML, Sainfort F, Pierskalla WP (eds) Operations research and health care. Springer, US, pp 593-612.
4. Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. A Bradford Book, Cambridge, Mass.
5. Kreke JE (2007) Modeling disease management decisions for patients with pneumonia-related sepsis [Online]. Available: http://d-scholarship. pitt.edu/8143/.
6. Liu JS (2004) Monte Carlo strategies in scientific computing. Springer, New York.
7. Zucchini W, MacDonald IL (2009) Hidden Markov models for time series: an introduction using R. Chapman and Hall/CRC, Boca Raton (2Rev Ed edition).
8. Alagoz O, Maillart LM, Schaefer AJ, Roberts MS (2004) The optimal timing of living-donor liver transplantation. Manag Sci 50 (10): 1420-1430.
9. Shechter SM, Bailey MD, Schaefer AJ, Roberts MS (2008) The optimal time to initiate HIV therapy under ordered health states. Oper Res 56 (1): 20-33.

10. Maillart LM, Ivy JS, Ransom S, Diehl K (2008) Assessing dynamic breast cancer screening policies. Oper Res 56 (6): 1411-1427.
11. Daniel PMG, Faissol M (2007) Timing of testing and treatment of hepatitis C and other diseases. Inf J Comput Inf.
12. Denton BT, Kurt M, Shah ND, Bryant SC, Smith SA (2009) Optimizing the start time of statin therapy for patients with diabetes. Med Decis Mak Int J Soc Med Decis Mak 29 (3): 351– 367.
13. Raffa JD, Dubin JA (2015) Multivariate longitudinal data analysis with mixed effects hidden Markov models. Biometrics 71 (3): 821-831.
14. Scott SL, James GM, Sugar CA (2005) Hidden Markov models for longitudinal comparisons. J Am Stat Assoc 100:359-369.
15. Srikanth P (2015) Using Markov chains to predict the natural progressión of diabetic retinopathy. Int J Ophthalmol 8 (1): 132-137.
16. Alterovitz R, Branicky M, Goldberg K (2008) Motion planning under uncertainty for image-guided medical needle steering. Int J Robot Res 27 (11-12): 1361-1374.
17. Morris S, Devlin N, Parkin D, Spencer A (2012) Economic analysis in healthcare, 2nd edn. Wiley, Chichester.
18. Nord E, Daniels N, Kamlet M (2009) QALYs: some challenges. Value Health 12 (Supplement1): S10-S15.
19. Torrance GW (1986) Measurement of health state utilities for economic appraisal. J Health Econ 5 (1): 1-30.
20. Drummond M, Sculpher M (2005) Common methodological flaws in economic evaluations. Med Care 43 (7 Suppl): 5-14.
21. Girard TD, Kress JP, Fuchs BD, Thomason JWW, Schweickert WD, Pun BT, Taichman DB, Dunn JG, Pohlman AS, Kinniry PA, Jackson JC, Canonico AE, Light RW, Shintani AK, Thompson JL, Gordon SM, Hall JB, Dittus RS, Bernard GR, Ely EW (2008) Efficacy and safety of a paired sedation and ventilator weaning protocol for mechanically ventilated patients in intensive care (awakening and breathing controlled trial): a randomised controlled trial. Lancet Lond Engl 371 (9607): 126-134.
22. Roberts DJ, Haroon B, Hall RI (2012) Sedation for critically ill or injured adults in the intensive care unit: a shifting paradigm. Drugs 72 (14): 1881-1916.

Capítulo 25

LA PRESIÓN SANGUÍNEA Y EL RIESGO DE LESIÓN RENAL AGUDA EN LA UCI: DISEÑO CASO-CONTROL VERSUS DISEÑO DE CASOS CRUZADOS

Li-wei H. Lehman, Mengling Feng, Yijun Yang y Roger G. Mark

Objetivos de Aprendizaje

Introducir dos enfoques diferentes, un diseño de casos y controles y otro de casos cruzados, para estudiar el efecto de la exposición transitoria de la hipotensión en el riesgo de desarrollo de insuficiencia renal aguda (IRA) en los pacientes de la unidad de cuidados intensivos (UCI).

25.1 Introducción

La insuficiencia renal aguda (IRA) se refiere a una rápida disminución de la función renal, que ocurre en un período de días. La presencia de IRA puede detectarse utilizando definiciones bien establecidas basadas en el aumento de la creatinina plasmática o en la reducción de la diuresis [1]. La insuficiencia renal aguda fue reportada en el 36% de los pacientes ingresados en la unidad de cuidados intensivos UCI [2,3]. Un estudio anterior mostró que pacientes hospitalizados con aumentos, incluso muy pequeños (0.3-0.4 mg/dL), de la creatinina plasmática tiene un riesgo 70% mayor de morir que pacientes sin aumentos de creatinina [4]. Aunque la relación entre la baja presión sanguínea y la función renal está bien documentada en un escenario experimental basado en datos de animales [5], la asociación entre hipotensión e insuficiencia renal aguda en un entorno de cuidados intensivos no se comprende completamente.

Este capítulo describe dos enfoques diferentes para estudiar la presión sanguínea y el riesgo del desarrollo de IRA en pacientes de UCI utilizando la base de datos MIMIC II [6]. En nuestro primer estudio, adoptamos un enfoque tradicional de caso-control y examinamos la asociación entre hipotensión e IRA comparando medidas de presión arterial de pacientes que tuvieron IRA (caso) con pacientes sin IRA (control) [7,8]. Las mediciones de la presión sanguínea inmediatamente antes del inicio de la IRA se compararon con las mediciones de la presión sanguínea de los controles muestreados en una ventana de tiempo similar.

En el segundo estudio, adoptamos un diseño de casos-cruzados en el que cada paciente sirve como su propio control. Las mediciones de presión sanguínea inmediatamente antes del inicio de la IRA se compararon con las mediciones de presión sanguínea del mismo paciente tomadas en una ventana de tiempo anterior mientras las funciones renales de ese paciente estaban todavía estables. En el resto del capítulo, destacamos las principales diferencias y la lógica de diseño de estos dos enfoques. Aplicamos estas técnicas de análisis para estudiar la relación entre la hipotensión y el desarrollo de IRA utilizando la base de datos MIMIC II, y presentamos nuestros hallazgos preliminares.

25.2 Métodos

25.2.1 Preprocesamiento de datos

Para el análisis se utilizaron muestras de presión arterial media (PAM) verificadas por enfermeras, registradas cada hora. Se incluyeron en el estudio las mediciones de la presión sanguínea tanto de catéter arterial invasivo como de los métodos oscilométricos automatizados y no invasivos. Nuestra elección de la PAM (en lugar de la presión arterial sistólica) para las mediciones de presión sanguínea fue motivada por trabajos anteriores [8] que demostraron que la PAM proporcionaba lecturas más consistentes a lo largo de diferentes modalidades de medición en la UCI. Las mediciones de presión sanguínea fueron filtradas para eliminar los valores por fuera de los límites razonables fisiológicamente (PAM entre 20 y 200 mmHg).

25.2.2 Un Estudio Caso-Control

En el enfoque de caso-control, examinamos el efecto de la exposición transitoria a hipotensión (definida como una caída de la presión sanguínea por debajo de umbrales específicos) y el riesgo del desarrollo de IRA comparando mediciones de presión sanguínea de los pacientes que desarrollaron IRA (caso) con pacientes que nunca desarrollaron IRA en la UCI (control). La IRA se definió como un incremento agudo en la creatinina sérica ≥ 0.3 mg/dl, o un incremento de ≥ 50% en la creatinina sérica en 48 hs, basado en la definición de AKIN (Acute Kidney Injury Network) [1]. Las mediciones de presión sanguínea tomadas antes del inicio de la IRA (a partir de una ventana de hasta 48 hs) se compararon con las mediciones de

presión sanguínea de los controles con una ventana de tiempo anterior a la última medición de la creatinina.

Los pacientes fueron seleccionados entre los adultos admitidos en UCI en la base de datos MIMIC II [8]. Examinamos las estadías de adultos en la UCI (pacientes ≥15 años de edad) con al menos 2 valores de creatinina sérica. Se excluyeron los pacientes con menos de 2 valores de creatinina en su estadía en la UCI o con evidencia de enfermedad renal en etapa terminal (ESRD).

Entre los 16728 pacientes adultos en UCI que tenían por lo menos 2 mediciones de creatinina sin evidencia de enfermedad renal en etapa terminal, 5207 (31%) desarrollaron IRA. Los 11521 casos restantes fueron identificados como controles. El promedio de tiempo de inicio de la IRA fue de 2,34 días después de la admisión en la UCI. Para los controles, el tiempo de la última muestra de creatinina fue, en promedio, de 2,76 días luego del ingreso a la UCI. La Figura 25.1 muestra el promedio de la población y el error estándar de la mediana de la PAM hasta 3 días antes del inicio de la IRA para la cohorte de IRA, o antes del último tiempo de medición de creatinina para los controles.

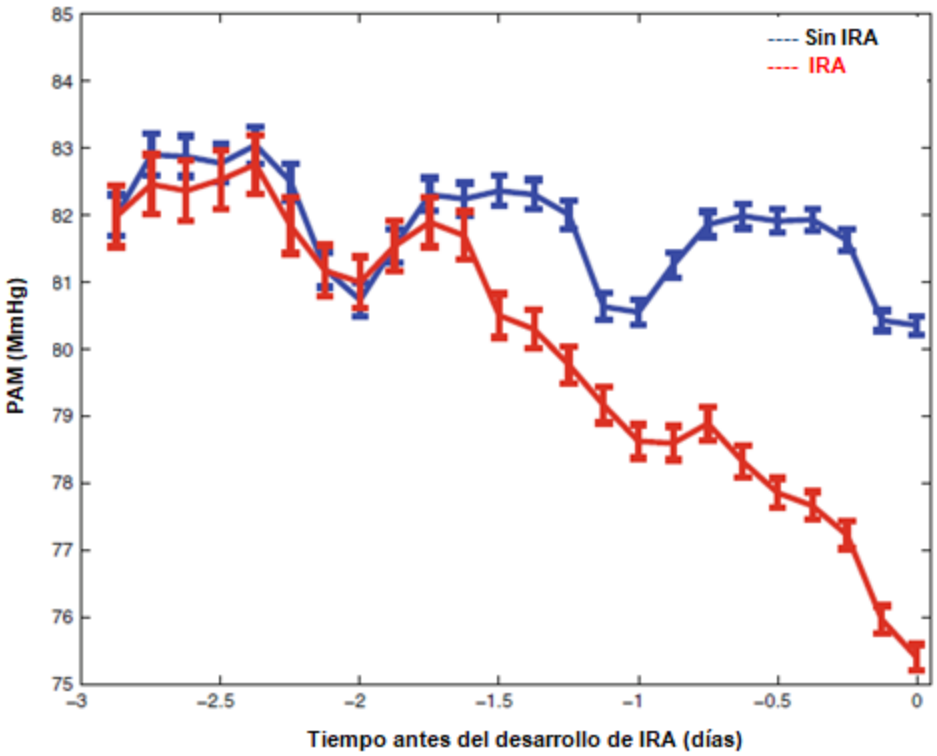

Fig 25.1 La media poblacional (y su error estándar) de la mediana de TAM hasta 3 días antes del desarrollo de IRA para la cohorte con IRA o antes de la última medición de creatinina para los controles. La TAM de la cohorte con IRA difirió de la de los controles desde el día 2 antes de la aparición de IRA y ambas cohortes presentaron importantes variaciones diurnas.

Nótese que la presión arterial media de la cohorte que desarrolló IRA difirió de la de los controles antes del inicio de la IRA.

Estudiamos el riesgo de IRA en pacientes de UCI en función tanto de la severidad como de la duración de la hipotensión. Se examinaron las características de la presión sanguínea obtenida de la ventana objetivo de 48 hs como principales predictores de la IRA, incluyendo la PAM mínima y el número máximo de horas que la PAM fue continuamente inferior a varios umbrales diferentes (de 80 a 45 mmHg). La duración de la hipotensión por debajo de un umbral específico se calculó en base a muestras de presión sanguínea interpoladas linealmente. Se consideró que los episodios hipotensivos comenzaban y terminaban cuando los valores de presión sanguínea interpolados interceptaban el umbral objetivo. Los episodios hipotensivos con menos de una hora de diferencia se fusionaron para formar un episodio continuo.

Se realizaron regresiones logísticas univariada y multivariable para encontrar correlaciones entre hipotensión e IRA. La edad, el SAPS-I, la creatinina de admisión y la presencia (basada en CIE-9) de falla renal crónica (585.9), hipertensión (401.9), diabetes (250.00), aterosclerosis coronaria (414.01), insuficiencia cardíaca congestiva (428.0) y shock séptico (785.52) o sepsis (038) fueron agregados como potenciales factores de confusión [9].

Nuestros resultados indican que las probabilidades de IRA estaban relacionadas con la gravedad de la hipotensión con un odds ratio (OR) de 1.03, un intervalo de confianza 95% (IC 95%) de 1.02-1.04 ($p<0.0001$) por cada 1 mmHg de disminución de la PAM$\leq$ 80 mmHg. El análisis multivariable de la duración de la hipotensión involucró a 3203 pacientes que tenían puntuaciones SAPS-I y con al menos 45 h de muestras de presión sanguínea en la ventana objetivo de 48 hs. Nuestros resultados indican que la duración de tiempo que la PAM de los pacientes era continuamente menor o igual a 70, 65, 60, 55 y 50 mmHg eran factores de riesgo significantes en el desarrollo de IRA. Además, a medida que el grado de hipotensión empeoraba, el aumento del riesgo de IRA por cada hora adicional de

hipotensión continua aumentaba, por cada descenso de 10 mmHg de la PAM por debajo de 80 mmHg. Por cada hora adicional que la PAM era menor que 70,60,50 mmHg, la probabilidad de IRA aumentaba un 2% (OR 1.02, IC 95% 1.00-1.03, p=0.0034), 5% (OR 1.05, IC 95% 1.02-1.08, p=0.0028), y 22% (OR 1.22, IC 95% 1.04-1.43, p=0.0122) respectivamente. A medida que el grado de hipotensión empeoraba, la probabilidad de IRA por cada hora adicional de hipotensión continua aumentaba más del doble, por cada descenso de 10 mmHg en la PAM debajo de 80. Nuestros resultados también sugieren que la severidad de la hipotensión acortaba significativamente el tiempo para el inicio de la IRA.

25.2.3 Diseño Casos-Cruzados

En el segundo estudio, adoptamos un diseño de cohorte de casos-cruzados para evaluar el efecto de la exposición transitoria a hipotensión y el riesgo de IRA. El diseño caso cruzado fue ideado para evaluar la relación entre variables de exposición transitorias y resultados agudos en situaciones donde la serie de control de un estudio de casos-controles es difícil de lograr. En el diseño caso-cruzado, los sujetos sirven como sus propios controles definidos por períodos de tiempo anteriores en el mismo sujeto. Dada una exposición transitoria con una prevalencia estable a lo largo del tiempo, el diseño de caso cruzado utiliza la diferencia en la tasa de exposición justo antes del evento (caso) con la de otros puntos del historial del sujeto (controles) para estimar un odds ratio del resultado asociado a la exposición. El diseño de caso cruzado fue propuesto en primer lugar por Maclure y col para estudiar los efectos de cambios transitorios en el riesgo de eventos agudos [10]. Una ventaja del diseño de caso cruzado es que evita el sesgo en la selección del control y elimina factores confundidores entre pacientes [10, 11]. En este diseño de estudio, la definición de IRA se basa en la diuresis por hora (en lugar de mediciones diarias de creatinina) con el fin de determinar un momento más preciso de la aparición aguda (oliguria).

Se incluyeron en el estudio pacientes adultos con función renal normal (ej., la diuresis se mantiene en 0.5 ml/kg/h o más) durante las primeras 12 hs en la UCI, quienes subsecuentemente desarrollaron IRA/oliguria (la diuresis permanece por debajo de 0.5 ml/kg/h por al menos 6 hs). Los

mismos pacientes, antes de desarrollar la IRA/oliguria, fueron utilizados como controles. El comienzo de la IRA/oliguria se definió como el inicio del período de 6 hs en el que la diuresis se mantuvo por debajo de 0.5 ml/kg/h.

La PAM mínima en el período de 3 hs antes del inicio de la IRA se utilizó como exposición de los casos. El mínimo de PAM del período de control de 3 hs durante las primeras 12 hs en la UCI, cuando la función renal del mismo paciente era todavía normal, fue utilizado como la exposición de los controles. Mientras que las mediciones de presión sanguínea durante las primeras 6 hs en pacientes en la UCI pueden ser escasas, elegimos que el período de control fuera la 7ª-9ª hora desde el comienzo de la estadía del paciente en la UCI. Las mediciones de presión sanguínea fueron filtradas para eliminar valores atípicos (*outliers)* como antes.

Los diseños de casos cruzados son típicamente analizados utilizando la regresión logística condicional, ya que da cuenta de la naturaleza pareada de los datos. Es análogo a un estudio de casos y controles pareado, donde se compara un "caso" de persona-momento con una serie de "controles" persona-momentos de diferentes sujetos, mientras que en el diseño de caso cruzado, el "control" de persona-momentos es del mismo sujeto. Aplicamos este último enfoque para analizar los datos de los estudios de casos cruzados. Además, los factores confundidores que varían con el tiempo (ventilación mecánica, vasopresores, temperatura, frecuencia cardíaca, recuento de glóbulos blancos, SatO2) fueron incluidos en el modelo multivariable de regresión logística condicional.

La cohorte total incluyó a 911 adultos en UCI (29.86% UCI médica, 21.73% UCI quirúrgica, 22.94% UCI cardiovascular, 25.47% UCI de recuperación de cardiocirugía) de la base de datos MIMIC II. El tiempo promedio del inicio de la IRA/oliguria fue de 45 hs. La mediana de la población del mínimo de las mediciones de PAM durante los períodos del control y caso fue de 73 mmHg con un rango intercuartílico de [65, 83] mmHg, y 70 [62,79] mmHg respectivamente. Una prueba T Student para muestras pareadas indica que la mínima PAM durante el período del caso fue más baja que durante el período de control, en forma estadísticamente significativa (valor de p =0.0001). Nuestros resultados indican que las probabilidades de IRA estaban relacionadas con la gravedad de la hipotensión con un odds ratio (OR) de 1.035, un de intervalo de confianza 95% (IC 95%) de 1.024-1.045 ($p< 0.0001$) por cada 1 mmHg de disminución

de la PAM mínima en la regresión logística condicional multivariable después de ajustar por la temperatura, la frecuencia cardíaca, la SatO2, el recuento de glóbulos blancos y el uso de ventilación mecánica y vasopresores. Además, realizamos un análisis similar para comprender si el aumento del riesgo de desarrollar IRA se asocia con el empeoramiento de la hipotensión, considerando la PAM mínima como una variable binaria con un corte de 70,65,60,55 y 50 mmHg. El odds ratio ajustado y el IC 95% para la mínima PAM< 70, PAM< 65, PAM< 60, PAM< 55, y PAM< 50 (vs. cuando la PAM es mayor o igual a sus respectivos límites) fue de 1.854 (1.44-2.38), 1.945 (1.502-2.519), 2.096 (1.532-2.896), 2.002 (1.307-3.065), y 2.107 (1.115-3.982), respectivamente. Estos hallazgos son consistentes con los resultados descritos en la sección previa utilizando un diseño de estudio de caso-control.

25.3 Discusión

El diseño de caso cruzado es una alternativa eficiente al enfoque de caso-control en el estudio de la asociación entre hipotensión e IRA.

El diseño de caso cruzado, basado exclusivamente en la serie de casos, realiza comparaciones entre las mediciones de presión sanguínea de los períodos de caso y de control en un mismo sujeto para estimar la razón de tasas del resultado IRA asociado con hipotensión. Este diseño inherentemente elimina en la razón de tasas estimada, los efectos de sesgo de los factores de confusión no medidos e invariables con el tiempo.

Muchos factores (incluyendo enfermedad renal crónica, hipertensión, diabetes) potencialmente podrían contribuir al desarrollo de IRA en un entorno de UCI. En un diseño de caso control tradicional, estos factores de confusión entre pacientes que no varían con el tiempo (así como los factores de confusión que varían con el tiempo) tendrían que ser incluidos para ajustar el riesgo de base de desarrollar IRA. En algunos casos, puede ser difícil determinar en una base de datos de ICU retrospectiva, estas variables de confusión. En un diseño de caso cruzado, cada presión sanguínea del paciente mientras la función renal es normal es comparada con la presión sanguínea del mismo paciente inmediatamente antes del inicio de la IRA, de modo que las características del paciente y los factores de confusión que varían con el tiempo son eliminados en el análisis. El

diseño de caso cruzado puede ser un enfoque más eficiente para investigar el efecto transitorio de la exposición (ej., baja presión sanguínea) en el riesgo de un resultado agudo (ej., desarrollo de una IRA), cuando la heterogeneidad en el riesgo de base puede ser difícil de explicar en un diseño convencional de caso-control.

Reconocemos las siguientes limitaciones en el estudio actual. Primero, este fue un estudio retrospectivo, y como tal, la incidencia de hipotensión anterior a la IRA no prueba un mecanismo causal. Segundo, no hemos tenido en cuenta en nuestro análisis multivariable la presencia de fluidos y varias intervenciones (ej., contrastes, AINEs, aminoglucósidos, IECA, etc.) que pueden perjudicar la función renal. Como parte de trabajos futuros, se podría incluir en el modelo otros factores de confusión que varían con el tiempo (tales como, el uso de Lasix (furosemida) dentro de las 6 hs, fluidos EV, creatinina, tiempo de inicio de la IRA).

25.4 Conclusiones

Hemos presentado dos enfoques diferentes, el caso control y el diseño de caso cruzado, para estudiar el efecto de la exposición transitoria a hipotensión en el riesgo del desarrollo de IRA en pacientes en UCI. Los resultados del análisis multivariable en ambos estudios indican que la hipotensión es un factor de riesgo estadísticamente significativo en el desarrollo de IRA en la UCI. El estudio sirve como un ejemplo para ilustrar la utilidad de los diseños de casos cruzados para estudiar la asociación entre un factor de riesgo y subsecuente desarrollo de enfermedad en un análisis clínico retrospectivo basado en HCE.

reglamento, los usuarios deberán obtener permiso del titular de la licencia para duplicar, adaptar o reproducir el material.

