Med Intensiva. 2019;43(7):416–426

medicina intensiva

http://www.medintensiva.org/

REVISIÓN

Big Data Analysis y Machine Learning en medicina intensiva

A. Núñez Reiz[a,*], **M.A. Armengol de la Hoz**[b,c,d] **y M. Sánchez García**[a]

[a] *Servicio de Medicina Intensiva, Hospital Universitario Clínico San Carlos, Madrid, España*
[b] *Department of Anesthesia, Critical Care and Pain Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, Estados Unidos*
[c] *Laboratory for Computational Physiology, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, Estados Unidos*
[d] *Biomedical Engineering and Telemedicine Group, Biomedical Technology Centre CTB, ETSI Telecomunicación, Universidad Politécnica de Madrid, Madrid, España*

Recibido el 5 de septiembre de 2018; aceptado el 21 de octubre de 2018
Disponible en Internet el 24 de diciembre de 2018

PALABRAS CLAVE
Big Data Analysis;
Machine Learning;
Inteligencia artificial;
Análisis secundario de datos clínicos electrónicos

Resumen La gran cantidad de información que se procesa informáticamente en el entorno de la medicina intensiva la convierte en un campo ideal para el empleo de técnicas conocidas como Big Data Analysis (BDA) y Machine Learning (ML), que pueden permitir en el futuro mejorar nuestra capacidad de investigación clínica y dirigir de manera más precisa las terapias que proporcionamos a nuestros pacientes.

En este artículo se revisan los conceptos fundamentales sobre BDA y ML, y se estudian sus posibles aplicaciones al ámbito de la medicina intensiva, desde un punto de vista del clínico. También se plantean potenciales estrategias para sacar el máximo partido a estas tecnologías emergentes, incluyendo la aparición de un nuevo tipo de profesional sanitario encargado de actuar como enlace entre la parte clínica y la ingeniería de datos.
© 2018 Elsevier España, S.L.U. y SEMICYUC. Todos los derechos reservados.

KEYWORDS
Big Data Analysis;
Machine Learning;
Artificial intelligence;
Secondary electronic health record data analysis

Big Data Analysis and Machine Learning in Intensive Care Units

Abstract Intensive care is an ideal environment for the use of Big Data Analysis (BDA) and Machine Learning (ML), due to the huge amount of information processed and stored in electronic format in relation to such care. These tools can improve our clinical research capabilities and clinical decision making in the future.

* Autor para correspondencia.
Correo electrónico: anunezreiz@gmail.com (A. Núñez Reiz).

https://doi.org/10.1016/j.medin.2018.10.007
0210-5691/© 2018 Elsevier España, S.L.U. y SEMICYUC. Todos los derechos reservados.

The present study reviews the foundations of BDA and ML, and explores possible applications in our field from a clinical viewpoint. We also suggest potential strategies to optimize these new technologies and describe a new kind of hybrid healthcare-data science professional with a linking role between clinicians and data.
© 2018 Elsevier España, S.L.U. y SEMICYUC. All rights reserved.

Introducción

Los profesionales de la medicina intensiva vivimos sumergidos en un mar de datos. En un servicio de medicina intensiva informatizado, como el del Hospital Clínico San Carlos (con tres unidades de cuidados intensivos y aproximadamente 2.400 ingresos al año), durante un día normal de actividad en nuestra Unidad se incorporan a nuestra base de datos informatizada 1.400 nuevas unidades de información en promedio por cada paciente ingresado, lo que supone unos diez millones de unidades de información al año (Nuñez A, comunicación personal).

El desarrollo exponencial de la informática y la irrupción de computadores con gran capacidad de almacenamiento y procesamiento a un coste asequible hacen que toda esa información quede registrada y pueda ser utilizada de diversas maneras. Incluso en aquellas unidades que todavía carecen de historia clínica electrónica o aplicaciones departamentales para medicina intensiva existe la posibilidad de utilizar el potencial de los métodos computacionales de última generación para mejorar la manera en que hacemos nuestro trabajo cada día. Un ejemplo es la aplicación de técnicas de procesamiento de lenguaje natural (*Natural Language Processing* [NLP]) a un conjunto de informes almacenados en Word, PDF u otro formato similar de texto libre (no estructurado)[1].