Nota: Las imágenes de este capítulo se encuentran disponibles a color en el ebook; disponible en www.hardineros.com

Apéndice: Código

El código usado en este caso de estudio se encuentra disponible en el repositorio de GitHub que acompaña este libro: https://github.com/MIT-LCP/critical-data-book. En este sitio web se encuentra disponible mayor información sobre el código.

Referencias

1. Mehta RL, Kellum JA, Shah SV, Molitoris BA, Ronco C, Warnock DG, Levin A (2007) Acute kidney injury network (AKIN): report of an initiative to improve outcomes in acute kidney injury. Crit Care 11: R31.
2. Bagshaw S, George C, Dinu I, Bellomo R (2008) A multi-center evaluation of the RIFLE criteria for early acute kidney injury in critically ill patients. Nephrol Dial Transplant 23:1203-1210.
3. Ostermann M, Chang R (2007) Acute kidney injury in the intensive care unit according to rifle. Crit Care Med 35:1837-1843.
4. Chertow G, Burdick E, Honour M, Bonventre J, Bates D (2005) Acute kidney injury, mortality, length of stay, and costs in hospitalized patients. J Am Soc Nephrol 16:3365-3370.
5. Kirchheim HR, Ehmke H, Hackenthal E, Löwe W, Persson P (1987) Autoregulation of renal blood flow, glomerular filtration rate and renin release in conscious dogs. Pflugers Archiv Eur J Physiol 410:441-449.
6. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman LH, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG (2011) Multiparameter intelligent monitoring in intensive care (MIMIC II): a public-access intensive care unit database. Crit Care Med (5): 952-960.
7. Lehman LH, Saeed M, Moody G, Mark R (2010) Hypotensión as a risk factor for acute kidney injury in ICU patients. In: Computing in cardiology 2010. IEEE Computer Society Press, Belfast, pp 1095-1098.
8. Lehman LH, Saeed M, Talmor D, Mark RG, Malhotra A (2013) Methods of blood pressure measurement in the ICU. Crit Care Med 41 (1): 3-40.
9. Abuelo G (2007) Normotensive ischemic acute renal failure. N Engl J Med 357:797-805.

10. Maclure M (1991) The case-crossover design: a method for studying transient effects on the risk of acute events. Am J Epidemiol 133:144-153.
11. Maclure M, Mittleman M (2000) Should we use a case-crossover design. Annu Rev Public Health 21:193-221.

Capítulo 26

ANÁLISIS DE SEÑALES PARA ESTIMAR LA FRECUENCIA RESPIRATORIA

Peter H. Charlton, Mauricio Villarroel
y Francisco Salguiero

Objetivos de Aprendizaje

Usar la base de datos MIMIC II para comparar el rendimiento de múltiples algoritmos para la estimación de la frecuencia respiratoria (FR) a partir de las formas de onda de señales fisiológicas.

1. Extraer de la base de datos MIMIC II las formas de onda del electrocardiograma (ECG), fotopletismógrafo (FPG) y neumografía por impedancia torácica (NI)
2. Identificar períodos con datos con formas de onda de baja calidad
3. Identificar los latidos cardíacos en el ECG y señales de FPG
4. Estimar la FR a partir de las señales
5. Mejorar la precisión de la estimación de la FR utilizando evaluación de la calidad y fusión de datos
6. Evaluar el rendimiento de los algoritmos de FR

26.1 Introducción

La frecuencia respiratoria (FR) es un importante parámetro fisiológico que proporciona valiosa información diagnóstica y pronóstica. Se ha descubierto que es predictivo de infecciones del tracto respiratorio inferior [1], indicativo de la gravedad de una neumonía [2] y que se asocia con la mortalidad en pacientes pediátricos en la unidad de cuidados intensivos (UCI) [3]. La frecuencia respiratoria se mide en respiraciones por minuto (rpm). La práctica rutinaria habitual para obtener mediciones de la FR fuera de Cuidados Intensivos implica contar en forma manual los movimientos torácicos [4]. Esta práctica requiere mucho tiempo, es inexacta, y se lleva a cabo de forma deficiente [6, 8]. Por lo tanto, existe la necesidad urgente de desarrollar un método preciso y automatizado para la medición de la FR en pacientes ambulatorios. Además, un método automatizado de medición de la FR facilitaría: (i) la monitorización objetiva del asma en el hogar dirigida por el paciente; (ii) detección de apnea del sueño obstructiva; y (iii)

detección de períodos de respiración desregulada durante el sueño, vistos ocasionalmente en la insuficiencia cardíaca congestiva avanzada.

Una solución posible es estimar la FR a partir de una señal, convenientemente no invasiva, modulada por la respiración, fácilmente medible y de preferencia medible en forma rutinaria. Dos de estas señales son el electrocardiograma (ECG) y la fotopletismografía (FPG). Ambas señales exhiben una variación de la línea de base (VB), modulación de la amplitud (MA) y modulación de la frecuencia (MF) debido a la respiración, como se muestra en la Fig. 26.1 (vea [9, 10] para más detalles). Además, las dos señales pueden obtenerse continuamente de pacientes ambulatorios que usan novedosos sensores portátiles. Por ejemplo, el sistema Sensium Vitals (Sensium Healthcare) proporciona un monitoreo continuo de ECG y FPG usando un parche liviano con una batería de hasta cinco días de duración. The VisiMobile (Sotera Wireless) proporciona un monitoreo continuo de ECG y FPG utilizando un monitor de muñeca con electrodos de ECG adicionales. Así mismo, está siendo desarrollada tecnología basada en video, sin contacto, para el monitoreo continuo de FPG sin necesidad de ningún equipamiento conectado al paciente [11].

Se han desarrollado muchos algoritmos para estimar la FR a partir del ECG y la FPG [10, 12], pero aún no se han adoptado ampliamente en la práctica clínica. En este caso de estudio demostramos la aplicación de técnicas de ejemplo para el ECG y el FPG. El desempeño de estas técnicas fue evaluado en un set de datos de ejemplo. El caso de estudio se acompaña del correspondiente código MATLAB, aportando al lector herramientas para desarrollar y probar sus propios algoritmos de estimación de FR a partir de trazados fisiológicos.

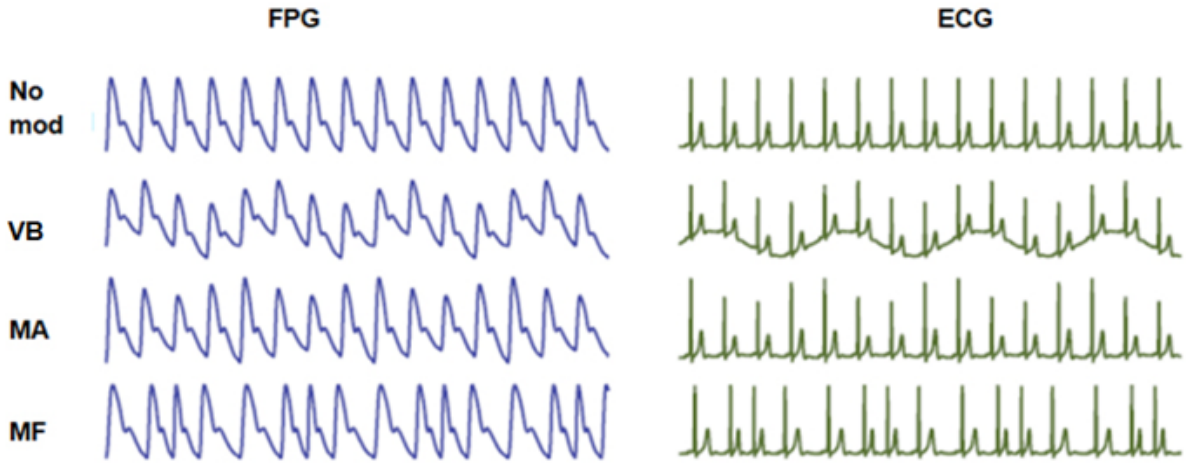

Fig 26.1 Modulaciones respiratorias idealizadas de la FPG (izquierda) y ECG (derecha). Durante 3 ciclos respiratorios, desde arriba no modulación, variación de la línea de base, modulación de amplitud (MA) y modulación de frecuencia (MF). Adaptado de [18,27,30]

26.2 Set de Datos del Estudio

La base de datos MIMIC II (Versión 3) de PhysioNet fue elegida para este estudio porque contiene simultáneamente trazados de ECG, FPG y neumografía por impedancia torácica (NI) [13,14]. Las señales de NI, usualmente medidas sólo en cuidados intensivos, pueden utilizarse para estimar las FR de referencia, ya que se pueden identificar las respiraciones individuales cuando la impedancia torácica aumenta durante la inhalación y disminuye durante la exhalación. Se utilizó `MIMICII_data_importer.m` en conjunción con el sistema de acceso libre *WFDB Toolbox*[3] para descargar los datos. Se descargaron cien registros de estadía en la Unidad de Cuidados Intensivos (UCI), cada uno de los cuales contiene datos de una estancia distinta en la UCI.

Tabla 26.1 Criterios para determinar si cada uno de los 100 registros descargados de la base de datos , estaban incluidos en el análisis

Criterio	Porcentaje de registros que cumplían con el criterio
Contiene todos los trazados de ondas requeridas (ECG, FPG e impedanciometría torácica)	76
Contiene todos los numéricos requeridos (frecuencia cardíaca FC, frecuencia de pulso, FP y frecuencia respiratoria FR)	64
Los trazados de ondas y numéricos requeridos tienen al menos 10 minutos	51

En el análisis se incluyeron los registros que cumplían los criterios de la Tabla 26.1. Los trazados de onda y números requeridos se extrajeron del 51% de los registros que cumplieron con esos criterios. Cada canal de datos se almacenó en dos vectores de valores y los correspondientes registros de tiempo. De este modo se aseguró que se preservara en el análisis cualquier faltante en los datos debida a cambios en el monitoreo de los pacientes o a fallos en la adquisición de datos.

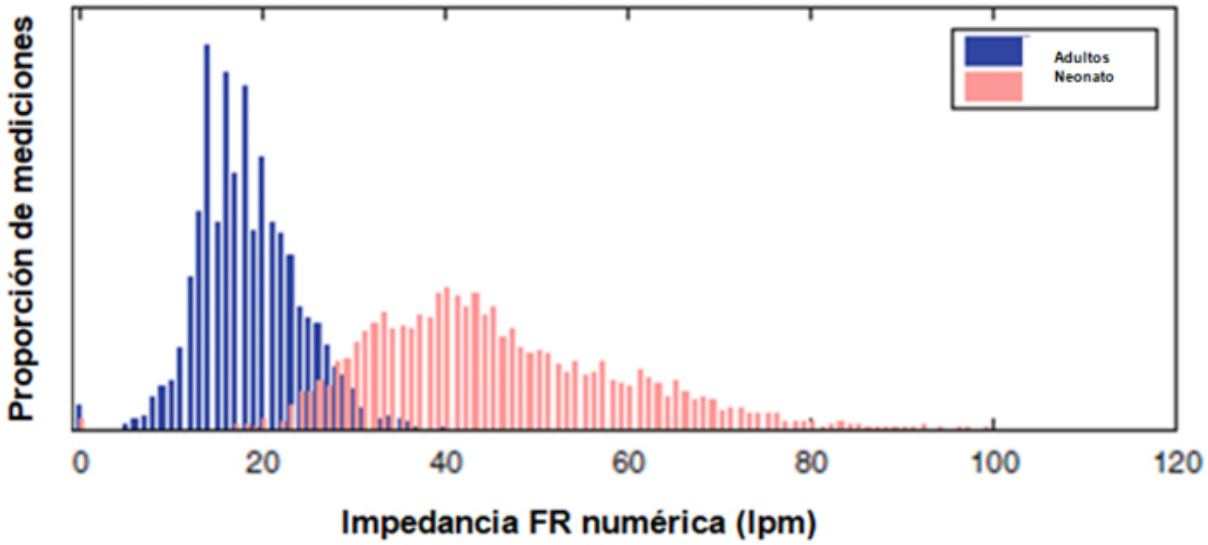

Fig 26.2 Mediciones de frecuencia respiratoria de referencia (FR), adquiridas usando impedancia torácica de adultos y neonatos. La disparidad de distribución de mediciones de FR adquirida de adultos (azul) y neonatos (roja) llevó a análisis de subgrupo de estas dos poblaciones de pacientes

La inspección del set de datos reveló una diferencia sustancial en las distribuciones de las mediciones de FR por NI adquiridas de pacientes adultos y neonatos, como se ve en la Fig. 26.2. Esto está en concordancia con los hallazgos anteriores [15], donde se informó que la FR de los niños disminuye desde una mediana de 43 rpm en los menores de 3 meses a una

mediana de 16 rpm en aquellos entre 15-18 años. Por lo tanto, decidimos restringir el análisis sólo a pacientes adultos.

26.3 PreProcesamiento

Los trazados de ondas obtenidos contenían períodos de alta y baja (confiable y poco confiable) calidad, como muestra la Fig. 26.3. Esto coincide con la literatura, donde está bien reportado que se puede esperar que las señales fisiológicas contengan periodos de artefacto en el escenario de Cuidados Intensivos [16]. Cada segmento de 10 s de los datos del ECG y la FPG fue clasificado como de alta o baja calidad utilizando el índice de calidad de la señal (ICS) que se informa en [17]. Este ICS determina la calidad de la señal en dos pasos. En primer lugar, se detectan los latidos del corazón para cuantificar la frecuencia cardíaca detectada. Cualquiera de los segmentos que contenga frecuencias cardíacas no plausibles fisiológicamente será considerado de baja calidad. En segundo lugar, se utiliza la coincidencia de patrones para cuantificar la correlación entre la morfología de un latido promedio y la de cada latido individual. Si el coeficiente de correlación promedio a través de un segmento es inferior a un umbral empírico, entonces se considera que la calidad de la señal es baja (como se muestra en la Fig. 26.4). Los segmentos de baja calidad fueron eliminados del análisis.

Las mediciones de FR proporcionadas por el monitor clínico no se utilizaron como referencias para evaluar la exactitud de los algoritmos de FR dado que son susceptibles de sufrir inexactitudes durante los períodos de artefacto de señal. En su lugar, se extrajeron FR de referencia de la señal de NI, excluyendo del análisis los períodos en que las FR de referencia eran poco fiables. Para hacer eso, la señal fue segmentada en ventanas no superpuestas de 32 s. Se utilizaron dos métodos independientes para estimar la FR para cada ventana de acuerdo con la metodología presentada en [18]. En primer lugar, se utilizó un análisis de Fourier para calcular la densidad espectral de potencia de la señal, como se describe en [19].

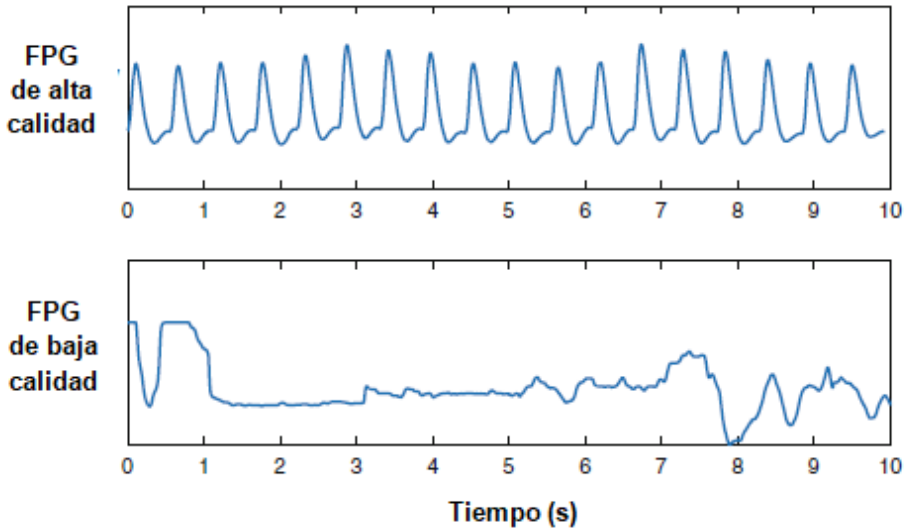

Fig 26.3 Períodos de señales de FPG de alta y baja calidad

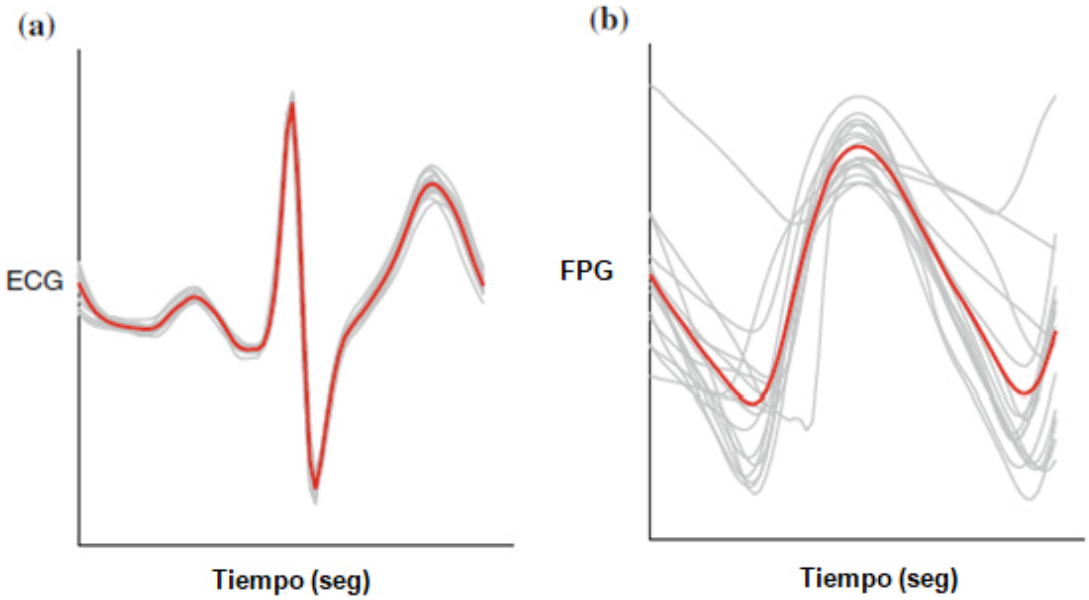

Fig 26.4 Uso de un modelo de coincidencia de índice de calidad de señales para determinar si un segmento de señales es de alta o baja calidad. a) en los latidos del ECG (gris) todos tienen morfología similar al latido promedio (rojo) y el segmento de ECG se califica como de alta calidad, b) los latidos de la FPG tienen una morfología muy variable, lo que indica baja señal.

Se obtuvo una primera FR estimada como la frecuencia correspondiente a la potencia máxima dentro del rango de frecuencias respiratorias plausibles (4-60 rpm). En segundo lugar, se utilizó el método de "count-orig" presentado en [20] para detectar las respiraciones individuales. El "count-

orig" implica la normalización de la señal, la identificación de los pares de máximos que superan un valor de umbral y la identificación de las respiraciones fiables como períodos de señal entre los pares de máximos que contienen sólo un mínimo bajo cero. Por último, si la diferencia entre las dos FR estimadas era <2 rpm, entonces la FR de referencia se calculó como el promedio entre las dos estimaciones. De lo contrario, la ventana fue excluida.

26.4 Métodos

Se han propuesto una gran cantidad de algoritmos para la estimación de la FR a partir del ECG o de la FPG. En este caso de estudio implementamos algoritmos ejemplo (utilizando `RRest.m`) que estiman la FR explotando una de las tres modulaciones respiratorias fundamentales, modeladas según el enfoque descrito en [19]. Los algoritmos de FR generalmente consisten en dos componentes obligatorios y dos componentes opcionales. Los componentes obligatorios son:

- extracción de una señal respiratoria (una serie temporal dominada por la modulación de la respiración) a partir de una señal en bruto y,
- estimación de la FR a partir de la señal respiratoria.

Se pueden utilizar 2 componentes opcionales, la evaluación de la calidad y la fusión para mejorar la precisión de la FR estimada.

La extracción de la señal respiratoria generalmente se realiza usando una técnica basada en las características, que extrae una serie temporal de mediciones de las características latido a latido. La figura 26.5 muestra los pasos involucrados.

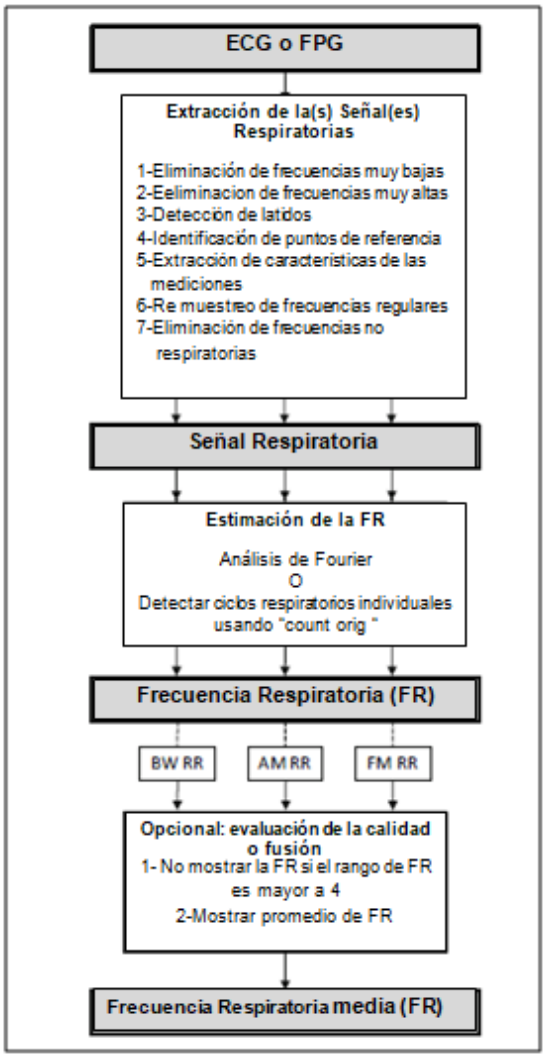

Fig 26. Los pasos en un algoritmo de Frecuencia respiratoria (FR). La extracción de señales y la estimación de la FR es obligatoria. El tercer paso que consiste en evaluación de la calidad y fusión es optativo.

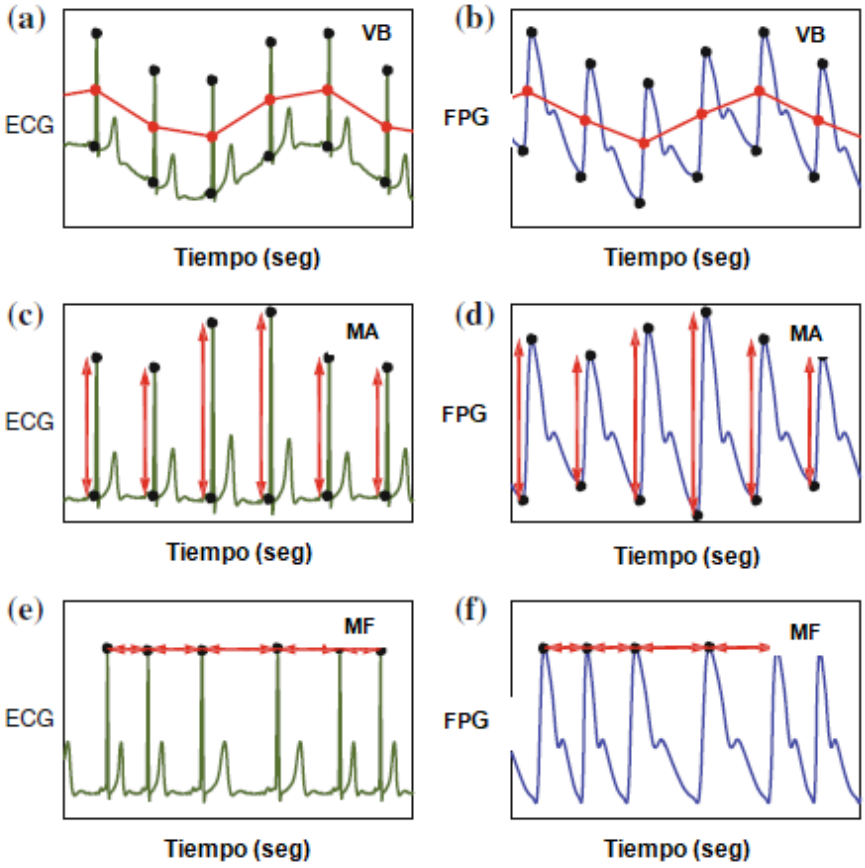

Fig 26.4 Medición de las características de los puntos de referencia de las señales de ECG y de la FPG; a y b mediciones de variación de línea de base (VB), promedio de las amplitudes de pico y valle de un latido, c y d modulación de amplitud (MA), la diferencia entre las amplitudes del pico y valle de cada latido; y f modulación de frecuencia (MF), el intervalo de tiempo entre picos consecutivos.

Los dos primeros pasos, la eliminación de frecuencias menores de 4 lpm y frecuencias muy altas (> 100 Hz y <35 Hz para el ECG y el FPG respectivamente) generalmente no son necesarios cuando se analizan datos de la HCE dado que generalmente son realizados por los monitores de los pacientes antes de la salida de la señal.

La detección de latidos en el ECG se realizó usando un detector de QRS basado en el algoritmo de Pan, Hamilton y Tompkins [21,22] y en el FPG usando el algoritmo de Segmentación y Fusión Incremental [23]. Se identificaron puntos de referencia para cada latido, como ondas R o picos de pulso y onda Q y valles de pulso. Luego se extrajeron de estos puntos de referencia tres mediciones características, como se muestra en la Figura 26.6

Las tres series temporales latido a latido de las características de las mediciones son muestreadas en forma irregular dado que hay una sola medición por latido cardiaco. Dado que el análisis del dominio frecuencia requiere que las señales sean muestreadas en forma regular, estas señales fueron remuestreadas a una frecuencia regular de 5 Hz usando interpolación lineal. Finalmente, se eliminaron las frecuencias espurias no respiratorias introducidas en el proceso de extracción, utilizando filtros paso banda en el rango de frecuencias respiratorias plausibles (4-60 rpm). Las frecuencias altas espurias provienen debido a interpolación lineal y las frecuencias bajas pueden ser causadas por cambios fisiológicos.

Se realizó la estimación de la FR del ECG y de la FPG en los dominios de frecuencia y de tiempo usando el análisis de Fourier y se usaron técnicas de detección de ciclos respiratorios para estimar las FR de referencia. En forma opcional se realizó un paso adicional de evaluación de calidad y fusión, el método “Smart Fusion” [19], en un intento de aumentar la precisión de las estimaciones de la FR. El primer paso de “Smart Fusion” es evaluar la calidad de las estimaciones de FR derivadas de las tres modulaciones. Si las tres estimaciones se encontraban entre 4 lpm entre una y la otra, luego se generó una estimación de FR como la media de las estimaciones. En caso contrario, no se entrega ningún resultado.

26.5 Resultados

La tabla 26.2 muestra el error absoluto medio (MAE por sus siglas en inglés) para todos los métodos analizados. Los algoritmos más precisos antes de implementar los pasos de evaluación de calidad y fusión tenían un MAE de 4,28 lpm. Este algoritmo extrajo la VB de la FPG y la FR estimada usando detección de la respiración. Los algoritmos que usaban las señales de VB superaron a aquellos que usaban MA, que a su vez se desempeñaron mejor que los algoritmos que usaban MF. Aún más, aquellos que usaban detección de la respiración para estimar la FR superaron el desempeño de aquellos que usaban análisis de Fourier.

Se observó una mejora en precisión cuando se sumo el paso adicional de evaluación de calidad y fusión a los algoritmos de detección de la respiración.

Los MAE para el ECG y FPG disminuyeron de 4.87 a 3.92 lpm y de 4.28 a 3.36 lpm respectivamente. Esto se logró a expensas del número de ventanas

a partir de las cuales se estimaron las FR. Al usar este paso adicional 44% de las ventanas del ECG y 63% de las ventanas de FPG se descartaron por la evaluación de calidad. En forma interesante, no se observó mejora en la precisión al sumar estos pasos a un algoritmo basado en Fourier. Debe notarse que una proporción sustancial de datos disponibles para el análisis se descartó antes del análisis. Se pudo obtener una FR de referencia en solo 10% de las ventanas. Sumado a esto, 44% de las ventanas de ECG y 30% de las ventanas de FPG se descartaron debido a señal de baja calidad, indicando probablemente la presencia de artefactos de movimiento o desconexión de sensores. En consecuencia, sólo 6% de los datos del ECG y 7% de los datos de FPG se incluyeron en el análisis.

Tabla 26.2 El desempeño de los algoritmos aplicados al ECG y la FPG, medidos usando el error absoluto medio (MAE) medido en respiraciones por minuto (rpm)

Especificación del Algoritmo		MAE (rpm)	
Señal respiratoria	Estimación de la FR	ECG	FPG
VB	Detección de la respiración	4.87	4.28
MA	Detección de la respiración	4.95	5.58
MF	Detección de la respiración	8.48	7.95
VB	Fourier	7.51	8.18
MA	Fourier	8.69	11.14
MF	Fourier	13.16	12.11
VB, MA, MF	Detección de la respiración + evaluación de calidad + fusión	3.92	3.36
VB, MA, MF	Fourier + evaluación de calidad + fusión	12.66	10.52

26.6 Discusión

La FR es ampliamente usada en distintos ámbitos clínicos para ayudar en el diagnóstico y pronóstico. A pesar de su importancia clínica, es el único signo vital que no se mide electrónicamente fuera de las unidades de cuidados críticos. En este caso de estudio se han presentado técnicas para la estimación de la FR a partir de dos señales fisiológicas medidas

rutinariamente y en forma fácil, el ECG y la FPG. Encontramos dos hallazgos importantes. En primer lugar, la suma de un paso de calidad de señal y fusión a los algoritmos de detección respiratoria, aumentó su precisión. En segundo lugar, los algoritmos de detección respiratoria en el dominio del tiempo superaron el desempeño de aquellos en el dominio de la frecuencia. Esto sugiere que se necesita más investigación en los métodos en el dominio del tiempo, que son mucho menos dependientes de la FR, siendo casi estacionarios.

Si se nota que un método se desempeña lo suficientemente bien, luego podría ser utilizado para medir la FR durante las evaluaciones fisiológicas de rutina para brindar alertas tempranas de deterioro clínico.

El set de datos usado en este caso de estudio es un recurso útil para mayor prueba de los algoritmos de FR. Su fortaleza es que contiene datos de trazados de miles de pacientes críticamente enfermos, con varios set de datos con duración de horas o días. De todas formas, la generalización de los resultados es limitada dado que se consideraron exclusivamente de pacientes críticamente enfermos. Esto es particularmente significativo, considerando que los algoritmos de FR con frecuencia serían utilizados fuera de los Cuidados Críticos. Es más, la señal de NI dio una referencia confiable para sólo el 10% de los pacientes. Esto resultó en un número bajo de ventanas de señales incluidas en el análisis, lo cual es una limitación significativa.

En consecuencia, este caso de estudio debería tratarse como un ejemplo de la metodología que podría usarse para desarrollar un estudio robusto, antes que ser un estudio robusto en sí mismo. Aun más, algunas incertezas permanecieron en la FR de referencia, dado que eran el promedio de dos estimaciones que podrían diferir hasta 2 lpm. Cuando se prueban algoritmos para la extracción de parámetros clínicos a partir de señales fisiológicas, es mejor que los valores de referencia sean lo mas precisos posible. En este estudio los MAE medidos son probablemente más altos que los verdaderos MAE del algoritmo debido a las imprecisiones en la FR de referencia. Un desafío central del análisis de formas de ondas es manejar datos de baja calidad. Un enfoque es detectar y excluir datos de baja calidad como se realizó al utilizar el paso de evaluación de calidad y fusión. Aquí se usó una simple plantilla ICS. Técnicas más complejas que funden los resultados de múltiples ICS para determinar la calidad de la señal podrían mejorar el

desempeño de los algoritmos de FR en la práctica clínica [24,25]. Un enfoque alternativo consiste en refinar las técnicas de análisis para asegurar que mantienen precisión aun usando datos de calidad baja. Por ejemplo, en [26] se presenta un algoritmo para la estimación de FR a partir del ECG durante el ejercicio, cuando es probable que la señal sea de baja calidad.

26.7 Conclusiones

Este caso de estudio demuestra la utilidad potencial del ECG y PSG para la medición de la FR en el ámbito clínico. Se presentan las herramientas necesarias para diseñar y probar los algoritmos de FR, permitiendo a los lectores interesados ampliar el trabajo. Los resultados sugieren dos áreas particulares para mayor desarrollo de algoritmos. En primer lugar, el uso de la calidad de la señal y la fusión para mejorar la precisión de los algoritmos de FR deberían explorarse más. En la literatura, se ha puesto mucho foco en la extracción de señales respiratorias y estimación de FR mientras que se ha conducido relativamente poca investigación en la evaluación de la calidad y la fusión. En segundo lugar, debería conducirse más investigación en el uso de técnicas de dominio de tiempo para identificar los ciclos respiratorios individuales. Es notable que en este estudio, la técnica de dominio de tiempo supero la técnica de dominio de frecuencia, mientras que en la literatura se reporta que las técnicas de dominio de tiempo rara vez son mas sofisticadas que la detección pico. De todas formas, baja tasa de inclusión de datos sugiere que se necesita más investigación para asegurar que las conclusiones son robustas.

26.8 Direcciones futuras

Hay dos preguntas de investigación urgentes referentes a la estimación de la FR a partir de señales fisiológicas. En primer lugar, no es claro que algoritmo de FR es el más preciso. Hasta hace poco tiempo, los estudios de validación habían comparado solo unos pocos de los muchos algoritmos existentes. La comparación entre estudios es difícil ya que los estudios generalmente se realizan en diferentes sets de datos recolectados de distintas poblaciones, usando diferentes medidas estadísticas. Un estudio reciente evaluó varios algoritmos en datos obtenidos de sujetos jóvenes, sanos. En segundo lugar, no es claro si los algoritmos más precisos se desempeñan lo suficientemente bien para uso clínico. Se necesitan más

estudios para responder dichas preguntas. Proponemos que los algoritmos deberían probarse previamente en una población sana, idealmente en condiciones quirúrgicas. Esto facilitaria la evaluación del mejor desempeño posible de los algoritmos. Si cualquier algoritmo se desempeña lo suficientemente bien para uso clínico, luego podría ser probado en poblaciones de pacientes en el ámbito clínico. Por el contrario, si no hay algoritmos que funcionen adecuadamente, entonces se debe llevar a cabo un mayor desarrollo algorítmico para intentar mejorar el rendimiento. La base de datos MIMIC II provee la oportunidad de probar algoritmos en un amplio rango de condiciones fisiológicas, como hiper e hipotensión, y fracción de eyección normal y disminuida. Esto puede brindar un panorama de las limitaciones de los algoritmos, asegurando que solo son usados en condiciones en que se espera que se desempeñen bien.

26.9 Estimación de signos vitales sin contacto

Como se presentó en este capítulo, los sistemas de monitoreo actuales requieren contacto con el sujeto para seguir los cambios de sus signos vitales en contexto de internación o en su domicilio. La mayoría de los pacientes que requieren monitorización regular encuentran que las sondas son difíciles de fijar y usar en forma adecuada [28].

El proceso de registro de signos vitales se torna molesto aun si solo lleva pocos minutos dado que usualmente debe realizarse en forma diaria. El bajo cumplimiento de los pacientes para usar sensores también es un obstáculo para el monitoreo exitoso. La tecnología ideal para estimar los signos vitales involucraría sensores sin contacto directo con el paciente, brindando varias ventajas sobre los métodos tradicionales dado que no se necesita participación de los sujetos para configurar el equipo, no requiere preparación de la piel, no causa irritación cutánea, disminuye el riesgo de infección y tiene el potencial de integrarse en el estilo de vida del paciente. Se han propuesto diversas tecnologías para el monitoreo sin contacto de los signos vitales, desde sistemas basados en radares hasta ECG sin contacto usando electrodos de acoplamiento capacitivo. Durante la última década, con el costo de las video cámaras digitales en constante descenso, a medida que la tecnología se hace omnipresente, se ha expandido la investigación en monitoreo de signos vitales sin contacto, con el uso de video cámaras portátiles. Las video cámaras pueden encontrarse en las laptops, teléfonos

móviles, televisores en el living del paciente, abriendo nuevas posibilidades de monitoreo de signos vitales

El monitoreo de signos vitales basado en video, excede el concepto de la FPG tradicional usando múltiples fotogramas presentes en un sensor de imágenes para registrar los cambios de volumen sanguíneo asociados al ciclo cardíaco. Estos cambios fisiológicos resultan en una onda de pulso conocida como imagen de fotopletismografia a partir de la cual pueden estimarse los signos vitales como la frecuencia cardiaca, la frecuencia respiratoria, la saturación de oxígeno y otros [11, 29]. La figura 26.7 representa una muestra de 15 FPGi junto con señales de FPG y PI medidas utilizando equipamiento de monitoreo convencional. El paciente estaba recibiendo tratamiento de hemodiálisis en el Hospital Churchill de Oxford. En este periodo el paciente tenia una frecuencia cardiaca de 60 lpm y una respiratoria de 15 rpm, ambas podían calcularse a partir del equipamiento de monitoreo convencional y la cámara usando los métodos explicados en este capítulo. Décadas de investigación exhaustiva, de la comunidad de visión por computadora han colaborado para desarrollar sistemas de imágenes capaces de realizar complejos cálculos (detección facial, control de acceso de identidad y otros rastreadores de objetos), son interactivos (como movimiento, gesto y seguimiento de cuerpo en los juegos) y pueden desarrollar operaciones de reconstrucción 3D. Por lo tanto, el monitoreo de signos vitales basado en señales tiene el potencial de expandir el rol del monitoreo mas allá de lo alcanzable por la tradicional oximetría de pulso.

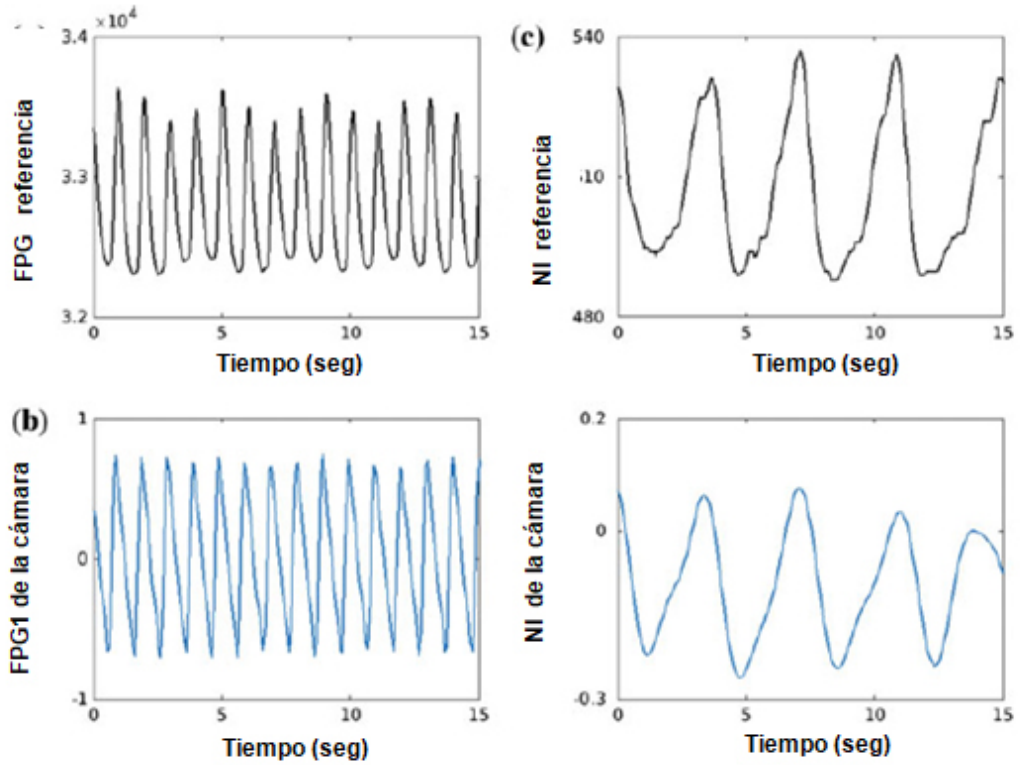

Fig 26.7 Muestra de datos de 15 segundos de un paciente en hemodiálisis en el Hospital Churchill en Oxford, a) Señal de referencia de FPG de un oxímetro de pulso Nonin b) Imagen de señal de FPG (FPGi) extraída de una videocámara c) Señal Respiratoria de Neumografía por Impedancia (NI) de referencia d) señal respiratoria extraída del trazado de FPGi. Durante el período el paciente tenía una frecuencia cardiaca de 60 latidos/minuto y una frecuencia respiratoria de 15 respiraciones/minuto (rpm)

Nota: Las imágenes de este capítulo se encuentran disponibles a color en el ebook; disponible en www.hardineros.com

Apéndice: Código

El código usado en este caso de estudio se encuentra disponible en el repositorio de GitHub que acompaña este libro: https://github.com/MIT-LCP/critical-data-book. En este sitio web se encuentra disponible más información acerca del código. Se utilizaron los siguientes scripts claves:

- MIMICII-data-importer.m: se uso para extraer datos de la base de datos MIMIC II.
- RRest.m: se usó para ejecutar los algoritmos RR y evaluar su desempeño.

Referencias

1. Shann F, Hart K, Thomas D (1984) Acute lower respiratory tract infections in children: possible criteria for selection of patients for antibiotic therapy and hospital admissión. Bull World Health Organ 62 (5): 749.
2. Lim WS, Van der Eerden MM, Laing R, Boersma WG, Karalus N, Town GI, Lewis SA, Macfarlane JT (2003) Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. Thorax 58 (5): 377-382.
3. Pollack MM, Ruttimann UE, Getson PR (1988) Pediatric risk of mortality (prism) score. Crit Care Med 16 (11): 1110-1116.
4. World Health Organization (WHO) (1990) Fourth Programme Report, 1988-1989: ARI Programme for Control of Acute Respiratory Infections. Technical Report, WHO, Geneva.
5. Lovett PB, Buchwald JM, Stürmann K, Bijur P (2005) The vexatious vital: neither clinical measurements by nurses nor an electronic monitor provides accurate measurements of respiratory rate in triage. Ann Emerg Med 45 (1): 68-76.
6. Chellel A, Fraser J, Fender V, Higgs D, Buras-Rees S, Hook L, Mummery L, Cook C, Parsons S, Thomas C (2002) Nursing observations on ward patients at risk of critical illness. Nurs Times 98 (46): 36-39.
7. Cretikos MA, Bellomo R, Hillman K, Chen J, Finfer S, Flabouris A (2008) Respiratory rate: the neglected vital sign. Med J Aust 188 (16): 657-659.
8. Hogan J (2006) Why don't nurses monitor the respiratory rates of patients? Br J Nurs 15 (9): 489-492.
9. Meredith DJ, Clifton D, Charlton P, Brooks J, Pugh CW, Tarassenko L (2012) Photoplethysmographic derivation of respiratory rate: a review of relevant physiology. J Med Eng Technol 36 (1): 1-7.
10. Bailon R, Sornmo L, Laguna P (2006) ECG-derived respiratory frequency estimation. In: Advanced methods and tools for ECG data analysis (Chap. 8). Artech House, London, pp 215-244.
11. Tarassenko L, Villarroel M, Guazzi A, Jorge J, Clifton DA, Pugh C (2014) Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. Physiol Meas 35 (5): 807-831.