Los datos que los intensivistas manejamos pueden provenir de diversas fuentes. Los profesionales sanitarios están añadiendo continuamente datos (estructurados o no) a la documentación del paciente. Se estima que los médicos dedican casi dos horas a la documentación por cada hora de atención directa al paciente[2]. Los médicos, las enfermeras y los técnicos los escriben como notas o informes de distintas especialidades en texto libre, o como información codificada sobre diagnósticos o procedimientos. También generamos órdenes de tratamiento y registros de administración de fármacos, y recibimos grandes cantidades de datos generados por nuestro aparataje medico: resultados de los analizadores de laboratorio, constantes vitales, datos de monitorización avanzada, parámetros del respirador, parámetros de funcionamiento de equipos complejos como bombas de perfusión, monitores de diálisis, dispositivos de ECMO (membrana de oxigenación extracorpórea), así como información en forma de imágenes, sonidos, vídeos y muchos más. Hasta algunos años esta información se perdía en la nada o como mucho en archivos de historias clínicas en papel. Hoy es posible almacenar y procesar esta información de manera automática en formato digital y extraer de ella nuevo conocimiento y orientación para mejorar el cuidado de los pacientes.

Existe un amplio espectro en cuanto a lo estructurados que están los datos que utilizamos. Es mucho más fácil hacer investigación clínica cuando los datos están en un formato estructurado, pero los profesionales sanitarios todavía no han alcanzado un consenso en cuanto a una manera unívoca de expresar cada concepto relacionado con la salud. Esto no quiere decir que no tengamos estándares. SNOMED CT[3], HL7[4], UMLS[5], DICOM[6], LOINC[7] y muchos otros son ejemplos de ello y son usados en diferentes campos sanitarios, permitiendo el tratamiento automatizado de los datos y el intercambio de información entre sistemas. Pero los estándares no lo son de manera absoluta: podemos expresar un diagnóstico en diferentes estándares, como MESH[8], UMLS, SNOMED CT, ICD-9[9], ICD-10[9], y frecuentemente es necesario hacer «traducciones» entre estándares (el término técnico es «mapeo») para trasladar información de unos sistemas a otros. Soluciones con nuevos enfoques del problema de la multiplicidad de estándares (por ejemplo, OMOP[10]) proporcionan herramientas para realizar estos mapeos de una manera sistematizada.

No solo los datos en sí son importantes, también lo es el contexto en el que están inmersos. A veces un dato no tiene valor alguno si no podemos asociarlo a más información relevante sobre el paciente o la situación clínica. Un registro de presión arterial de 90/60 mmHg se interpreta de manera muy diferente en una mujer joven sometida a cirugía plástica que en un anciano hipertenso con una hemorragia. El mismo concepto puede además plasmarse de formas distintas por distintos profesionales, en distintas situaciones y —en el caso de valores numéricos— con distintas unidades.

Nuestros datos estructurados pueden ser categóricos, valores numéricos enteros o con decimales, fechas, horas, duraciones, o bien pueden agruparse (por ejemplo, la presión arterial con sus dos componentes, sistólico y diastólico) de una manera similar a las listas, tuplas u objetos que vemos como estructuras de datos en lenguajes de programación. A veces nuestros datos son matrices de bytes organizados para representar una imagen o un vídeo, como por ejemplo un fichero DICOM que puede utilizarse para mostrar una imagen de CT. En este ejemplo concreto aparece otro concepto interesante, que es el de metadatos: información acerca de los propios datos (por ejemplo, en el caso de la imagen del CT, metadatos serían los parámetros de adquisición).