12. Garde A, Karlen W, Ansermino JM, Dumont GA (2014) Estimating respiratory and heart rates from the correntropy spectral density of the photoplethysmogram. PLoS ONE 9 (1): e86427.
13. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE (2000) Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. Circulation 101 (23): E215-E220.
14. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG (2011) Multiparameter intelligent monitoring in intensive care II: apublic-access intensive care unit database. Crit Care Med 39 (5): 952-960.
15. Fleming S, Thompson M, Stevens R, Heneghan C, Plüddemann A, Maconochie I, Tarassenko L, Mant D (2011) Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: a systematic review of observational studies. Lancet 377 (9770): 1011-1018.
16. Nizami S, Green JR, McGregor C (2013) Implementation of artifact detection in critical care: a methodological review. IEEE Rev Biomed Eng 6:127-142.
17. Orphanidou C, Bonnici T, Charlton P, Clifton D, Vallance D, Tarassenko L (2015) Signal-quality indices for the electrocardiogram and photoplethysmogram: derivation and applications to wireless monitoring. IEEE J Biomed Health Inform 19 (3): 832-838.
18. Pimentel MAF, Charlton PH, Clifton DA (2015) Probabilistic estimation of respiratory rate from wearable sensors. In: Mukhopadhyay SC (ed) Wearable electronics sensors, vol 15. Springer International Publishing, pp 241-262.
19. Karlen W, Raman S, Ansermino JM, Dumont GA (2013) Multiparameter respiratory rate estimation from the photoplethysmogram. IEEE Trans Biomed Eng 60 (7): 1946-1953.
20. Schäfer A, Kratky KW (2008) Estimation of breathing rate from respiratory sinus arrhythmia: comparison of various methods. Ann Biomed Eng 36 (3): 476-485.
21. Pan J, Tompkins WJ (1985) A real-time QRS detection algorithm. IEEE Trans Biomed Eng 32 (3): 230-236.
22. Hamilton PS, Tompkins WJ (1986) Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database. IEEE Trans Biomed Eng 33 (12): 1157-1165.
23. Karlen W, Ansermino JM, Dumont G (2012) Adaptive pulse segmentation and artifact detection in photoplethysmography for mobile applications. In: Proceedings of the annual international conference of the IEEE engineering in medicine and biology society, vol 2012. EMBS, pp 3131-3134.
24. Behar J, Oster J, Li Q, Clifford GD (2013) ECG signal quality during arrhythmia and its application to false alarm reduction. IEEE Trans Biomed Eng 60 (6): 1660-1666.
25. Li Q, Clifford GD (2012) Dynamic time warping and machine learning for signal quality assessment of pulsatile signals. Physiol Meas 33 (9): 1491-1501.
26. Bailón R, Sörnmo L, Laguna P (2006) A robust method for ECG-based estimation of the respiratory frequency during stress testing. IEEE Trans Biomed Eng 53 (7): 1273-1285.

27. Charlton PH, Bonnici T, Tarassenko L, Clifton DA, Beale R, Watkinson PJ (2016) An assessment of algorithms to estimate respiratory rate from the electrocardiogram and photoplethysmogram. Physiol Measur 37 (4): 610-626.
28. Bonnici T, Orphanidou C, Vallance D, Darrell A, Tarassenko L (2012) Testing of wearable monitors in a real-world hospital environment: what lessons can be learnt? In: 2012 ninth international conference on wearable and implantable body sensor networks, pp 79-84.
29. Villarroel M, Guazzi A, Jorge J, Davis S, Watkinson P, Green G, Shenvi A, McCormick K, Tarassenko L (2014) Continuous non-contact vital sign monitoring in neonatal intensive care unit. Healthc Technol Lett 1 (3): 87-91.
30. Addison PS, Watson JN, Mestek ML, Mecca RS (2012) Developing an algorithm for pulse oximetry derived respiratory rate (RR (oxi)): a healthy volunteer study. J Clin Monit Comput 26 (1): 45-51.

[3] Las herramientas WFDB se encuentran disponibles en PhysioNet: http://physionet.org/physiotools/matlab/wfdb-appmatlab/.

Capítulo 27

PROCESAMIENTO DE SEÑALES: REDUCCIÓN DE FALSAS ALARMAS

Qiao Li y GariD. Clifford

Objetivos de aprendizaje

Usar la fusión de datos y el aprendizaje automático para eliminar falsas alarmas de arritmias.

Este caso de estudio introduce conceptos que deberían mejorar la comprensión de lo siguiente:

1. Extraer características relevantes de formas de onda clínicas
2. Evaluar la calidad de señal de los datos clínicos
3. Desarrollar un modelo de aprendizaje automático, entrenarlo y validarlo usando una base de datos clínica

27.1 Introducción

Los sistemas modernos de monitoreo de pacientes en terapia intensiva producen frecuentemente falsas alarmas que conducen a una disrupción en la atención, que impacta tanto en el paciente como en personal a través de perturbaciones de ruido, desensibilización a las alarmas y disminución en el tiempo de respuesta [1, 2]. Esto conduce a una disminución en la calidad del cuidado [3, 4], privación de sueño [1, 5, 6], disrupción en la estructura del sueño [7, 8], estrés para los pacientes y el personal de salud [9-12] y depresión del sistema inmune [13]. En unidades de cuidados intensivos (UCI) se han reportado índices de falsas alarmas mayores al 90% [14], solamente el 8% de las alarmas fueron consideradas alarmas con significado clínico [15] y más del 94% podrían no ser clínicamente importantes [16]. Hay dos razones principales que explican el alto índice de falsas alarmas. Una es que los datos fisiológicos pueden ser severamente corrompidos por artefactos (por ejemplo, por el movimiento), ruidos (por la interferencia eléctrica) y datos faltantes (por ejemplo, por inadecuado funcionamiento del transductor que lleva a cambios en la impedancia o la presión y una saturación de la señal resultante). La figura 27.1 ilustra las formas de onda de los monitores (o datos de alta resolución) registradas alrededor de una falsa alarma de taquicardia ventricular (la línea vertical indica el momento en el

cual el monitor gatilla la alarma). La alarma es causada por un ruido significativo que afecta las derivaciones del electrocardiograma (ECG). Sin embargo, los latidos pulsátiles regulares presentes en la derivación de la presión arterial (PA) indican claramente que es una falsa alarma (la disminución de la función de bomba durante esta arritmia debería causar una caída significativa en la amplitud del pulso y un incremento en la frecuencia). La otra razón de la alta frecuencia de falsas alarmas es que los monitores actuales utilizan en forma predominante algoritmos de alarma univariados y umbrales numéricos simples. La razón de esto es un artefacto histórico, en el que los fabricantes desarrollaron sistemas diferentes embebidos con un hardware hecho a medida y transductores unimodales. Los algoritmos de detección de alarmas univariados, por lo tanto, consideran una sola forma de onda monitorizada a la vez. La alarma es gatillada generalmente cuando una variable (por ejemplo, frecuencia cardíaca) derivada desde la forma de onda (por ejemplo, ECG) está por encima o por debajo de un umbral prefijado (o ajustable) para un determinado tiempo, en forma independiente de si el cambio es causado por un cambio en el estado fisiológico, por un artefacto o por intervenciones médicas, como mover o posicionar al paciente, realizar una extracción de sangre y limpiar la vía arterial, o desconectar al paciente del ventilador para la aspiración endotraqueal. Además, los umbrales de alarma se ajustan frecuentemente de forma “ad hoc” de acuerdo a qué tan molesta sea percibida la alarma por el equipo clínico de turno. Hay poca evidencia de que los umbrales de alarma sean optimizados para cualquier población o individuo, particularmente en un sentido multivariado.

Con la finalidad de suprimir falsas alarmas se utilizaron varios algoritmos de cancelación de ruidos, como filtros de mediana [17] o filtros de Kalman [18]. Aunque el ruido transitorio puede ser removido por un filtro de mediana, el mismo es brutalmente no-adaptativo. El filtro de Kalman, por otro lado, es un método de estimación de estado óptimo, que ha sido usado para mejorar la estimación de la frecuencia cardíaca (FC) y la presión arterial (PA) durante periodos de ruido y arritmias [18].

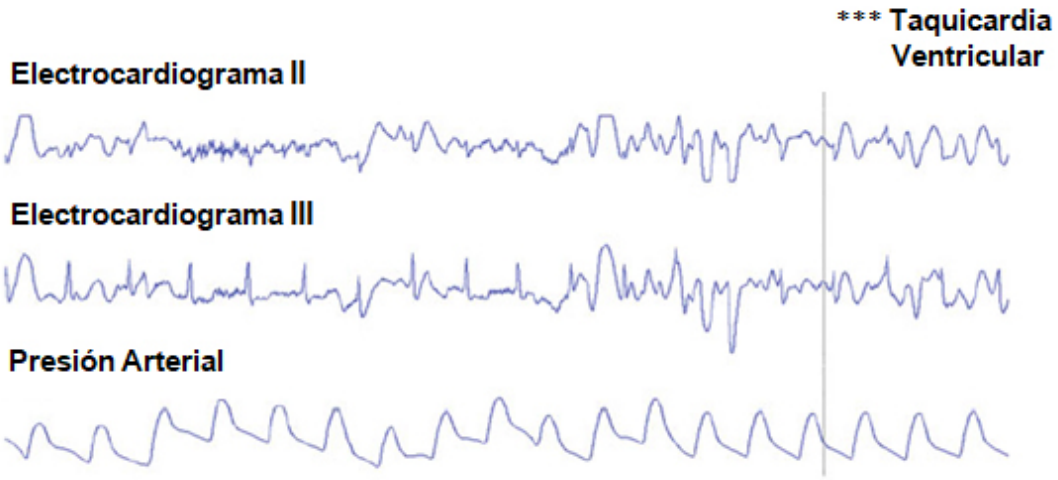

Fig. 27.1 Falsa alarma de taquicardia ventricular, "llamada" en el punto donde se sitúa la línea vertical, en una captura de 30 s de dos derivaciones de ECG (ECG II y ECG III) y una señal de presión arterial (PA). La alarma es gatillada por el fuerte ruido que se manifiesta como oscilaciones de alta amplitud (+/-2 mV) en el ECG de aproximadamente 5 Hz que comienza un poco por delante de la mitad de la captura (y un poco antes de 10s del señalador vertical de TV vertical). Note que la PA continúa normal, sin cambios significativos en el ritmo o la morfología.

De todas formas, la detección de alarmas cambió poco en décadas, persistiendo el paradigma de algoritmo de alarma univariado. Una solución prometedora al problema de las falsas alarmas viene de la fusión de datos de múltiples variables, como la estimación de FC fusionando la información de ECG, PA, y fotopletismografía (FPG) de la cual deriva la saturación de oxígeno [18]. Otero y col. [19] propusieron un modelo de perfil multivariable temporal *fuzzy* que describía un set de criterios de monitoreo de evolución temporal de las variables fisiológicas de los pacientes de FC, saturación de oxígeno (SpO2) y PA. Aboukhalil y col. [14] y Deshmane [20] usaron señales sincrónicas de PA y FPG para suprimir falsas alarmas de ECG. Zong et al. [21] redujeron las falsas alarmas de PA usando la relación entre ECG y PA. Además de los parámetros fisiológicos calculados, los índices de calidad de señal (ICS), que evalúan la utilidad de la forma de onda o los niveles de ruido de las formas de onda, también pueden extraerse de los datos crudos y usarse como factores de ponderación para permitir diversos niveles de confianza en los parámetros derivados. Behar y col. [22] y Li y Clifford [23] suprimieron falsas alarmas de ECG evaluando la calidad de señal de ECG, PA y FPG. Monasterio y col. [24] usaron máquinas de Vector Soporte (MVS) para

fusionar datos de señales respiratorias, frecuencia cardíaca y saturación de oxígeno derivado de ECG, FPG y neumograma de impedancia, además de varios ICS, para reducir falsas desaturaciones relacionadas a apneas.

27.2 Set de datos de estudio

En este estudio se utilizó un set de datos tomado de la base de datos de PhysioNet, MIMIC II [25, 26], que contenía registros simultáneos de ECG, PA y FPG, con 4107 alarmas de arritmias amenazantes para la vida señaladas por expertos [asistolia (AS), bradicardia extrema (BE), taquicardia extrema (TE) y taquicardia ventricular (TV)] en 182 admisiones a UCI. Se encontró un total de 2301 alarmas seleccionando las mismas cuando estaban disponibles el ECG, PA, y FPG. Los índices de falsa alarma fueron del 91.2% para AS, 26.6% para BE, 14.4% para TE, y 44.4% para TV respectivamente, y 45% en total. Las admisiones a UCI fueron divididas en dos sets separados para entrenamiento y evaluación, asegurando que la frecuencia de alarmas en cada categoría fuera dentro de todo igual a lo largo del ranking de frecuencia y separando las señales numeradas entre par e impar. La tabla 27.1 detalla la frecuencia relativa de cada categoría de alarma y sus índices asociados de verdaderas y falsas alarmas. Para cada alarma se extrajo la forma de onda de los datos desde los 30 segundos previos hasta los 10 segundos después de la alarma, para ayudar a la verificación de los expertos (dado que las guías de la Asociación para el Avance de la Instrumentación Médica (AAMI, del inglés Association for the Advancement of Medical Instrumentation) requieren que una alarma responda dentro de los 10 segundos del inicio de cualquier evento de alarma [27]). Se requirió un consenso de tres expertos para catalogar cada alarma como verdadera o falsa. Sólo se utilizaron los datos de los 10 segundos previos al comienzo de la alarma para la extracción de atributos automatizada y clasificación del modelo.

En el resto del capítulo nos enfocaremos en cómo reducir la falsa alarma de TV, dado que fue considerada en la literatura el tipo de falsa alarma más difícil de suprimir, con un índice bajo de reducción de falsas alarmas y un índice alto de supresión de alarmas verdaderas [14, 20-23, 28]. Para consultar métodos referidos a reducir falsas alarmas en los otros tipos de alarmas, los lectores interesados pueden dirigirse a Li y Clifford [23].

Tabla 27.1 Distribución de alarmas en el set de datos y en los set de entrenamiento y prueba.

Tipo de alarma	Total				Set de entrenamiento				Set de testing			
	F	V	Total	Tasa FA (%)	F	V	Total	Tasa FA (%)	F	V	Total	Tasa FA(%)
AS	260	25	285	91.2	166	14	180	92.2	94	11	105	89.5
BE	62	171	233	26.6	58	108	166	34.9	4	63	67	6.0
TE	37	220	257	14.4	19	116	135	14.1	18	104	122	14.8
TV	677	849	1526	44.4	306	478	784	39.0	371	371	742	50.0
Todas	1036	1265	2301	45.0	549	716	1265	43.4	487	549	1036	47.0

F falsa , V verdadera , FA falsa alarma AS asistolia, BE bradicardia extrema , TE taquicardia extrema , VT taquicardia ventricular

27.3 Preprocesamiento

Se extrajeron 147 características y medidas de ICS de las señales de ECG, PA, FPG, y SpO2 dentro de los 10 segundos de la ventana de análisis. Estas características en general fueron elegidas basándose en investigaciones previas de los autores y otros [14, 20-24 ,28-32]. Las características típicas incluían FC (extraída de ECG, PA, y FPG), presión arterial (sistólica, diastólica, media), saturación de oxígeno (SpO2), y la amplitud de FPG. Cada característica tenía 5 subcaracterísticas calculadas dentro de la ventana de 10 segundos: incluyendo el mínimo, máximo, la mediana, la desviación, y gradiente (derivado de un ajuste de mínimos cuadrados robusto sobre toda la ventana). Además de las características típicas, se extrajeron el área de diferencia de latidos (ADL), el radio de área de latidos (RAL) en el ECG, PA y FPG y trece métricas de fibrilación ventricular (tomadas de [29]). El área de cada latido fue definida como el área entre la forma de onda y el eje x, desde el comienzo del latido en el ECG hasta 0.6 veces la media del intervalo latido-latido (iLL). Note que el comienzo del latido en el ECG fue tomado como la posición del pico de R-0.2*iLL. El ADL fue calculado comparando cada latido con la mediana de los latidos en la ventana, como se muestra en la Fig. 27.2. El ADL usó cuatro sub-características; el ADL promedio de cinco latidos con los intervalos latido-latido más cortos, el máximo de la ADL promedio de cinco latidos consecutivos, la varianza y el gradiente del ADL. El RAL usó

cinco sub-características; el radio entre el área promedio de los cinco latidos más pequeños y los cinco latidos más grandes del ECG (RAL_{ECG}), PA (RAL_{PA}), y FPG (RAL_{FPG}), el ratio entre RAL_{ECG} y $RAL_{PA,}$ y el ratio entre RAL_{ECG} y RAL_{FPG}. La descripción de las trece métricas de las fibrilaciones ventriculares pueden encontrarse en Li y col. [29], e incluyen características de dominio espectral y de tiempo mostradas para permitir la clasificación altamente precisa de la TV. Las métricas de ICS del ECG incluyeron trece medidas [30], basados en momentos estándares, estadísticas de dominio de frecuencia y la coincidencia entre detectores de evento con diferente sensibilidad de ruido. Las métricas de ICS de la PA incluyeron un índice de anormalidad de señal con sus nueve sub-medidas [31] y una alineación temporal dinámica (ATD) basada en el enfoque de ICS con sus cuatro sub-medidas [32]. El ICS basado en ATD remuestrea cada latido para que coincida con una plantilla de latidos que se está ejecutando usando la ATD. El ICS fue entonces dado por el coeficiente de correlación entre la plantilla y cada latido. Las medidas de ICS para FPG incluían las ICSs basadas en ATD [32] y los primeros dos parámetros de Hjorth [20] que estimaban la frecuencia dominante y la mitad del ancho de banda de la distribución espectral de la FPG. Mientras éstas no representan necesariamente una lista de características exhaustivas, sí representan la amplia mayoría de características identificadas como útiles en estudios previos.

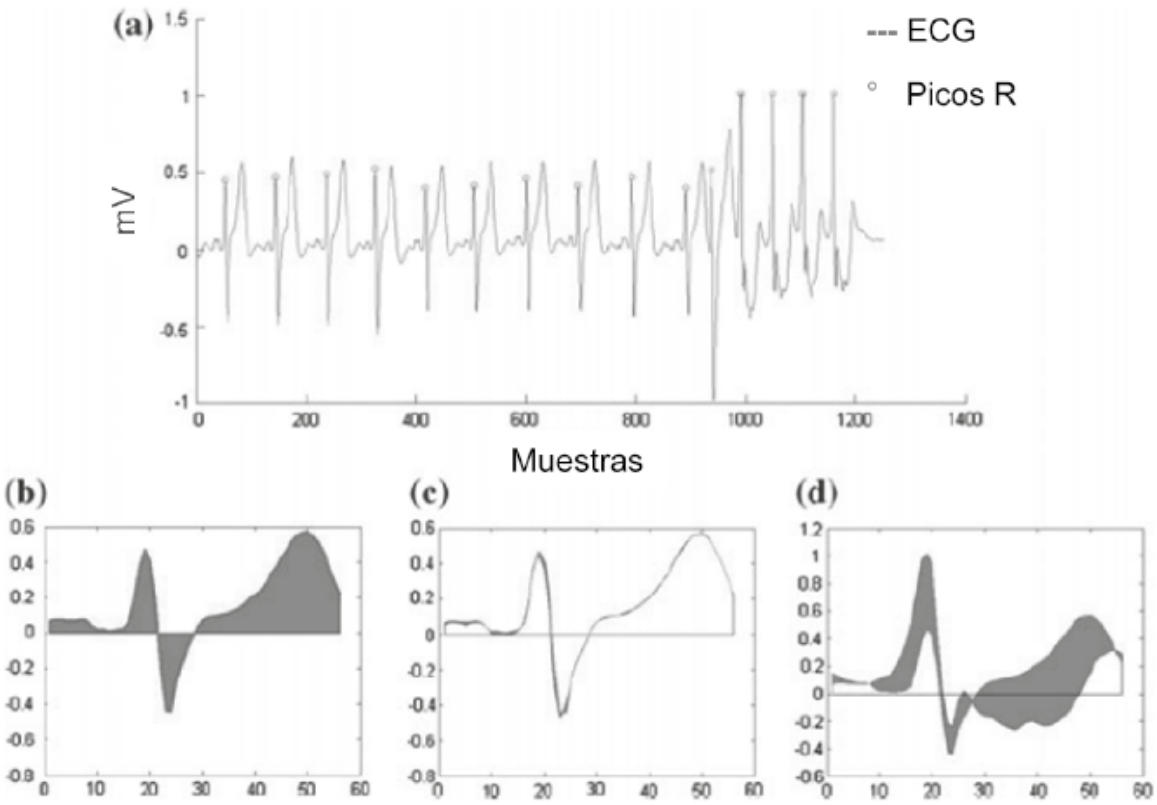

Fig. 27.2 Ejemplo de cálculo de diferencia de área de latidos. **a** ECG en una ventana de 10s. **b** La mediana de latidos en la ventana (área *gris* muestra el área entre la forma de onda y el eje x). **c** ADL de un latido normal (el primer latido, el área *gris* muestra el ADL). **d** ADL de un latido anormal (el último latido)

27.4 Métodos

Se utilizó un clasificador de ramdom forest (RF) modificado, previamente descripto por Johnson y col. [33]. El RF [34] es un método de aprendizaje automático ensamblado para clasificación que construye un número de árboles de decisión en el momento del entrenamiento y produce la clase, que es el modo de las clases de los árboles individuales. El principio básico es que un grupo de "aprendices débiles" puede combinarse para formar un "aprendiz robusto". Los RFs corrigen los defectos de los árboles de decisión de sobreajustar y de agregar sesgo a su set de entrenamiento. Cada árbol selecciona un subgrupo de observaciones por medio de dos regresiones derivadas de la división. A estas observaciones se les da una contribución igual al valor de la observación un número aleatorio constante de veces, para un atributo elegido más un intercepto aleatorio. Las contribuciones a través

de todos los árboles se suman para proveer la contribución para un solo "forest" o bosque, donde bosque se refiere a un grupo de árboles más un término intercepto. El resultado de la función de probabilidad predicha (*L*) por el bosque es el logaritmo inverso de la suma de la contribución de cada árbol más el término intercepto (27.1). El término intercepto es establecido al logit del resultado promedio observado.

$$L = \sum_{i=1}^{N} \left((-t_i) * \log\left(\text{logit}^{-1}(s_i)\right) - (1 - t_i) * \log\left(1 - \text{logit}^{-1}(s_i)\right) \right) \qquad (27.1)$$

Donde t_i es el objetivo del set de entrenamiento, s_i es la suma de la contribución del árbol, i=1… N es el número de observaciones en el set de entrenamiento. El núcleo del nuevo modelo RF que usamos es la muestra personalizada de la cadena Markov Monte Carlo (MCMC, del inglés Markov Chain Monte Carlo) que iterativamente optimiza el bosque. Este proceso de muestreo construye la cadena Markov por un proceso iterativo sin memoria que selecciona aleatoriamente dos árboles de los bosques actuales y actualiza su estructura. El MCMC muestrea aleatoriamente el espacio de observación por un gran número de iteraciones definidas por usuario. Después de estandarizar los datos de entrenamiento a una distribución normal estándar, el bosque es iniciado a un modelo nulo, sin contribuciones asignadas para ninguna observación.

En cada iteración, el algoritmo selecciona aleatoriamente dos árboles en el bosque y genera una aleatorización en su estructura. Es decir, aleatoriamente re-selecciona primero dos atributos que el árbol usa para dividir, el valor en el cual el árbol divide estos atributos, el tercer atributo usado para contribuir en los cálculos, y las constantes multiplicativas y aditivas que se aplican a este tercer atributo. La contribución total del bosque es entonces recalculada y se utiliza un paso de aceptación Metropolis-Hastings para determinar si se acepta la actualización. Se calculó la probabilidad predicha para el bosque previo (L_i) y la probabilidad del bosque con los dos árboles actualizados (L_{i+1}).

Si $e^{(Li + Li+1)}$ es mayor que un número real aleatorio uniformemente distribuido dentro de una unidad de intervalo, la actualización es aceptada. Si la actualización es aceptada, los dos árboles se conservan en el bosque, de otra forma están descartados y el bosque permanece sin cambios. Después

de una fracción del set del número total de iteraciones para permitir al bosque aprender la distribución objetivo (generalmente 20%), el algoritmo empieza a almacenar bosques en un intervalo fijado, es decir una vez por cada número de iteraciones. Una vez que se alcanza el número de iteraciones definidas por el usuario, el bosque es reiniciado como antes, y recomienza el proceso iterativo. Nuevamente, después del periodo de exposición del set, los bosques empiezan a ser guardados en un intervalo fijado. El resultado final de este algoritmo es un set de bosques, de los cuales cada uno contribuirá a la clasificación del modelo final. El diagrama de flujo del algoritmo RF es mostrado en la Fig. 27.3.

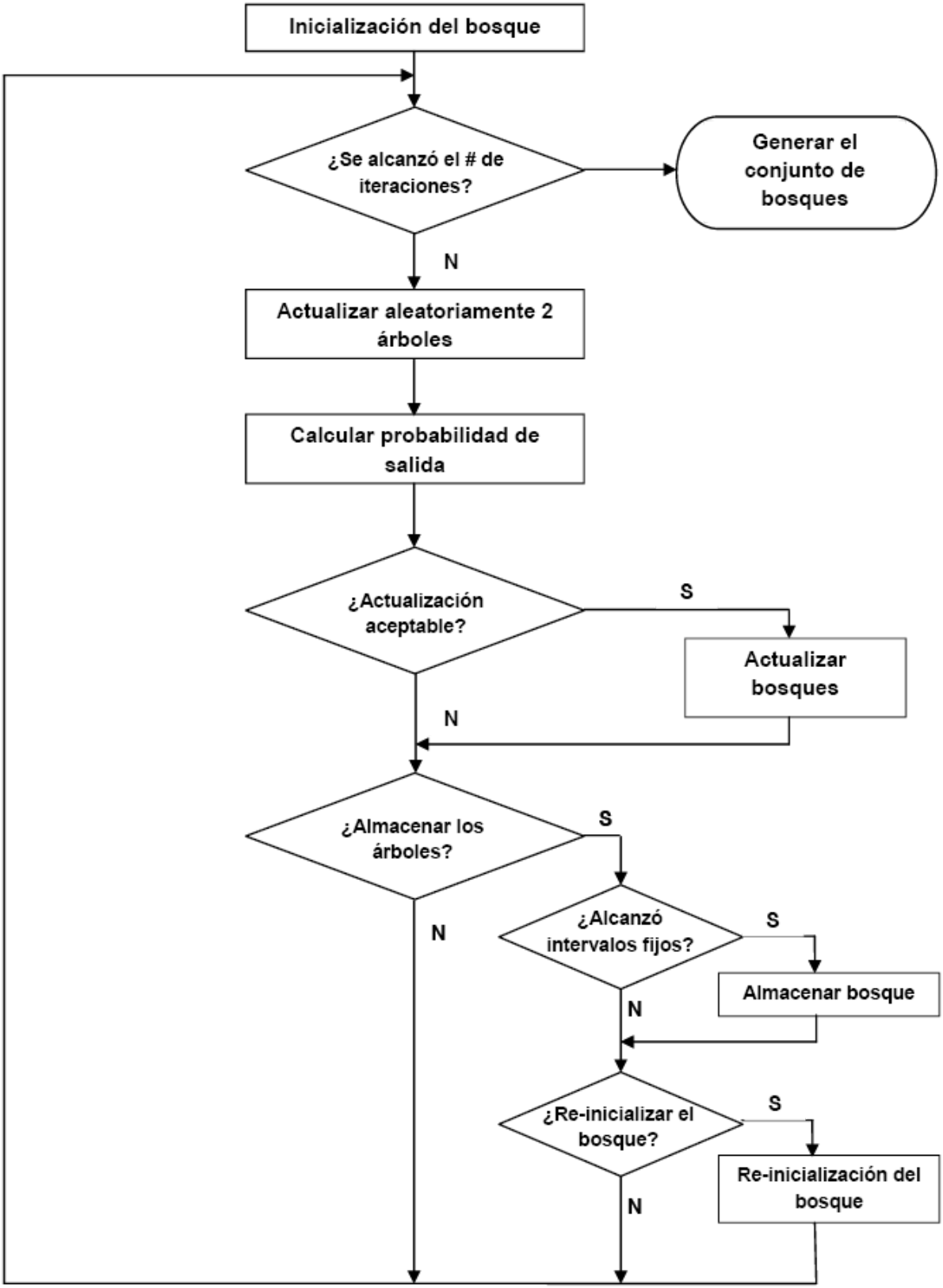
Inicialización del bosque
¿Se alcanzó el # de iteraciones?
Generar el conjunto de bosques
N
Actualizar aleatoriamente 2 árboles
Calcular probabilidad de salida
¿Actualización aceptable?
S
Actualizar bosques
N
¿Almacenar los árboles?
S
N
¿Alcanzó intervalos fijos?
S
Almacenar bosque
N
¿Re-inicializar el bosque?
S
Re-inicialización del bosque
N

Fig. 27.3 El diagrama de flujo del algoritmo de random forest.

27.5 Análisis

El modelo RF fue optimizado en el set de entrenamiento y evaluado en cuanto a precisión fuera-de-muestra en el set de prueba. Durante la fase de entrenamiento, se estableció un modelo de 320 bosques con 500 árboles en cada bosque. El resultado del modelo provee una probabilidad entre 0 y 1, que es un valor equivalente estimado para una falsa o verdadera alarma respectivamente. Se extrajo la curva de la Característica Operativa del Receptor (ROC) aumentando el umbral de la probabilidad en el cual cambiamos de falso a verdadero desde 0 a 1-es decir la probabilidad mayor que el umbral indica una verdadera alarma y por debajo (o igual) indica una falsa alarma. El punto de operación óptimo fue seleccionado en la curva ROC cuando la sensibilidad iguala 1 (no supresión de alarma verdadera) con la mayor especificidad. De todas formas, un punto de operación sub-óptimo fue también seleccionado con sensibilidad aceptable para balancear la especificidad, por ejemplo sensibilidad igual a 99%. (La razón para esto es que anecdóticamente, los expertos clínicos indicaron sería aceptable un 1% de tasa de supresión de alarma verdadera (o aumento en la tasa de supresión de alarma verdadera) (ver la discusión en las conclusiones del estudio). El modelo fue entonces evaluado en el set de prueba con los puntos de operación seleccionados.

En la fase de validación del algoritmo, la clasificación del desempeño del algoritmo se evaluó utilizando validación cruzada de 10 iteraciones. El proceso separaba el set de datos del estudio en diez partes estratificadas aleatoriamente por admisiones a UCI en vez de por las alarmas. Luego, nueve partes fueron usadas para un modelo de entrenamiento y la última fue usada para validación. Este proceso fue repetido diez veces como un procedimiento integral, con cada una de las partes usada exactamente una vez como datos de validación. Se usó el desempeño promedio para evaluación. Notamos de todas formas, que esto puede ser subóptimo y que el voto de todas las partes puede producir un mejor rendimiento.

27.6 Visualización

En la figura 27.4 se muestra la curva ROC en el set de entrenamiento. El punto óptimo de operación (marcado con un círculo) muestra sensibilidad

100% y especificidad 24.5%, indicando que suprimimos 24.5% de las falsas alarmas sin suprimir alarmas verdaderas. El punto sub óptimo de operación (marcado con una estrella) muestra una sensibilidad de 99.2% y una especificidad de 53.3%, indicando una reducción de 53.3% con solo un 0.8% de índice de supresión de alarmas verdaderas. Cuando el modelo fue usado en el set de prueba por el punto de operación óptimo, se logró una sensibilidad de 99.9% y una especificidad de 17%, con una sensibilidad del 99.5% y una especificidad del 44.2% para el punto de operación sub óptimo. En la tabla 27.2 se muestra el resultado de la validación cruzada de 10 iteraciones con diferentes opciones de puntos de operación.

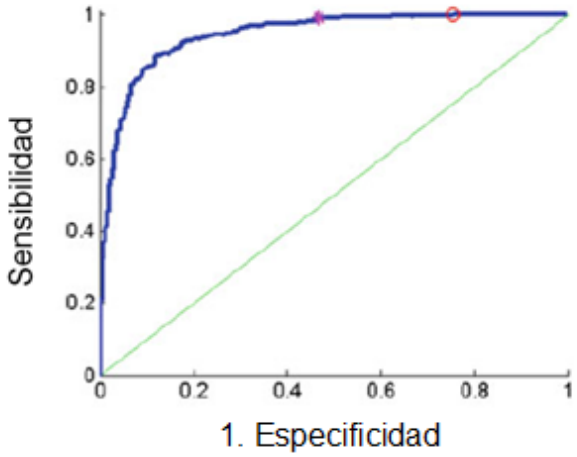

Fig. 27.4 Curva ROC para el set de entrenamiento. El *círculo* indica el punto de operación óptimo (en términos de aceptabilidad clínica) y una *estrella* un punto de operación sub óptimo que, de hecho, puede ser preferible.

Tabla 27.2 Resultado de la validación cruzada de 10 iteraciones del modelo de clasificación con diferentes puntos de operación.

Punto de operatividad (por sensibilidad) (%)	Entrenamiento (en 9 iteraciones)		Validación (en 1 iteración extendida)	
	Sensibilidad (%)	Especificidad (%)	Sensibilidad (%)	Especificidad (%)
99.00	99.06 +/- 0.04	56.41 +/- 5.60	95.82 +/- 5.62	51.68 +/- 16.88
99.50	99.56 +/- 0.04	49.08 +/- 5.37	96.50 +/- 5.39	45.19 +/- 17.94

99.60	99.66 +/- 0.04	43.49 +/- 6.45	98.72 +/- 2.06	38.14 +/- 17.25
99.70	99.75 +/- 0.03	39.50 +/- 7.39	98.75 +/- 2.08	32.07 +/- 16.19
99.80	99.87 +/- 0.02	34.57 +/- 9.02	98.87 +/- 2.11	28.16 +/- 15.80
100.0	100.0 +/- 0.00	27.85 +/- 6.17	99.04 +/- 2.02	18.10 +/- 9.87

27.7 Conclusiones

Mostramos aquí que un enfoque prometedor para suprimir las falsas alarmas, parece ser a través del uso de algoritmos multivariados, los cuales fusionan fuentes de datos sincrónicas y estimativos de la calidad subyacente para tomar decisiones. Las falsas alarmas de TV son las más difíciles de suprimir sin causar ninguna supresión de alarmas verdaderas dado que las ondas de PA y FPG pueden tener cambios de morfología indicando cambios hemodinámicos durante la TV. También mostramos que puede implementarse un modelo basado en Random Forest, con mucha confianza de que pocas alarmas verdaderas serían suprimidas (a pesar de que es imposible decir “nunca”). Un punto de operación práctico, puede ser seleccionado cambiando el umbral del modelo para balancear la sensibilidad y especificidad. Observamos que Aboukhalil et al. [14] y Sayadu y Shamsollahi [28] reportaron los mejores resultados en alarmas de TV, quienes lograron índices de supresión de falsas alarmas de TV de 33% y 66.7% respectivamente. Sin embargo, los índices de supresión de VA (verdaderas alarmas) que alcanzaron (9.4 y 3.8% respectivamente) son claramente demasiado altos para hacer sus algoritmos aceptables para esta categoría de alarma. En comparación con nuestros estudios previos usando algunos algoritmos de aprendizaje automático como máquina de Vector de Soporte [22] y de Vector de Relevancia [23], el algoritmo de Random Forest, que fusionaba los atributos extraídos de fuentes de datos sincrónicas como ECG, PA y FPG, proveía menor índice de supresión de VA y mayor índice de supresión de FA. Además, se necesita un procedimiento de validación sistemático, como la validación cruzada en k-iteraciones, para evaluar el algoritmo y notamos que trabajos previos no siguieron un protocolo similar. Sin dicha validación, es difícil creer que el algoritmo funcionará bien en datos no vistos a causa del sobreajuste. Esto es extremadamente importante de tener en cuenta, que es poco probable que se mantenga siempre un 0% de

supresión de alarma verdadera, y por eso probablemente sea aceptable una pequeña supresión de alarmas verdaderas. En discusiones privadas con nuestros asesores clínicos, frecuentemente se sugirió un 1%. En el trabajo presentado aquí, mostramos que con sólo suprimir 0,5% de verdaderas alarmas, pueden suprimirse casi la mitad de las falsas alarmas. Este índice de supresión de alarma verdadera es probablemente despreciable comparado con el número actual de alarmas perdidas inducidas-por-ruido desde el propio monitor. (Ningún monitor es perfecto, y se reportaron índices de falsos negativos de entre 0.5 y 5% [35]). También señalamos que el algoritmo propuesto aquí usaba sólo 10 segundos de datos antes de la alarma, lo cual cumple el requerimiento estándar de AAMI de 10 segundos [27]. En el trabajo reciente de PhysioNet/Computing in Cardiology Challenge 2015, se mostró que extender esta ventana un poco puede llevar a mejoras significativas en la supresión de falsas alarmas [36]. A pesar de que los entes reguladores necesitarían aprobar estos cambios, y esto frecuentemente es visto como poco probable, hacemos notar que la regla de 10 segundos es de alguna forma arbitraria y dicho trabajo puede influenciar los cambios en la aceptación regulatoria. Notamos varias limitaciones en nuestro estudio. Primero, el número de alarmas es todavía relativamente bajo, y provienen de una sola base de datos/fabricante. Segundo, la historia clínica, la demografía, y otros datos médicos no estaban disponibles y por ende no fueron utilizados para ajustar los umbrales. Finalmente, no se usó la información referente a las alarmas repetidas para ajustar la supresión de falsas alarmas de una manera dinámica basada en frecuencia de alarmas previas durante la misma estadía en UCI. Este último punto es particularmente complicado, dado que usar los datos de alarmas previas como información previa puede ser totalmente engañoso cuando los índices de falsa alarma no son despreciables.

27.8 Direcciones Futuras / Potenciales estudios de seguimiento

El problema de las falsas alarmas perturbó el monitoreo del paciente y preocupó a los fabricantes de monitores por muchos años, aun así el manejo de alarmas no vio el mismo progreso que el resto de la tecnología de monitoreo médico. Una razón importante es que en el ambiente actual legal y regulatorio, se puede alegar que los fabricantes tienen presión externa para proveer los algoritmos de alarma más sensibles de forma tal que ningún

evento crítico pase sin ser detectado [4]. De igual forma, uno podría argumentar que los médicos también tienen un imperativo para asegurar que ninguna alarma crítica pase sin ser detectada, y estamos dispuestos a aceptar un gran número de falsas alarmas para evitar un solo evento perdido. Un gran número de algoritmos y métodos emergieron en esta área [4, 14, 17-24, 28, 37, 38]. De todas formas, la mayoría de estos enfoques están todavía en un paso experimental y hay todavía un largo camino por recorrer antes de que los algoritmos se encuentren listos para aplicación clínica.

El evento PhysioNet/Computing in Cardiology Challenge de 2015 apuntó a promover el desarrollo de algoritmos para reducir la incidencia de falsas alarmas en UCI [36]. En este desafío se usaron datos de los monitores que incluían un total de 1250 alarmas de arritmias amenazantes para la vida, registradas de unidades de monitoreo de los tres principales fabricantes de monitores de las unidades de cuidados intensivos. Es probable que estos desafíos estimulen en la industria de monitoreo el interés renovado por el problema de las falsas alarmas. Además, el compromiso de la comunidad científica delineará otras cuestiones más sutiles. Quizás las tres cuestiones claves que faltan ser mencionadas son: (1) ¿Cuántas alarmas deberían ser anotadas y por cuántos expertos? (ver Zhu y col [39] para una discusión detallada de este punto); (2) ¿Cómo deberíamos tratar las alarmas repetidas, pasando información de una alarma a la siguiente?; (3) ¿Qué datos adicionales deberían ser aportados al monitor como información previa a la alarma? Esto podría incluir una historia de taquicardia, hipertensión, dosaje de droga, intervenciones y otra información relacionada incluyendo puntajes de severidad de enfermedad. Finalmente, notamos que las alarmas por eventos que comprometen la vida son muchos menos frecuentes que otras alarmas menos críticas, y por lejos el mayor contribuyente a la contaminación por alarmas en cuidados críticos viene de estas alarmas menos importantes. También se necesita un enfoque sistemático para estas alarmas menos urgentes, tomando prestado el marco de trabajo aquí presentado. En forma más prometedora, la tolerancia a la supresión de verdaderas alarmas, es probable que sea mucho más alta para alarmas menos importantes, y por lo tanto esperamos ver tasas muy altas de supresión de falsas alarmas. Esto es particularmente importante, dado que las técnicas descriptas aquí son generales y podrían aplicar a la mayoría de

las falsas alarmas no-críticas, que constituyen la mayoría de tales eventos en UCI. A pesar de que la competencia no aborda estos cuatro puntos (y de hecho los datos que se necesitan para hacerlo quedan por estar disponibles en grandes números), proveerá un estímulo para dichas discusiones y las herramientas (datos y códigos) ayudarán a continuar la evolución en este campo.

Nota: Las imágenes de este capítulo se encuentran disponibles a color en el ebook; disponible en www.hardineros.com

Referencias

1. Chambrin MC (2001) Review: alarms in the intensive care unit: how can the number of false alarms be reduced? Crit Care 5 (4): 184-188.
2. Cvach M (2012) Monitor alarm fatigue, an integrative review. Biomed Inst Tech 46 (4): 268-277.
3. Donchin Y, Seagull FJ (2002) The hostile environment of the intensive care unit. Curr Opin Crit Care 8 (4): 316-320.
4. Imhoff M, Kuhls S (2006) Alarm algorithms in critical care monitoring. Anesth Analg 102 (5): 1525-1537.
5. Meyer TJ, Eveloff SE, Bauer MS, Schwartz WA, Hill NS, Millman RP (1994) Adverse environmental conditions in the respiratory and medical ICU settings. Chest 105 (4): 1211-1216.
6. Parthasarathy S, Tobin MJ (2004) Sleep in the intensive care unit. Intensive Care Med 30 (2): 197-206.
7. Johnson AN (2001) Neonatal response to control of noise inside the incubator. Pediatr Nurs 27 (6): 600-605.

8. Slevin M, Farrington N, Duffy G, Daly L, Murphy JF (2000) Altering the NICU and measuring infants' responses. Acta Paediatr 89 (5): 577-581.
9. Cropp AJ, Woods LA, Raney D, Bredle DL (1994) Name that tone. The proliferation of alarms in the intensive care unit. Chest 105 (4): 1217-1220.
10. Novaes MA, Aronovich A, Ferraz MB, Knobel E (1997) Stressors in ICU: patients'evaluation. Intensive Care Med 23 (12): 1282-1285.
11. Topf M, Thompson S (2001) Interactive relationships between hospital patients' noise induced stress and other stress with sleep. Heart Lung 30 (4): 237-243.
12. Morrison WE, Haas EC, Shaffner DH, Garrett ES, Fackler JC (2003) Noise, stress, and annoyance in a pediatric intensive care unit. Crit Care Med 31 (1): 113-119.
13. Berg S (2001) Impact of reduced reverberation time on sound-induced arousals during sleep. Sleep 24 (3): 289-292.
14. Aboukhalil A, Nielsen L, Saeed M, Mark RG, Clifford GD (2008) Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform. J Biomed Inform 41 (3): 442-451.
15. Tsien CL, Fackler JC (1997) Poor prognosis for existing monitors in the intensive care unit. Crit Care Med 25 (4): 614-619.
16. Lawless ST (1994) Crying wolf: false alarms in a pediatric intensive care unit. Crit Care Med 22 (6): 981-985.
17. Mäkivirta A, Koski E, Kari A, Sukuvaara T (1991) The median filter as a preprocessor for a patient monitor limit alarm system in intensive care. Comput Meth Prog Biomed 34 (2-3): 139-144.
18. Li Q, Mark RG, Clifford GD (2008) Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter. Physiol Meas 29 (1): 15-32.
19. Otero A, Felix P, Barro S, Palacios F (2009) Addressing the flaws of current critical alarms: a fuzzy constraint satisfaction approach. Artif Intell Med 47 (3): 219-238.
20. Deshmane AV (2009) False arrhythmia alarm suppressión using ECG, ABP, and photoplethysmogram.M.S. thesis, MIT, USA.
21. Zong W, Moody GB, Mark RG (2004) Reduction of false arterial blood pressure alarms using signal quality assessment and relationships between the electrocardiogram and arterial blood pressure. Med Biol Eng Comput 42 (5): 698-706.
22. Behar J, Oster J, Li Q, Clifford GD (2013) ECG signal quality during arrhythmia and its application to false alarm reduction. IEEE Trans Biomed Eng 60 (6): 1660-1666.
23. Li Q, Clifford GD (2012) Signal quality and data fusión for false alarm reduction in the intensive care unit. J Electrocardiol 45 (6): 596-603.
24. Monasterio V, Burgess F, Clifford GD (2012) Robust classification of neonatal apnoea-related desaturations. Physiol Meas 33 (9): 1503-1516.

25. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE (2000) Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation 101 (23): e215-e220.
26. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG (2011) Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database. Crit Care Med 39 (5): 952-960.
27. American National Standard (ANSI/AAMI EC13:2002) (2002) Cardiac monitors, heart rate meters, and alarms. Association for the Advancement of Medical Instrumentation, Arlington, VA.
28. Sayadi O, Shamsollahi M (2011) Life-threatening arrhythmia verification in ICU patients using the joint cardiovascular dynamical model and a Bayesian filter. IEEE Trans Biomed Eng 58 (10): 2748-2757.
29. Li Q, Rajagopalan C, Clifford GD (2014) Ventricular fibrillation and tachycardia classification using a machine learning approach. IEEE Trans Biomed Eng 61 (6): 1607-1613.
30. Li Q, Rajagopalan C, Clifford GD (2014) A machine learning approach to multi-level ECG signal quality classification. Comput Meth Prog Biomed 117 (3): 435-447.
31. Sun JX, Reisner AT, Mark RG (2006) A signal abnormality index for arterial blood pressure waveforms. Comput Cardiol 33:13-16.
32. Li Q, Clifford GD (2012) Dynamic time warping and machine learning for signal quality assessment of pulsatile signals. Physiol Meas 33 (9): 1491-1501.
33. Johnson AEW, Dunkley N, Mayaud L, Tsanas A, Kramer AA, Clifford GD (2012) Patient specific predictions in the intensive care unit using a Bayesian ensemble. Comput Cardiol 39:249-252.
34. Breiman L (2001) Random forests. Mach Learn 45 (1): 5-32.
35. Schapira RM, Van Ruiswyk J (2002) Reduction in alarm frequency with a fusión algorithm for processing monitor signals. Meeting of the American Thoracic Society. Sessión A56, Poster H57.
36. Clifford GD, Silva I, Moody B, Li Q, Kella D, Shahin A, Kooistra T, Perry D, Mark RG (2006) The PhysioNet/computing in cardiology challenge 2015: reducing false arrhythmia alarms in the ICU. Comput Cardiol 42:1-4.
37. Borowski M, Siebig S, Wrede C, Imhoff M (2011) Reducing false alarms of intensive care online-monitoring systems: an evaluation of two signal extraction algorithms. Comput Meth Prog Biomed 2011:143480.
38. Li Q, Mark RG, Clifford GD (2009) Artificial arterial blood pressure artifact models and an evaluation of a robust blood pressure and heart rate estimator. Biomed Eng Online 8:13.
39. Zhu T, Johnson AEW, Behar J, Clifford GD (2014) Crowd-sourced annotation of ECG signals using contextual information. Ann Biomed Eng 42 (4): 871-884.

Capítulo 28

MEJORANDO LA IDENTIFICACIÓN DE COHORTES DE PACIENTES MEDIANTE EL PROCESAMIENTO DEL LENGUAJE NATURAL

Raymond Francis Sarmiento
y Franck Dernoncourt

Objetivos de aprendizaje

Comparar y evaluar el desempeño del método de extracción estructurada de datos y el procesamiento del lenguaje natural al identificar cohortes de pacientes usando la base de datos MIMIC-III

1. Identificar una cohorte especifica de pacientes de la base de datos MIMIC-III buscando las tablas estructuradas utilizando códigos de diagnósticos y procedimientos CIE-9
2. Identificar una cohorte especifica de pacientes de la base de datos MIMIC-III buscando texto libre no estructurado contenido en las notas clínicas usando una herramienta clínica de procesamiento de lenguaje natural (PLN) que aprovecha la detección de negación y el Sistema de Lenguaje Médico Unificado (UMLS, del inglés Unified Médical Language System) para encontrar términos médicos sinónimos
3. Evaluar el desempeño del método de extracción estructurada de datos y el método de PLN usados para la identificación de cohortes de pacientes

Ambos autores han contribuido en forma igual a este trabajo.

MIT Critical Data, SecondaryAnalysis of ElectronicHealth Records,
DOI 10.1007/978-3-319-43742-2-28

28.1 Introducción

Un área activa de investigación en la comunidad informática biomédica involucra el desarrollo de técnicas para identificar cohortes de pacientes para estudios clínicos e investigación que involucra el uso secundario de datos de las HCEs. En ese sentido, la expansión de las bases de datos a partir de HCEs que contienen tanto información estructurada como no estructurada ha sido beneficiosa para los investigadores clínicos. Ha colaborado para identificar

individuos que podrían ser elegibles para estudios clínicos así como ha permitido conducir estudios retrospectivos para validar potencialmente los resultados de los estudios clínicos prospectivos en menor tiempo y costo [1]. También ha ayudado a los médicos clínicos a identificar pacientes con mayor riesgo de desarrollar enfermedades crónicas, especialmente aquellas que podrían beneficiarse de tratamiento precoz [2]. Varios estudios han investigado la precisión de los datos administrativos estructurados como los códigos de facturación de la novena revisión de Clasificación Internacional de Enfermedades (CIE-9) de la Organización Mundial de la Salud (OMS) para identificar cohortes de pacientes [3-11]. La extracción de información estructurada utilizando los códigos de la CIE-9 ha mostrado tener una buena detección, precisión y especificidad [3, 4] para identificar poblaciones de pacientes distintas. De todas formas, para grandes bases de datos, la extracción de la información puede requerir mucho tiempo, ser costosa y poco práctica cuando se conduce en diferentes fuentes de datos [12] y se aplica a grandes cohortes de pacientes [13]. Utilizar consultas estructuradas para extraer información de las bases de datos de las HCE permite recuperar datos fácilmente y de manera más eficiente. Los datos estructurados de las HCE son generalmente útiles, pero pueden contener información incompleta o inexacta, en especial cuando cada elemento se toma en forma aislada. Por ejemplo [14], para justificar una orden de un laboratorio particular o test radiológico los médicos clínicos con frecuencia asignan a los pacientes códigos diagnósticos que corresponden a una condición sospechada. Pero aún cuando los resultados de la prueba diagnóstica muestren que el paciente no tenía dicha condición, el código de diagnóstico generalmente permanece en el registro médico del paciente.

Cuando se ve el código de diagnóstico sin el contexto (esto es, sin el beneficio de comprender los detalles del caso que se brindan en la narrativa de la historia del paciente), esto se convierte en un problema porque limita la capacidad de los investigadores de identificar con precisión las cohortes de pacientes y de utilizar todo el potencial estadístico de las poblaciones disponibles. En comparación con la narrativa de las notas clínicas, confiar solamente en datos estructurados como códigos diagnósticos puede ser poco fiable porque podrían no ser capaces de brindar información acerca del contexto clínico general. Sin embargo, el análisis automático de grandes volúmenes de notas clínicas requiere el uso de procesamiento de lenguaje

natural (PLN). El campo de estudio del analisis automatizado de datos de texto no estructurados es conocido como PLN y ya se ha utilizado con cierto éxito en el campo de la medicina. En este capítulo nos enfocaremos en cómo puede usarse el PLN para extraer información de datos no estructurados para identificación de cohortes. El PLN es un campo de las ciencias de la computación y lingüística que tiene como objetivo comprender el lenguaje humano (natural) y facilita interacciones más efectivas entre los humanos y las máquinas [13, 15].

En el campo de la clínica, el PLN ha sido utilizado para extraer información relevante como resultados de laboratorio, medicaciones y diagnósticos desde el texto libre de registros médicos de-identificados, para identificar cohortes de pacientes que cumplan con criterios de elegibilidad para estudios de investigación clínica [16].

Cuando se compara con la revisión de historias clínicas por humanos, el PLN produce resultados más rápidos [17-20]. Las técnicas de PLN también han sido usadas para identificar pacientes con posible cáncer de pulmón basado en sus informes radiológicos [21] y extraer características de la enfermedad para pacientes con cáncer de próstata [22].

Nosotros consideramos las condiciones crónicas en las que era probable que el diagnóstico de la enfermedad y el de la intervención se encontraran juntos en un intento por destacar mejor las diferencias entre las técnicas de recuperación estructuradas y no estructuradas, en especial dado el limitado número de estudios que han mirado como resultados a las intervenciones o tratamientos antes que a la enfermedad [14]. La población diabética fue de particular interés para esta tarea de PLN dado que las numerosas complicaciones cardiovasculares, oftalmológicas y renales asociadas con la diabetes mellitus eventualmente requieren tratamientos de intervención o procedimientos como hemodiálisis en este caso. Además, las notas clínicas frecuentemente contienen abreviaturas médicas y acrónimos, donde el uso de PLN puede ayudar a capturar y ver estas informaciones correctamente en los registros médicos.

Por lo tanto, en este caso de estudio, intentamos determinar si el uso de PLN en notas clínicas no estructuradas de esta población podría ayudar a mejorar la extracción de datos estructurada. Nosotros identificamos una cohorte de pacientes diabéticos críticamente enfermos que presentaban

falla renal terminal que realizaron hemodiálisis utilizando la base de datos MIMIC-III [23].

28.2 Métodos

28.2.1 Set de datos de estudio y preprocesamiento

Todos los datos de este estudio fueron extraídos de la base de datos MIMIC-III, la cual contiene datos de-identificados [24] por las reglas de privacidad de HIPPA (en inglés, Health Insurance. Portability and Accountability Act) [25], de más de 58.000 ingresos hospitalarios en la Unidad de Cuidados Intensivos (UCI) del Centro Médico Beth Israel entre junio del 2001 y octubre del 2012 [26]. Elegimos MIMIC-III porque, además de ser accesible públicamente, contiene datos detallados de la HCE de pacientes críticamente enfermos con probabilidad de presentar múltiples condiciones crónicas, incluyendo aquellas con complicaciones que podrían requerir intervenciones de tratamiento para salvar la vida. Excluímos todos los pacientes menores de 18 años, con diagnóstico de diabetes insípida aislada, sin diabetes mellitus, aquellos que recibieron solamente diálisis peritoneal sin hemodiálisis o aquellos con diagnóstico de condiciones transitorias como diabetes gestacional o diabetes inducida por esteroides sin historia médica de diabetes mellitus. También excluimos aquellos pacientes que recibieron hemodiálisis antes de su ingreso hospitalario pero que no la recibieron durante la internación. De los sujetos restantes, incluimos aquellos diagnosticados con diabetes mellitus y aquellos que requirieron hemodiálisis durante su internación en la UCI. Extrajimos los datos de dos fuentes primarias: las tablas MIMIC-III estructuradas (diagnósticos y procedimientos de alta) y notas clínicas no estructuradas.

28.2.2 Extracción de datos estructurados de las tablas de la base de datos MIMIC-III

Utilizando los códigos diagnósticos CIE-9 de la tabla diagnósticos de egreso y los códigos de procedimientos CIE-9 de la tabla procedimientos, buscamos una base de datos de acceso público para encontrar códigos de diagnósticos de enfermedad y de procedimientos relacionados con diabetes y con hemodiálisis según se muestra en la Tabla 28.1

Utilizamos lenguaje de consultas estructuradas (SQL) para encontrar pacientes en cada una de las tablas de datos estructuradas basándonos en códigos específicos CIE-9.

Tabla28.1 Códigos CIE-9 y descripciones que indican que un paciente fue diagnosticado diabetesmellitus y que potencialmente realizóhemodiálisisa partir de tablas de datos estructurados enMIMIC-III

Tabla de datos estructurada	**Codigo CIE-9 y descripcion**
Diabetes mellitus	
Códigos diagnósticos de egreso	249 diabetes mellitus secundaria (incluye los siguientes códigos: 249, 249.0, 249.00, 249.01, 249.1, 249.10, 249.11, 249.2, 249.20, 249.21, 249.3, 249.30, 249.31, 249.4, 249.40, 249.41, 249.5, 249.50, 249.51, 249.6, 249.60, 249.61, 249.7, 249.70, 249.71, 249.8, 249.80, 249.81, 249.9, 249.90, 249.91)
	250 diabetes mellitus (incluye los siguientes códigos : 250, 250.0, 250.00, 250.01, 250.02, 250.03, 250.1, 250.10, 250.11, 250.12,250.13, 250.2, 250.20, 250.21, 250.22, 250.23, 250.3, 250.30, 250.31, 250.32, 250.33, 250.4, 250.40, 250.41, 250.42, 250.43, 250.5, 250.50, 250.51, 250.52, 250.53, 250.6, 250.60, 250.61, 250.62, 250.63, 250.7, 250.70, 250.71, 250.72, 250.73, 250.8, 250.80, 250.81, 250.82, 250.83, 250.9, 250.90, 250.91, 250.92, 250.93)
Hemodiálisis	
Códigos diagnósticos de egreso	585.6 enfermedad renal terminal(requiere diálisis crónica)
	996.1 complicación mecánica de otro dispositivo vascular, implante e injerto
	996.73 otras complicaciones debidas a dispositivos de diálisis renal, implantes e injertos
	E879.1diálisis renal como la causa de reacción anormal de un paciente o complicación tardía sin mención de complicación en el momento del procedimiento
	V45.1 estado postquirurgico de diálisis renal
	V56.0 encuentro para diálisis extracorporeo
	V56.1 ajuste y adaptación de catéter de diálisis extracorporeo
Códigos de procedimientos	38.95 cateterización venosa para diálisis renal
	39.27 arteriovenostomia para diálisis renal
	39.42 revisión de shunt arteriovenoso para diálisis renal
	39.43 retiro de shunt arteriovenoso para diálisis renal
	39.95 Hemodiálisis

28.2.3 Extracción de datos no estructurados de notas clínicas

Las notas clínicas no estructuradas incluyen las epicrisis (n=52.746), las notas de enfermería (n=812.128), notas médicas (n= 430.629), informes de electrocardiogramas (ECG) (n=209.058), informes de ecocardiogramas (n= 45.794) e informes radiológicos (n=896.478). Excluimos las notas clínicas relacionadas con cualquier resultado de imágenes (ECG-Report, Echo-Report, and Radiology-Report). Extrajimos las notas de MIMIC-III con los siguientes elementos: número de identificación de paciente (SUBJECT_ID), número de identificación de ingreso hospitalario (HADM_IDs), número de identificación de estadía en la Unidad de Cuidados Intensivos (ICUSTAY_ID), tipo de nota, nota de día/hora y nota de texto.

Usamos una consulta SQL para extraer información pertinente de todas las notas de los pacientes que serán útiles para identificar un paciente como alguien que pertenece a una cohorte, luego escribimos un script en Python para filtrar las notas, buscando palabras claves e implementamos heurística para refinar los resultados de nuestra búsqueda.

Como parte de nuestra estrategia de búsqueda, eliminamos las secciones de historia familiar al buscar notas clínicas y nos aseguramos que la búsqueda de acrónimos no rescatara aquellos que eran parte de otra palabra.

Por ejemplo, nuestros filtros no recuperaban aquellos donde aparecía “DM” como parte de otras palabras como en “admission’ o ‘admit’. Finalmente usamos cTAKES [28,29] versión 3.2 con acceso a los conceptos del Sistema de Lenguaje Médico Unificado (UMLS) [30] para utilizar el anotador de detección de negación al buscar en la nota de texto. La característica de detección de negación en cTAKES funciona tratando de identificar aquellas entidades que son negadas en el texto. Ejemplos de palabras de negación que pueden encontrarse en las notas clínicas incluyen: ‘no’, ‘nunca’, ‘rechazar’, ‘declinar’ (en inglés ‘not’, ‘no’, ‘never’, ‘hold’, ‘refuse’, ‘declined’).

Por ejemplo, en este caso de estudio, si “DM” o “HD” es negado consistentemente al buscar en las notas clínicas, el paciente no debería ser considerado parte de la cohorte.

El Metathesaurus [31] en UMLS contiene vocabularios de salud y biomédicos, ontologias y terminologías standard, incluyendo CIE. Cada término es asignado a uno o más conceptos en UMLS. Diferentes términos de diferentes vocabularios u ontologías que tienen significados similares y se asignan con el mismo identificador de concepto único (CUI) se consideran sinónimos en UMLS [32].

A los fines de identificar los pacientes con diabetes mellitus que recibieron hemodiálisis durante su estadía en UCI, revisamos las notas clínicas que contenían los términos “diabetes mellitus” y “hemodiálisis”. Utilizamos el Metathesaurus UMLS para obtener sinónimos de estos términos porque usando solamente estos 2 términos restringirían los resultados de nuestra búsqueda.

cTAKES es un sistema de procesamiento natural de lenguaje que extrae información de texto libre almacenado en las HCEs. Acepta tanto documentos de texto plano como archivos con arquitectura de documentos clínicos (CDA) que cumplen con las normas de lenguaje de marcas extensible (XML, del inglés eXtensible Markdown Language) y consta de distintos anotadores, como el extractor de atributos (anotadores de aserción), gestión de documentos clínicos, chunker, parser de poblaciones, generadores de tokens dependientes del contexto, analizadores de dependencia y semántica, detección de negación, preprocesador de documentos, extractor de relaciones y búsqueda en el diccionario, entre otros [33].

Cuando se realiza el reconocimiento de entidades nombradas o la identificación de conceptos, cada entidad nombrada es mapeada con un concepto de terminología específico a través del componente de búsqueda del diccionario de cTAKES [28], que utiliza el UMLS como diccionario.

Refinamos nuestros parámetros de consulta de forma iterativa y buscamos las notas clínicas que contenían nuestros parámetros de búsqueda finales basados en los sinónimos en UMLS para diabetes y hemodiálisis.

Estos han sido los siguientes: (A) incluimos documentos que contenían cualquiera de los siguientes términos: diabetes, diabetes mellitus, DM; (B) incluimos documentos que contenían cualquiera de los siguientes términos: hemodialisis, dialisis renal, diálisis extracorpórea, en HD, HD hoy, HD tunelizada, HD continua, HD cont; (C) finalizamos el set de documentos para correr en cTAKES sólo incluyendo documentos que contenían al menos uno de los términos del grupo A y al menos uno de los términos del grupo B; y (D) excluimos documentos usando el anotador de detección de negación en cTAKES para detectar negaciones como: evitar, rechazar, nunca, declinar, etc. que aparecían cerca de cualquiera de los términos listados en los grupos A y B.

28.2.4 Análisis

Revisamos en forma manual todas las notas de todos los pacientes identificados por el método de extracción de datos estructurada y/o el método de PNL clínico, como aquellos potenciales de tener diagnóstico de diabetes mellitus y que han requerido hemodiálisis durante su estadía en UCI para crear una base de datos de validación que contiene los pacientes identificados positivamente en la población de pacientes de MIMIC-III. Usamos esta base de datos de validación para evaluar la precisión y la recuperación tanto del método de extracción de datos estructurados como del método de PLN clínico. Comparamos los resultados de ambos métodos en la base de validación para determinar los verdaderos positivos, falsos positivos, la recuperación y la precisión. Definimos estos parámetros usando la siguiente ecuación

Recuperación = VP/ (VP + FN), donde VP = verdaderos positivos y FN = falsos negativos; y Precisión = VP/ (VP + FP), donde FP = falsos positivos. En este caso definimos recuperación como la proporción de pacientes diabéticos que requirieron hemodiálisis en la base de validación y que fueron identificados como tal. Definimos precisión como la proporción de pacientes identificados como diabéticos y que requirieron hemodiálisis cuyos diagnósticos fueron ambos confirmados en la base de validación.

28.3 Resultados

En el método de extracción de datos estructurados usando SQL como se ilustra en la Fig 28.1, encontramos 10.494 pacientes diagnosticados con diabetes mellitus usando códigos CIE-9; 1216 pacientes que requirieron hemodiálisis usando códigos de diagnóstico y procedimiento CIE-9; y 1691 pacientes que recibieron hemodiálisis al buscar las tablas de datos estructurados usando la secuencia '%hemodial%'. La figura 28.2 muestra el número de pacientes identificados usando método de PLN: 13.816 pacientes diagnosticados con diabetes mellitus y 3.735 pacientes identificados como habiendo recibido hemodiálisis durante su estadía en UCI. En la base de datos de validación había 1879 pacientes, de los cuales 1847 (98,3%) eran pacientes diabéticos confirmados que se habían sometido a hemodiálisis. Identificamos 1032 pacientes (54,9% de 1879) cuando se utilizó sólo SQL y 1679 (89.4% de 1879) cuando se utilizó cTAKES. De ellos, 832 (44,3% de 1879) se encontraron con ambos enfoques según se muestra en la Fig 28.3.

La tabla 28.2 muestra el resultado de los 2 métodos usados para identificar cohortes de pacientes comparados con la base de validación.

El método de PLN clínico tuvo mejor precisión comparado con el método de extracción de datos estructurados. El método de PLN clínico también identificó menos FP (0,8% de 1679) comparados con el método de extracción de datos estructurados (1,8% de 1032).

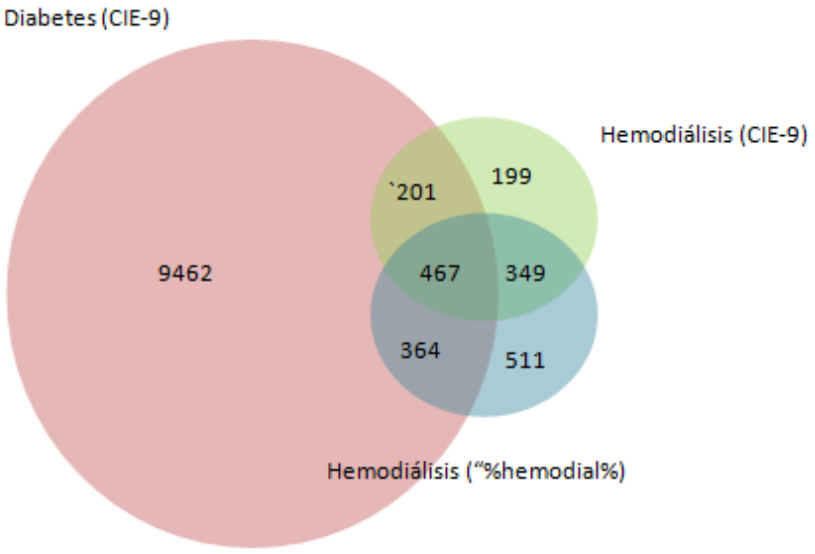

Fig. 28.1 Pacientes identificados por extracción de datos estructurados, en el sentido de las agujas del reloj desde la izquierda, diagnosticados con diabetes mellitus utilizando CIE-9, recibieron hemodiálisis

usando códigos diagnósticos y de procedimientos de alta CIE-9 y recibieron hemodiálisis usando la secuencia'%hemodial%'

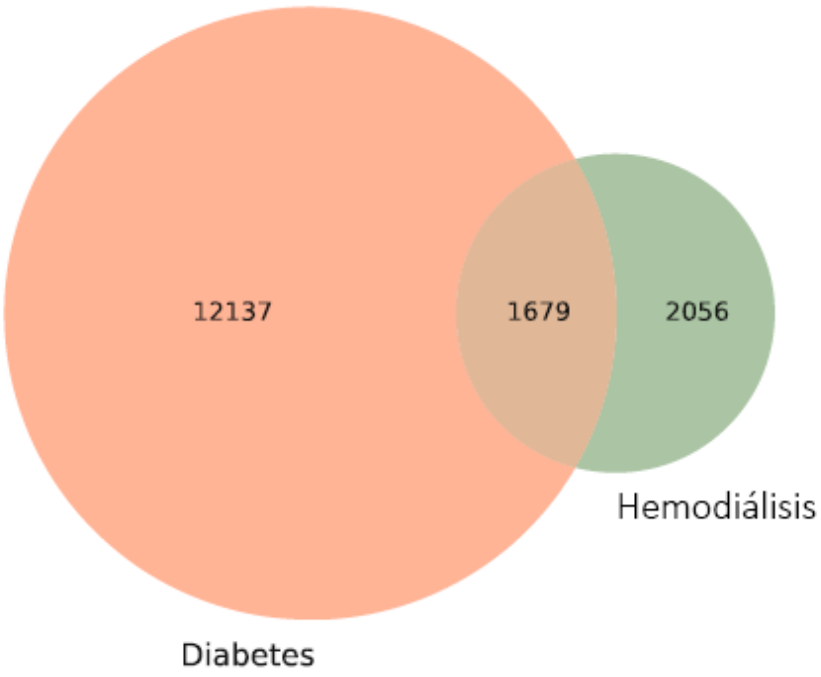

Fig. 28.2 Pacientes identificados por el método de PLN, desde la izquierda, diagnosticados con diabetes, diagnosticados con diabetes y que recibieron hemodiálisis, y aquellos que recibieron hemodiálisis

En este caso de estudio, el valor de recuperación no pudo ser calculado. Pero dado que la recuperación se calcula dividiendo VP por la suma de VP y FN, y el denominador de ambos métodos es el mismo, podemos usar la cuenta de VP como proxy para determinar qué método muestra mayor recuperación.