Algunos datos son subjetivos. Por ejemplo, un médico puede expresar el razonamiento que le llevó a un determinado diagnóstico mediante una serie de frases en una nota de texto libre. Otros datos son objetivos: nuestro hemofiltro ha extraído exactamente 52 mL de ultrafiltrado al paciente

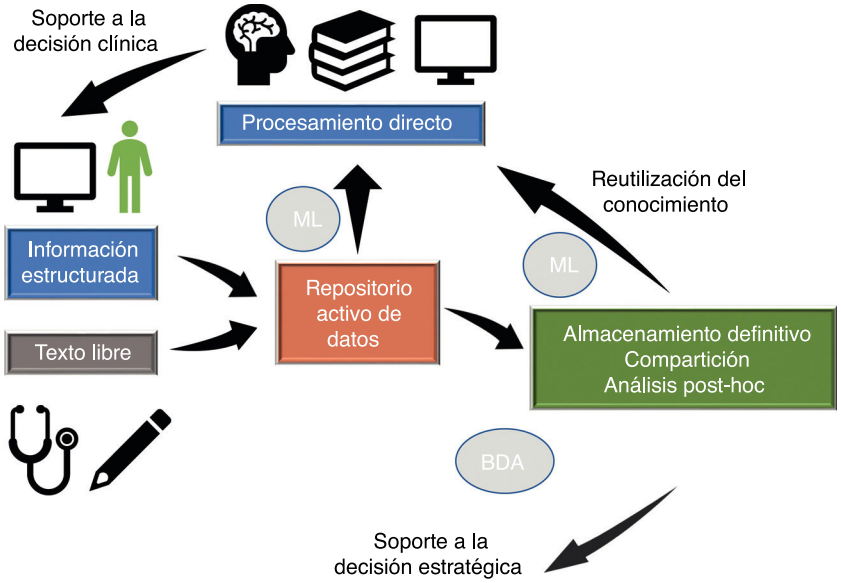

Figura 1 Flujo de datos en medicina intensiva. La calidad de la asistencia y de la toma de decisiones estratégicas depende del procesamiento de la información presente en el repositorio activo de datos (información que se va generando en cada momento y que está disponible para tomar decisiones) y de la reutilización de conocimiento previo almacenado en bases de datos locales o compartidas. El procesamiento automático de la información mediante Machine Learning (ML) y la posibilidad de acceso rápido a grandes cantidades de datos de estructura heterogénea mediante Big Data Analysis (BDA) permite mejorar los procesos asistenciales y extraer conocimiento de los datos.
Adaptada de Celi et al.[12], con permiso de los autores.

durante la última hora, y pasa esta información a la gráfica electrónica a través de la conexión de red.

Para llevar a cabo análisis secundarios de datos clínicos, son diversos los aspectos que han de tenerse en cuenta en cuanto a privacidad, seguridad y sensibilidad. Para un paciente, el que se conozca públicamente su frecuencia respiratoria puede ser perfectamente tolerable, pero no suele opinar lo mismo sobre su seropositividad al virus de la inmunodeficiencia humana.

Los datos vienen de distintos orígenes porque se obtienen por distintos motivos (contabilidad, investigación o manejo clínico del paciente, por ejemplo, son algunas de las opciones), y por tanto hay distintos actores implicados en el proceso (gestores, clínicos, investigadores, proveedores de servicios).

Si queremos ser capaces de extraer conocimiento válido de toda esta mezcolanza, para poder ayudar al profesional sanitario a tomar mejores decisiones y para orientar nuestra investigación clínica, hemos de ser capaces de trabajar de manera eficiente con este rompecabezas clínico de miles de pequeñas piezas que nuestro quehacer con el paciente genera cada día.

Flujo de datos en medicina intensiva

La información fluye en nuestro entorno a través de varias fases que vamos a estudiar brevemente (fig. 1), ya que en distintos momentos del proceso de atención clínica se producen eventos críticos que influyen sobre nuestra capacidad de utilizar eficientemente la información.