Basado en los resultados, encontramos que se identificaron más VPs usando PLN comparado con el enfoque de datos estructurados.

Por lo tanto, el método de PLN clínico tuvo una recuperación más alta que el método de datos estructurados.

También analizamos las notas clínicas de 19 pacientes identificados como FP usando el método de extracción de datos estructurados. Encontramos 14 pacientes que fueron identificados en forma incorrecta como diabéticos, 3 pacientes fueron identificados incorrectamente como haber recibido hemodiálisis y 2 pacientes no eran diabéticos ni requirieron hemodiálisis durante su estadía en la UCI. En los 13 pacientes identificados como FP al

usar el método de PLN, también analizamos las notas clínicas y encontramos que 5 no recibieron hemodiálisis durante su estadía en UCI, 2 inicialmente tuvieron hemodiálisis pero suspendieron por haber presentado complicaciones y 6 no tenían diabetes (3 no tenían historia de diabetes, 1 inicialmente se presumió que tenía diabetes según relato de la familia del paciente pero no era real, 1 tenía diabetes gestacional sin historia previa de diabetes y 1 recibió insulina en varias oportunidades durante su estadía en la UCI pero no tenía diagnóstico previo de diabetes ni tampoco diagnóstico de diabetes de reciente registrado en ninguna de las notas).

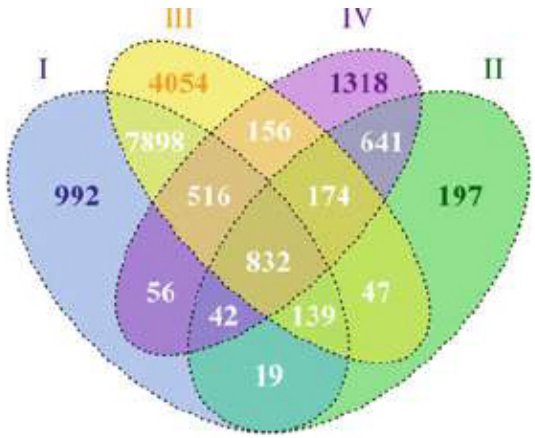

Fig. 28.3 Pacientes identificados por los métodos de extracción de datos estructurados y de PLN clínicos: I-pacientes con diabetes encontrados usando SQL; II-pacientes que requirieron hemodiálisis usando SQL; III– pacientes diabéticos usando cTAKES y; IV-pacientes que recibieron hemodiálisis encontrados usando cTAKES

Tabla 28.2 Precisión de identificar cohortes de pacientes usando extracción de datos estructurados y PLN comparados con la validación de la base de datos

Base de datos de validación (n = 1879)	**Método de extracción de datos estructurados, positivo (n = 1032)**	**Método de PLN Clinico, positivo (n = 1679)**
Positivo	VP=1013	VP=1666
Negativo	FP=19	FP=13
Precisión	98.2%	99.2%

28.4 Discusión

Tanto el método de extracción de datos estructurados como el método de PLN clínica alcanzaron alta precisión para identificar pacientes diabéticos que recibieron hemodiálisis durante su estadía en UCI. No obstante, el método de PLN clínico mostró mejor precisión y mayor recuperación en una forma más eficiente y en menor tiempo en comparación con la técnica de extracción de datos estructurados. Identificamos distintas variables que pueden causar una menor precisión cuando se usa sólo SQL para identificar cohortes de pacientes como el tipo de enfermedad y tipo de intervención, la presencia de otras condiciones similares a la diabetes (por ej, diabetes insípida, diabetes gestacional), y la presencia de otras intervenciones similares a la hemodiálisis (por ej., diálisis peritoneal, terapia de reemplazo renal continuo). La característica temporal de la intervención también sumo complejidad al proceso de identificación de la cohorte. Extraer y usar los sinónimos UMLS para “diabetes mellitus” y “hemodiálisis” al realizar PLN en las notas clínicas ayudó a aumentar el número de pacientes incluidos en la cohorte final. El saber que los médicos usan frecuentemente acrónimos, como”DM” para referirse a la diabetes mellitus y “HD”para hemodiálisis y abreviaturas como “cont” para la palabra ‘continua’ al tomar notas clínicas, nos ayudó a refinar nuestros parámetros de consulta finales.

Hay varias limitaciones en este caso de estudio. La especificidad no pudo calcularse porque para determinar los VP y FN necesitaríamos validar manualmente toda la base de datos MIMIC-III. Aunque se puede argumentar que los casos perdidos en la base de validación por cualquiera de los métodos podrían considerarse como FN, estos podrían no ser el recuento real de FN en MIMIC-III porque no se han incluído los que se pudieron encontrar fuera de la base de datos de validación.

Además, considerando que la base de datos de validación usada no fue independiente de ambos métodos, es posible que se haya sobrestimado el recuento de VP y de FP, así como la precisión y la recuperación. Otra limitación es la falta de una base de datos *gold standard* para la cohorte de pacientes específica que investigamos. Sin ella, no pudimos evaluar completamente los métodos de identificación de cohortes que implementamos. La creación de una base de datos *gold standard*, validada por médicos clínicos y que incluya pacientes en la base de datos MIMIC-III que han sido correctamente identificados como VN y FN para esta cohorte

particular de pacientes, ayudará a evaluar mejor el desempeño de los métodos usados en este caso de estudio. Disponer de una base de datos *gold standard* también ayudará a calcular la especificidad para ambos métodos.

Otra limitación es que nos enfocamos en diagnósticos y procedimientos de alta, especialmente en el método de extracción de datos estructurados. Otras fuentes de datos en MIMIC-III, como resultados de laboratorio y registro de medicamentos, podrían ayudar a sustentar los hallazgos o hasta aumentar el número de pacientes identificados al usar SQL.

Además, a pesar de que usamos una base de datos de gran tamaño, nuestros datos procedían de una única fuente. Comparar nuestros resultados encontrados usando MIMIC-III con otras bases de datos públicas disponibles que contengan datos de HCEs podría ayudar a evaluar la generalizabilidad de nuestros resultados.

28.5 Conclusiones

El método de PLN es una forma eficiente de identificar cohortes de pacientes en grandes bases de datos clínicas y produce mejores resultados cuando se compara con la extracción de datos estructurados. La combinación de uso de sinónimos de UMLS y anotadores de detección de negación en una herramienta de PLN clínico puede ayudar a los investigadores clínicos a realizar mejor las tareas de identificación de cohortes de pacientes utilizando datos de múltiples fuentes dentro de una gran base de datos clínica.

Direcciones futuras

Estudiar cómo los investigadores clínicos podrían aprovechar el PLN en la minería de datos clínicos podría ser beneficioso para la comunidad científica. En este caso de estudio, encontramos que el uso de PLN alcanza mejores resultados para las tareas de identificación de pacientes comparado con la extracción de datos estructurados.

Utilizar PLN, potencialmente podría ser útil para otras tareas de investigación clínica que consumen tiempo e involucran datos recolectados de HCE en departamentos de consultas externas, salas de internación, departamentos de emergencias, laboratorios y otras diversas fuentes de datos médicos. La detección automática de hallazgos anormales

mencionados en los resultados de test diagnósticos como Rayos X y electrocardiogramas podrían ser usados sistemáticamente para mejorar la calidad de las grandes bases de datos. Los análisis de series temporales también podrían mejorarse usando PLN para extraer más información de las notas clínicas de texto libre.

Notas

1. cTAKES se encuentra disponible en el sitio web Ctakes Apache: http://ctakes.apache.org/downloads.cgi.
Se puede encontrar una descripción de los componentes decTAKES 3.2 en la página wiki de cTAKES: https://cwiki.apache.org/confluence/display/CTAKES/cTAKES+3.2+Component+Use+Guide [28].

Nota: Las imágenes de este capítulo se encuentran disponibles a color en el ebook; disponible en www.hardineros.com

Apéndice: Código

Todas las consultas SQL para contar el número de pacientes por cohorte así como el archivo de configuracióncTAKES XML usado para analizar las notas están disponibles en el repositorio de GitHub de este libro: https://github.com/MIT-LCP/critical-data-book.

Más información del código se encuentra disponible en este sitio web.

Se usaron los siguientes scripts

- *cohort_diabetic_hemodialysis_icd9_based_count.sql*: Número total de pacientes diabéticos que realizaron hemodiálisis basado en los códigos diagnósticos
- *cohort_diabetic_hemodialysis_notes_based_count.sql*: Lista de pacientes diabéticos que realizaron hemodiálisis basado en las notas clínicas no estructuradas
- *cohort_diabetic_hemodialysis_proc_and_notes_based_count.sql*: Número total de pacientes diabéticos que realizaron hemodiálisis basado en notas clínicas no estructuradas y códigos de procedimientos.
- *cohort_diabetic_hemodialysis_proc_based_count.sql:* Número total de pacientes diabéticos que realizaron hemodiálisis basado en los códigos de procedimientos
- *cohort_diabetic_icd9_based_count_a.sql:* Lista de pacientes diabéticos basado en los códigos CIE-9
- *cohort_hemodialysis_icd9_based_count-b.sql:* Lista de pacientes que realizaron hemodiálisis basado en los códigos CIE.9
- *cohort_hemodialysis_proc_based_count_c.sql*: Lista el número de pacientes que realizaron hemodiálisis basado en la etiqueta del procedimiento
- *CPE_physician_notes.xml*: archivo de configuración cTAKES XML para procesar notas de pacientes. Algunas rutas necesitan ser adaptadas a la configuración del desarrollador

Referencias

1. Kury FSP, Huser V, Cimino JJ (2015) Reproducing a prospective clinical study as a computational retrospective study in MIMIC-II. In: AMIA Annual Symposium Proceedings, pp 804-813.
2. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G (2014) Big data in healthcare: using analytics to identify and manage high-risk and high-cost patients. Health Aff (Millwood) 33 (7): 1123-1131.
3. Segal JB, Powe NR (2004) Accuracy of identification of patients with immune thrombocytopenic purpura through administrative records: a data validation study. Am J Hematol 75 (1): 12-17.
4. Eichler AF, Lamont EB (2009) Utility of administrative claims data for the study of brain metastases: a validation study. J Neuro-Oncol 95 (3): 427-431.
5. Kern EF, Maney M, Miller DR, Tseng CL, Tiwari A, Rajan M, Aron D, Pogach L (2006) Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease indiabetes. Health Serv Res 41 (2): 564-580.

6. Zhan C, Eixhauser A, Richards CL Jr, Wang Y, Baine WB, Pineau M, Verzier N, Kilman R, Hunt D (2009) Identification of hospital-acquired catheter-associated urinary tract infections from Medicare claims: sensitivity and positive predictive value. MedCare 47 (3): 364-369.
7. Floyd JS, Heckbert SR, Weiss NS, Carell DS, Psaty BM (2012) Use of administrative data to estimate theincidence of statin-related rhabdomyolysis. J Am Med Assoc 307 (15): 1580-1582CodeAppendix 415.
8. van Walraven C, Austin PC, Manuel D, Knoll G, Jennings A, Forster AJ (2010) The usefulness of administrative databases for identifying disease cohorts is increased with a multivariatemodel. J Clin Epidemiol 63 (12): 1332-1341.
9. Tieder JS, Hall M, Auger KA, Hain PD, Jerardi KE, Myers AL, Rahman SS, Williams DJ, Shah SS (2011) Accuracy of administrative billing codes to detect urinary tract infection hospitalizations. Pediatrics 128:323-330.
10. Rosen LM, Liu T, Merchant RC (2012) Efficiency of International Classification of Diseases, Ninth Revisión, billing code searches to identify emergency department visits for blood andbody fluid exposures through a state wide multicenter database. Infect Control HospEpidemiol 33:581-588.
11. Lamont EB, Lan L (2014) Sensitivity of Medicare claims data formeasuring use of standard multiagent chemotherapy regimens. MedCare 52 (3): e15-e20.
12. Bache R, Miles S, Taweel A (2013) An adaptable architecture for patient cohort identification from diverse data sources. J Am Med Inform Assoc 20 (e2): e327-e333.
13. Sada Y, Hou J, Richardson P, El-Serag H, Davila J (2013) Validation of case finding algorithms for hepatocellular cancer from administrative data and electronic health records using natural language processing. MedCare.
14. Abhyankar S, Demner-Fushman D, Callaghan FM, McDonald CJ (2014) Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. J Am Med Inform Assoc 21 (5): 801-807.
15. Jurafsky D, Martin H (2008) Speech and language processing, 2nd edn. Prentice Hall, Englewood Cliffs, NJ.
16. Voorhees EM, Tong RM (2011) Overview of the TREC 2011 medical records track. In: The twentieth text retrieval conference proceedings (TREC 2011). National Institute for Standards and Technology, Gaithersburg, MD.
17. Wilbur WJ, Rzhetsky A, Shatkay H (2006) New directions in biomedical text annotation: definitions, guidelines and corpus construction. BMC Bioinform 7:356.
18. Buchan NS, Rajpal DK, Webster Y, Alatorre C, Gudivada RC, Zheng C, Sanseau P, Koehler J (2011) The role of translational bioinformatics in drug discovery. Drug Discov Today 16:426-434.
19. Nadkarni PM, Ohno-Machado L, Chapman WW (2011) Natural language processing: an introduction. J Am Med Inform Assoc 18:544-551.

20. Uzuner Ö, South BR, Shen S, Duvall SL (2011) 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 18 (5): 552-556.
21. Danforth KN, Early MI, Ngan S, Kosco AE, Zheng C, Gould MK (2012) Automated identification of patients with pulmonary nodules in an integrated health system using administrative health plan data, radiology reports, and natural language processing. J Thorac Oncol 7:1257-1262.
22. Thomas AA, Zheng C, Jung H, Chang A, Kim B, Gelfond J, Slezak J, Porter K, Jacobsen SJ, Chien GW (2014) Extracting data from electronic medical records: validation of a natural language processing program to assess prostate biopsy results. World J Urol 32 (1): 99-103.
23. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman LW, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG (2011) Multiparameter intelligent monitoring in intensive care II: a public-access intensive care unit database. Crit Care Med 39 (5): 952-960.
24. Neamatullah I, Douglass MM, Lehman LW, Reisner A, Villarroel M, Long WJ, Szolovits P, Moody GB, Mark RG, Clifford GD (2008) Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 8:32.
25. Standards for Privacy of Individually Identifiable Health Information; Final Rule, 45 CFR Parts 160 and 164 (2002) Disponible en http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/privruletxt. txt. [Consultado 6 Octubre 2015].
26. MIMIC. Disponible en https://mimic.physionet.org/gettingstarted/access. [Consultado 19 Febrero 2016].
27. The Web's Free 2015 Medical Coding Reference. Disponible en http://www.icd9data.com. [Consultado 7 Octubre 2015].
28. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG (2010) Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 17 (5): 507-513.
29. Apache cTAKESTM. http://cTAKES. apache.org/index.html. [Consultado 3 Oct 2015].
30. Lindberg DA, Humphreys BL, McCray AT (1993) The unified medical language system. Meth Inf Med 32 (4): 281-291.
31. Unified Medical Language System® (UMLS®) The Metathesaurus. Disponible en https://www.nlm.nih.gov/research/umls/new-users/online-learning/Meta-001. html. [Consultado 7 Octubre 2015].
32. Griffon N, Chebil W, Rollin L, Kerdelhue G, Thirion B, Gehanno JF, Darmoni SJ (2012) Performance evaluation of unified medical language system®'s synonyms expansión to query PubMed. BMC Med Inform Decis Mak 12:12.
33. cTAKES 3.2 Component Use Guide. Disponible en https://cwiki.apache.org/confluence/display/CTAKEScTAKES+3.2+Component+Use+Guide. [Consultado 7 Octubre 2015] J Am Med Inform Assoc 21 (5): 801-807.

Capítulo 29

SELECCIÓN DE HIPERPARÁMETROS

Franck Dernoncourt, Shamim Nemati,
Elias Baedorf Kassis y Mohammad Mahdi Ghassemi

Objetivos de Aprendizaje

Nivel alto:

Aprender cómo elegir los hiperparámetros óptimos en un ensamblaje de aprendizaje automático para predicción médica.

Nivel bajo:

1. Aprender la intuición que subyace a la optimización Bayesiana
2. Entender el algoritmo genético y el algoritmo de búsqueda dispersa con arranque múltiple
3. Aprender el atributo de entropía multiescala

29.1 Introducción

El uso de algoritmos y atributos para analizar datos médicos con el objetivo de predecir una enfermedad o un resultado comúnmente implica la elección de hiperparámetros. A grandes rasgos, un hiperparámetro puede ser definido como un parámetro que no es ajustado durante la fase de aprendizaje que optimiza la función del objetivo principal en el set de entrenamiento. Mientras que una simple red de búsqueda produciría los hiperparámetros óptimos probando todas las combinaciones posibles de hiperparámetros, éste no escala a medida que aumenta el número de hiperparámetros y del tamaño del set de datos. Como resultado, los investigadores típicamente eligen los hiperparámetros en forma arbitraria, después de una serie de ensayos manuales, los cuales pueden arrojar dudas sobre los resultados, dado que los investigadores podrían haberse tentado a ajustar los parámetros específicamente para el set de prueba. En este capítulo, presentamos tres técnicas con base matemática para optimizar hiperparámetros en forma automática: la optimización Bayesiana, los algoritmos genéticos y la búsqueda dispersa con arranque múltiple.

Para ejemplificar el uso de estos métodos de selección de hiperparámetros, nos enfocamos en la predicción de la mortalidad

hospitalaria en pacientes en la unidad de cuidados intensivos (UCI) con sepsis severa. El resultado que consideramos es binario: el paciente murió en el hospital o sobrevivió. Los pacientes con sepsis tienen alto riesgo de mortalidad (aproximadamente 30% [1]), y la capacidad de predecir los resultados es de gran interés clínico. El puntaje APACHE [2] se utiliza habitualmente para la predicción de mortalidad, pero tiene significativas limitaciones en términos de uso clínico, dado que en general no logra predecir en forma precisa los resultados de pacientes individuales, y no toma en cuenta mediciones fisiológicas dinámicas. Para remediar este inconveniente, investigamos el uso de la entropía multiescala (EMS) [3, 4] aplicado a la señal de la frecuencia cardiaca (FC) como un predictor de resultados: la EMS mide la complejidad de series de tiempo de duración finita. Para calcular la EMS es necesario especificar un set de parámetros, específicamente el factor de escala máximo, la diferencia entre factores de escala consecutivos, la longitud de las secuencias a comparar y el umbral de similitud. Mostramos que usando métodos de selección de hiperparámetros, la EMS puede predecir el resultado de cada paciente con mayor precisión que el puntaje APACHE.

29.2 Set de datos de estudio

Usamos la base de datos Medical Information Mart for Intensive Care II (MIMIC II) la cual está disponible online gratuitamente y fue introducida por [5, 6]. La base MIMIC II está dividida en dos sets de datos diferentes:

- la Base de Datos Clínicos, que es una base de datos relacional que contiene información estructurada como datos demográficos del paciente, fechas de admisión y egreso hospitalarios, seguimiento de lugar de internación, fecha de muerte, medicaciones, pruebas de laboratorio y notas del personal médico.
- la Base de Datos de Señales, que es un set de archivos planos que contienen hasta 22 tipos diferentes de señales para cada paciente, incluyendo señales de electrocardiogramas (ECG).

Seleccionamos pacientes que presentaban sepsis severa, definidos como pacientes con una infección identificada con evidencia de disfunción orgánica e hipotensión con necesidad de vasopresores y/o resucitación con fluidos [7]. Refinamos más la cohorte de pacientes eligiendo aquellos que

tuvieran señales de ECG completas en sus primeras 24hs en UCI. Para cada paciente, extrajimos el resultado binario (es decir, si fallecieron en el hospital) de la base de datos clínica. Las señales de FC fueron extraídas de las señales de ECG, y se eliminaron los pacientes con señales baja calidad de FC.

29.3 Métodos

Comparamos el poder predictivo de los siguientes tres sets de atributos para predecir los resultados de los pacientes: estadísticas descriptivas básicas en la serie de tiempo (media y desviación estándar), puntaje APACHE IV y EMS. Dado que estos atributos se calculan en series de tiempo, para cada set de atributos obtuvimos un vector de atributos de las series de tiempo. Una vez que se calcularon estos atributos agrupamos pacientes basándonos en estos vectores usando agrupamiento espectral. El número de grupos (*clusters*) se determinó usando los valores silueta [8]. Esto nos permitió identificar la alta heterogeneidad de los datos resultante del hecho de que los pacientes de MIMIC provenían de diferentes unidades de cuidado. Por último, para cada *cluster*, entrenamos una máquina de vector de soporte de clasificación (MVS). Para clasificar un nuevo paciente, calculamos la distancia a cada centro de *cluster*, y calculamos el resultado de cada MVS de clasificación: para tomar la decisión final en el resultado predicho, calculamos un promedio ponderado del resultado de cada MVS de clasificación, donde los pesos eran la distancia a cada centro de *cluster*. Este método de combinación de agrupamiento (*clustering)* con MVS es llamado MVS transductivo. Usamos el área bajo la curva ROC (AUROC, frecuentemente mencionada como AUC) como la medida de desempeño para la clasificación. La Figura 29.1 ilustra el funcionamiento de las MVS transductivos.

La EMS puede ser entendida como el set de valores de entropía de la muestra para una señal que se promedia sobre varias longitudes de segmento crecientes. La EMS, *y,* se calculó como se muestra a continuación:

$$y_j^{\tau} = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} x_i$$

Dónde:

- x_i es el valor de señal en la muestra *I,*
- *j* es el índice de la ventana a calcular,

- τ es el factor de escala,
- Y es la longitud de las secuencias a comparar,
- Z es el umbral de similitud.

Adicionalmente, tenemos los siguientes parámetros:

- El factor de escala máximo,
- El aumento de escala que es la diferencia entre factores de escala consecutivos,
- Los criterios o umbrales de similitud, nombrados r

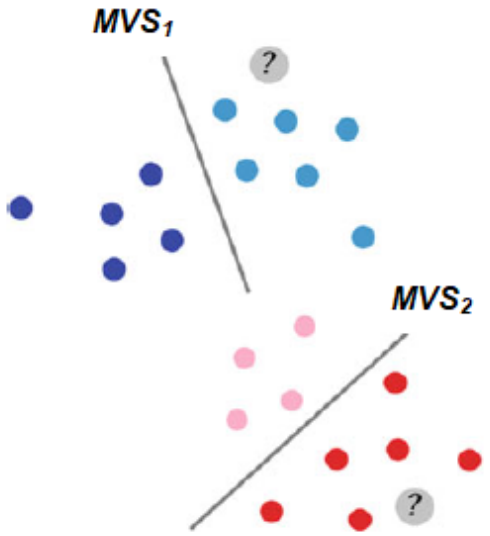

Fig. 29.1 MVS transductivo: en primer lugar se realiza el agrupamiento, luego se utiliza una combinación convexa de los resultados de MVS para obtener la predicción final de probabilidad.

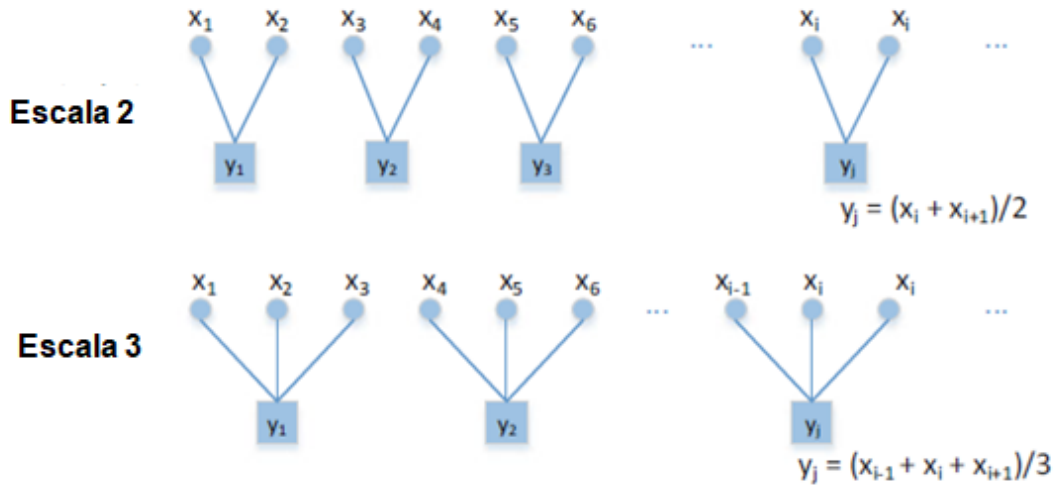

Fig. 29.2 Ilustración de varias escalas de Costa y col. Sólo se muestran las escalas 2 y 3. x_j es el valor de señal en la muestra *i*.

La Figura 29.2 muestra cómo se calcula "*y*" para diferentes escalas. Para seleccionar los mejores hiperparámetros para EMS, comparamos tres técnicas de optimización de hiperparámetros: la optimización Bayesiana, algoritmos genéticos, y búsqueda dispersa con múltiple arranque.

La optimización Bayesiana construye la distribución P (y_{test} | y_{train}, x_{train}, x_{test}), donde x_{train} es el set de parámetros de EMS que fueron usados para obtener el AUROC de $y_{train,}$ x_{test} es un nuevo set de parámetros de EMS, e y_{test} es el AUROC que se obtendría usando los nuevos parámetros de EMS. Para decirlo de otra forma, basada en las observaciones previas de los parámetros de EMS y las AUROCs obtenidas, la optimización Bayesiana predice qué AUROC arrojará un nuevo set de parámetros de EMS. Cada vez que se calcula una nueva AUROC, tanto el set de parámetros EMS como la AUROC se agregan a x_{test} e y_{test}. En cada iteración, podemos explorar, es decir calcular y_{test} para el cual la distribución *p* tiene una alta variabilidad, o aprovechar, es decir calcular y_{test} para el cual la distribución tiene una baja varianza y una alta expectativa. En [9] puede encontrarse una implementación.

Un algoritmo genético es una optimización de un algoritmo basado en el principio de selección natural Darwiniano. Una población está compuesta de sets de parámetros EMS. Cada set de parámetros EMS es evaluado basado en el AUROC que logró y se eliminan los sets de parámetros de EMS con bajas AUROCs. Los sets de parámetros EMS sobrevivientes son mutados, es

decir cada parámetro es apenas modificado, para crear nuevos sets de parámetros EMS, los cuales forman una nueva población. Iterando a través de este proceso, los nuevos sets de parámetros EMS incluyen AUROCs cada vez más altas. Configuramos un tamaño de población de 100 y corrimos la optimización por 30 minutos. La primera población fue tomada de manera aleatoria.

La búsqueda dispersa con arranque múltiple es similar al algoritmo genético; la única diferencia reside en el uso de un proceso determinista como el descenso de gradiente para identificar los individuos de la siguiente población.

La Figura 29.3 resume el uso de información de aprendizaje automático presentado en esta sección.

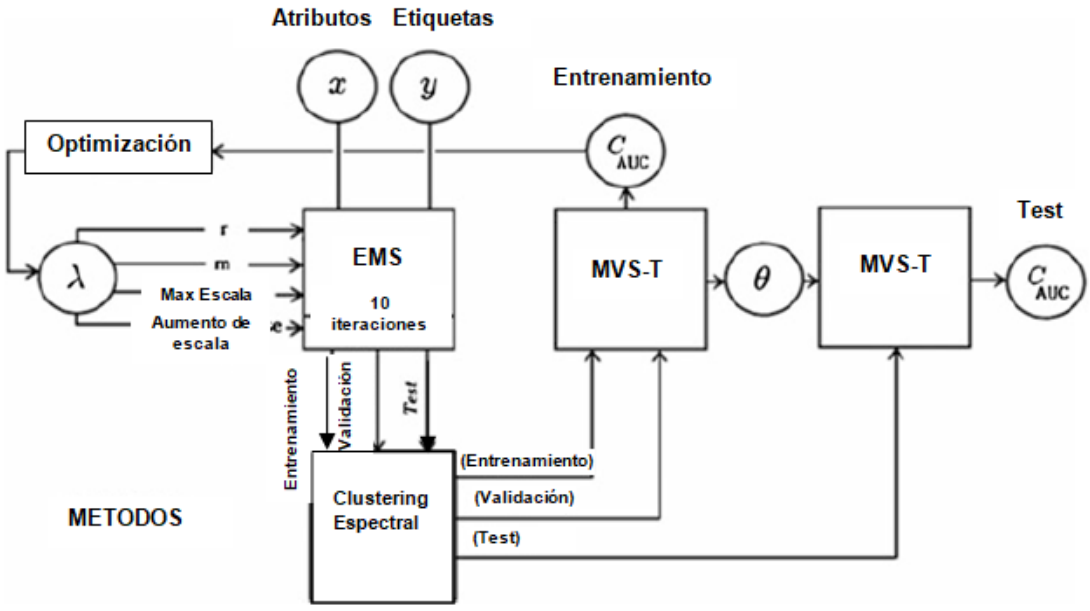

Fig. 29.3 El ensamblaje de aprendizaje automático completo. Los atributosde EMS son calculadas desde el *input x* usando los parámetros *r, m*, escala max y aumento de escala. Se crean 10 iteraciones.

El set de datos se dividió en sets de prueba (20%), validación (20%) y entrenamiento (60%). Para asegurar la robustez del resultado, usamos validación cruzada de 10 iteraciones y el promedio del AUROC de las 10 iteraciones. Para hacer justa la comparación, cada técnica de optimización de hiperparámetros fue corrida durante la misma cantidad de tiempo, esto es 30 minutos.

29.4 Análisis

La Tabla 29.1 contiene los resultados de los tres sets de atributos que consideramos. Para los atributos de la EMS, la Tabla 29.1 presenta los resultados logrados manteniendo los hiperparámetros predeterminados, u optimizándolos usando una de las tres técnicas de optimización de hiperparámetros que presentamos en la sección previa.

El primer set de atributos, específicamente las estadísticas descriptivas básicas (media y desviación estándar), arroja un AUROC de 0.54 en el set de prueba, que es muy bajo dado que un clasificador aleatorio produce un AUROC de 0.50. El segundo set de atributos, APACHE IV, alcanza un AUROC mucho más alto, 0.68, lo cual no es sorprendente dado que APACHE IV fue diseñado para estimar la mortalidad hospitalaria en pacientes críticamente enfermos. El tercer set de atributos basado en EMS actúa sorprendentemente bien con los valores predeterminados (AUROC de 0.66), e incluso mejor cuando se optimiza con cualquiera de las tres técnicas de optimización de hiperparámetros. La optimización Bayesiana logra el AUROC más alta, 0.72.

Tabla 29.1 Comparación de los atributos APACHE, media y desviación estándar de las series de tiempo, y el atributo EMS con parámetros predeterminados u optimizados con la optimización Bayesiana, algoritmos genéticos y búsqueda dispersa de múltiple arranque, para la predicción del resultado del paciente.

	Max Escala	Aumento de Escala	r	m	AUROC (entrenamiento)	AUROC (testing)
Series de tiempo: promedio y desviación estandar					0.56 (0.52–0.56)	0.54 (0.45–0.60)
APACHE IV					0.77 (0.75–0.79)	0.68 (0.55–0.77)
EMS (predeterminado)	20	1	0.15	2	0.77 (0.73–0.78)	0.66 (0.60–0.72)
EMS (Bayesiana)	17.62 (8.68)	2.59 (0.93)	0.11 (0.07)	2.58 (0.85)	0.77 (0.69–0.79)	0.72 (0.63–0.78)
EMS (genético)	23.54 (14.34)	2.56 (1.12)	0.18 (0.15)	2.07 (0.70)	0.77 (0.67–0.84)	0.67 (0.44–0.78)
EMS (multiarranque)	19.03 (12.57)	2.35 (0.87)	0.18 (0.128)	2.53 (0.87)	0.73 (0.69–0.76)	0.69 (0.53–0.72)

Para cada parámetro EMS reportamos media y desviación estándar de cada iteración (con la desviación estándar entre paréntesis). Para el AUROC reportado, mostramos el percentilo 50 en la mitad superior de la celda y los percentilos 25 y 75 en la mitad inferior.

Para cada parámetro EMS reportamos media y desviación estándar de cada iteración (con la desviación estándar entre paréntesis). Para el AUROC reportado, mostramos el percentilo 50 en la mitad superior de la celda y los percentilos 25 y 75 en la mitad inferior.

29.5 Visualizaciones

La Figura 29.4 brinda una mirada dentro de los parámetros EMS seleccionados por las tres técnicas de selección de hiperparámetros sobre la validación cruzada de 10 iteraciones. Cada punto representa el valor de un parámetro optimizado por una técnica de selección de hiperparámetros dada para una única iteración de datos. Para todos los 4 parámetros EMS, observamos una gran variabilidad: esto indica que no hay un óptimo global claro, en lugar de eso existen muchos sets de parámetros EMS que producen una alta AUROC.

Llamativamente, en este experimento la optimización Bayesiana es más robusta a la variabilidad de parámetros, como se muestra en los intervalos de confianza alrededor de las AUROCs: la mayoría de las AUROCs alcanzadas por la optimización Bayesiana son altas, a diferencia de los algoritmos genéticos y la búsqueda dispersa de múltiple arranque. Las dos últimas técnicas son susceptibles a presentar convergencia prematura, mientras que la optimización Bayesiana tiene mejor negociación “exploración-explotación”.

También notamos que la máxima escala y los valores *r* alcanzados por la optimización Bayesiana tienen una menor variabilidad que los algoritmos genéticos y la búsqueda dispersa de múltiple arranque. Se podría plantear la hipótesis de que la heterogeneidad entre pacientes se podría reflejar más en el aumento en escala y los parámetros *m* de EMS que en la escala máxima y los parámetros *r*.

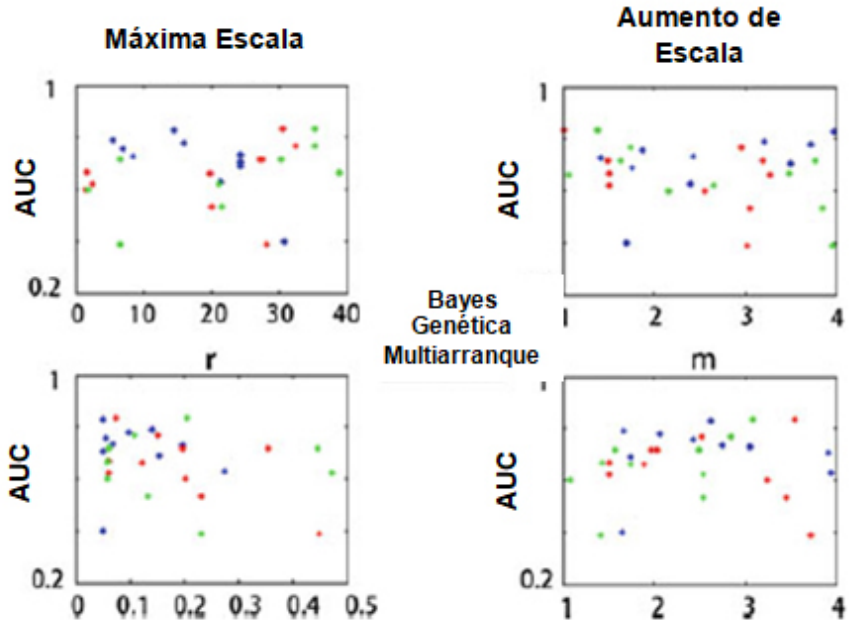

Fig. 29.4 El impacto de los parámetros EMS en el AUROC de la predicción del resultado.

29.6 Conclusiones

Los resultados de este estudio de caso demuestran dos puntos principales. El primero, desde un punto de vista médico, subraya los posibles beneficios de utilizar medidas fisiológicas dinámicas en la predicción de resultados para pacientes en UCI con sepsis severa: los datos del estudio efectivamente sugieren que utilizar estas variables fisiológicas dinámicas a través de EMS con hiperparámetros optimizados logra una mejor predicción de la mortalidad comparado con el score APACHE IV. Es necesario contar con un muestreo de señales fisiológicas de alta frecuencia para que los atributos de la EMS sean significativos, resaltando la necesidad de recolección de datos de alta resolución, en oposición a algunos métodos existentes de recolección de datos donde las señales se agregan a nivel de segundo o minuto, si no más, antes de ser grabadas.

Segundo, desde un punto de vista metodológico, los resultados observados son un argumento fuerte para el uso de técnicas de selección de hiperparámetros. No es sorprendente que los resultados obtenidos con los atributos de EMS sean altamente dependientes de los hiperparámetros EMS. Si no hubiéramos usado una técnica de selección de hiperparámetros y en vez de eso hubiéramos conservado los valores predeterminados, habríamos concluido que APACHE IV provee una mejor mirada predictiva que EMS, y por lo tanto nos hubiéramos perdido la importancia de las variables fisiológicas dinámicas para la predicción del resultado del paciente. La optimización Bayesiana parece producir mejores resultados que los algoritmos genéticos y la búsqueda dispersa de múltiple arranque.

29.7 Discusión

Todavía hay mucho lugar para investigación. Nos enfocamos en pacientes en UCI con sepsis severa, pero también valdría la pena investigar muchas otras cohortes de pacientes críticamente enfermos. A pesar de que restringimos nuestro estudio sólo al uso de EMS y FC, sería interesante integrar y combinar otras características de las enfermedades y señales fisiológicas. Por ejemplo, [10] usó optimización Bayesiana para encontrar los parámetros de onda pequeña óptimos para predecir episodios de hipotensión aguda. Quizás la combinación de pequeñas ondas de presión arterial dinámica con EMS de la FC, e incluso con otros datos dinámicos como variación de presión de pulso, optimizarían y mejorarían aún más el modelo de predicción de mortalidad. Además existen otros puntajes para predecir la mortalidad poblacional como SOFA y SAPS II, que proveerían líneas de base útiles sumadas al APACHE [11].

La escala de nuestros experimentos fue satisfactoria para los objetivos del caso de estudio, pero otras investigaciones podrían requerir un set de datos de mayor magnitud. Esto podría llevarnos a adoptar un enfoque distribuido para implementar las técnicas de selección de hiperparámetros. Por ejemplo, [12] usó un enfoque distribuido para la optimización de hiperparámetros en 5000 pacientes y más de un billón de puntos de presión arterial. [13, 14] presentan otro sistema de gran escala para usar algoritmos genéticos para la predicción de la presión arterial.

Por último, una comparación más exhaustiva entre las técnicas de selección de hiperparámetros ayudaría a comprender por qué una técnica de

selección de hiperparámetros actúa mejor que otras para un problema de predicción particular. Especialmente, las técnicas de selección de hiperparámetros también tienen parámetros, una mejor comprensión del impacto de estos parámetros en los resultados requiere más investigación.

29.8 Conclusiones

En este capítulo presentamos los tres métodos de selección de hiperparámetros principales. Para ilustrar su uso, los aplicamos a EMS, que calculamos en señales fisiológicas. En forma más amplia, estos métodos pueden ser usados para cualquier algoritmo y atributo en que se necesite ajustar los hiperparámetros.

Los datos de UCI proveen una oportunidad única para este tipo de investigación con variables que se recolectan rutinariamente en forma continua, incluyendo ondas de ECG, ondas de presión arterial de líneas arteriales, variación de presión de pulso, oximetría de pulso, además de datos de los dispositivos de ventilación mecánica. Estas mediciones fisiológicas dinámicas podrían potencialmente ayudar a descubrir mejores resultados de mediciones y mejorar el manejo de decisiones en pacientes con síndrome de distress respiratorio agudo (SDRA), shock séptico, insuficiencia hepática o paro cardiaco, y otros pacientes extremadamente enfermos en la UCI. Fuera de la UCI, los datos fisiológicos dinámicos son recolectados rutinariamente durante las cirugías por los equipos de anestesia, en unidades cardiológicas con telemetría continua y en las unidades de cuidado neurológico con mediciones de electroencefalografía (EEG) rutinarias para pacientes con convulsiones o con alto riesgo de padecerlas. Como tal, las aplicaciones potenciales de EMS con optimización de hiperparámetros son amplias.

usuarios deberán obtener permiso del titular de la licencia para duplicar, adaptar o reproducir el material.

Nota: Las imágenes de este capítulo se encuentran disponibles a color en el ebook; disponible en www.hardineros.com

Referencias

1. Angus DC, Linde-Zwirble WT, Lidicker J, Clermont G, Carcillo J, Pinsky MR (2001) Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. Crit Care Med 29 (7): 1303-1310.
2. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L, Moody GB, Heldt T, Kyaw TH, Moody BE, Mark RG (2011) Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access ICU database. Crit Care Med 39 (5): 952-960. doi: 10.1097/CCM.0b013e31820a92c6.
3. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101 (23): e215-e220 [Circulation Electronic Pages; http://circ. ahajournals.org/cgi/content/full/101/23/e215].
4. Mayaud L, Lai PS, Clifford GD, Tarassenko L, Celi LA, Annane D (2013) Dynamic data during hypotensive episode improves mortality predictions among patients with sepsis and hypotension*. Crit Care Med 41 (4): 954-962.
5. Ng AY, Jordan MI, Weiss Y et al (2002) On spectral clustering: analysis and an algorithm. Adv Neural Inf Process Syst 2:849-856.
6. Snoek J, Larochelle H, Adams RP (2012) Practical Bayesian optimization of machine learning algorithms. Adv Neural Inf Process Syst 2951-2959.
7. Dernoncourt F, Veeramachaneni K, O'Reilly U-M (2015) Gaussian process-based feature selection for wavelet parameters: predicting acute hypotensive episodes from physiological signals. In: Proceedings of the 2015 IEEE 28th international symposium on computer-based medical systems. IEEE Computer Society.
8. Castella X et al (1995) A comparison of severity of illness scoring systems for intensive care unit patients: results of a multicenter, multinational study. Crit Care Med 23 (8): 1327-1335.
9. Dernoncourt F, Veeramachaneni K, O'Reilly U-M (2013c) BeatDB: a large-scale waveform feature repository. In: NIPS 2013, machine learning for clinical data analysis and healthcare workshop.
10. Hemberg E, Veeramachaneni K, Dernoncourt F, Wagy M, O'Reilly U-M (2013) Efficient training set use for blood pressure prediction in a large scale learning classifier system. En: Proceeding of the fifteenth annual conference companion on genetic and evolutionary computation conference companion. ACM, New York, pp 1267-1274.

11. Hemberg E, Veeramachaneni K, Dernoncourt F, Wagy M, O'Reilly U-M (2013) Imprecise selection and fitness approximation in a large-scale evolutionary rule based system for blood pressure prediction. En: Proceeding of the fifteenth annual conference companion on genetic and evolutionary computation conference companion. ACM, New York, pp 153-154.
12. Knaus WA et al (1981) APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med 9 (8): 591-597.
13. Costa M, Goldberger AL, Peng C-K (2005) Multiscale entropy analysis of biological signals. Phys Rev E 71:021906.
14. Costa M, Goldberger AL, Peng C-K (2002) Multiscale entropy analysis of physiologic time series. Phys Rev Lett 89:062102.