Entrada de datos

Una parte del proceso de introducción de los datos se realiza de manera automática y en un formato estructurado que permite el empleo de los datos de manera eficiente. Sin embargo, existe una gran ventana de oportunidad de mejora en la adquisición de los datos para los que hoy en día se utiliza el texto libre. Intentemos explicar esto con un ejemplo:

«Paciente varón de 58 años que acude hace 1 hora a urgencias por un cuadro de disnea y dolor centrotorácico irradiado al brazo izquierdo...»

En este breve párrafo de 24 palabras existen al menos 8 conceptos que pueden expresarse de manera estructurada en el siguiente documento xml (un formato estándar para mostrar información[11]):

```
<Episodio id=xxxx>
<Paciente>
<ID>xxxx</ID>
<Sexo>Varón</Sexo>
<Edad unidad=''años''>58</Edad>
</Paciente>
<Lugar>Urgencias</Lugar>
<Momento>2018-08-12 15:30:00</Momento>
<Fecha registro<2018-08-12 16:30:00</Fecha registro>
<Síntomas>
<Disnea/>
<Dolor>
<Localización>Centrotorácica</Localización>
<Irradiación>Brazo izquierdo</Irradiación>
</Dolor>
</Síntomas>
</Episodio>
```

Para un médico suele ser mucho más fácilmente interpretable el párrafo de texto libre, mientras que para su manejo en un computador la segunda manera de representar la información es mucho más eficiente. Lo ideal sería encontrar una herramienta que permita al humano introducir la información de una manera al menos tan rápida y cómoda como el texto libre, pero que también permita su almacenamiento directo en un formato estructurado, facilitando su lectura y comprensión tanto por humanos como por ordenadores.

Procesamiento de los datos en vivo

Los datos que se van introduciendo en el sistema quedan en un estado «pasivo», y es el profesional sanitario el que tiene que hacer inferencias tras recopilar los datos que le parecen relevantes en un momento concreto de la evolución del paciente.

Sin embargo, existe la posibilidad de transformar los datos almacenados en datos «activos», que desencadenen respuestas del sistema, a través de un sistema de soporte a la decisión clínica (*Clinical Decission Support System* [CDSS]). Siguiendo con el ejemplo previo, supongamos que hemos introducido el párrafo previo y en el sistema existe un concepto almacenado:

<Síndrome>
<Nombre>Síndrome coronario agudo</Nombre>
<Síntomas>
<Dolor>
<Localización>Centrotorácica</Localización>
<Irradiación>Brazo izquierdo</Irradiación>
</Dolor>
</Síntomas>
</Síndrome>

Al introducir la información del paciente el sistema puede alertarnos directamente de que el paciente muestra un síntoma compatible con el síndrome coronario agudo. Este es un ejemplo muy sencillo, pero lo realmente útil de las herramientas informáticas es que este tipo de procesamiento puede realizarse de manera automática para miles de conceptos y de una manera mucho más sistemática que la que realiza un cerebro humano. Para entenderlo con una analogía, es como si el sistema estuviera realizando un *checklist* continuo con nuestros pacientes. Adicionalmente, las técnicas actuales de inteligencia artificial (IA) permiten utilizar la experiencia adquirida de pacientes previos en forma de datos estructurados en la valoración del paciente individual subsiguiente. Esto puede tener en el futuro un gran impacto sobre la práctica clínica.

Otro ejemplo es el empleo de sistemas de control en tiempo real, por ejemplo, para ajustar continuamente las dosis de insulina en perfusión de un paciente en función de la sensibilidad a la insulina que ha mostrado previamente, su situación de estrés valorado por distintos datos de la gráfica y la analítica, y el aporte calórico y de hidratos de carbono que se está realizando en ese momento. El sistema puede ir aprendiendo con el tiempo sobre el propio paciente y optimizando progresivamente el control de la glucemia dentro de unos parámetros de seguridad.

Almacenamiento de los datos

Existen dos maneras de almacenar la información de nuestros pacientes: a) en bases de datos relacionales, que utilizan el lenguaje de consulta SQL para poder recuperar y procesar los datos almacenados, y b) en repositorios de datos no estructurados (NoSQL). Hoy en día existen herramientas que permiten extraer y procesar información también de este tipo de datos. Tipos específicos de datos pueden almacenarse en bases de datos especializadas, como por ejemplo un PACS para almacenamiento de imágenes clínicas que utiliza el formato DICOM.

Análisis de los datos a posteriori y compartición de la información

Una vez que tenemos almacenados los datos, podemos procesarlos de diversas formas. Por ejemplo, podemos obtener un informe automático de actividad de nuestra unidad utilizando una consulta a nuestra base de datos SQL con los datos demográficos y clínicos de nuestros pacientes, o podemos averiguar qué antibióticos utilizamos con más frecuencia, o qué tipo de patologías se atienden primordialmente en nuestra unidad, sin tener que revisar de nuevo uno por uno nuestros pacientes. Podemos utilizar esta información para llevar a cabo *benchmarking* (comparación de resultados entre unidades o dentro de la misma unidad al cambiar procesos o recursos) o para planificar actuaciones estratégicas.

Recomendamos consultar el artículo de Celi et al.[12], en el que se revisa todo este ciclo de flujo de información y se aborda el concepto de «cerrar el círculo» *(closing the loop)*, explicando de qué manera la IA puede contribuir al desarrollo de la medicina intensiva.

¿Qué es Big Data?

El planteamiento clásico del clínico con respecto al manejo de los datos para investigación clínica se basa en la premisa de que es necesario obtener información de calidad para conseguir resultados fiables y aplicables a los pacientes. La dificultad para obtener la información hace que se intente optimizar este proceso aplicando diseños prospectivos, aleatorización y una hipótesis de trabajo previa al análisis de los datos.

Como alternativa al uso exclusivo de datos recogidos de manera ortodoxa, la novedad que las técnicas de análisis de Big Data (*Big Data Analysis* [BDA]) ofrecen es la detección de la estructura y el conocimiento subyacente en cantidades ingentes de información, incluso aunque aparentemente no esté estructurada.

Una definición popular del concepto[13] establece que «se considera Big Data a conjuntos de datos caracterizados por un volumen tan grande y una variedad tan amplia en su estructura, que hace necesario el uso de tecnología y métodos analíticos específicos para su procesamiento y transformación en conocimiento o valor».

Las técnicas específicas de BDA se han aplicado con éxito a campos como el marketing, la toma de decisiones estratégicas en el mundo de los negocios, la banca, el transporte, la logística, los seguros o la detección de fraude en el comercio electrónico. No hay ningún motivo para pensar que no puedan aplicarse a nuestro entorno, donde continuamente

	A	B	C	D	E	F	G	H
1	episodio	edad	sexo	apache2	SOFA	tipo ingreso	lactico	exitus (S/N)
2	3654	67	H	22	3	QP	1.2	N
3	2342	34	H	11	2	QP	0.9	N
4	1156	81	M	34	8	M	2.4	S
5	9856	59	M	30	7	QP	3.0	N
6	2317	75	H	21	1	M	4.9	N
7	9042	75	M	16	3	QU	2.9	N
8	4510	82	M	25	4	QP	0.9	N
9	5557	60	H	24	3	M	1.7	N
10	7623	63	H	12	1	QP	0.5	N
11	1995	71	H	19	3	M	1.8	N
12	5489	47	H	30	5	QU	5.0	S
13	1115	69	M	21	3	QU	1.2	N
14	6737	78	H	12	1	QP	0.5	N
15	2499	74	M	12	2	QP	1.2	N
16	2009	45	M	10	3	QP	1.9	N

Figura 2 Estructura típica de un *dataframe*, constituido por filas por cada registro con un identificador anonimizado y columnas donde se almacenan los valores de distintas variables para cada caso. En la última columna en este caso se muestra la variable objetivo (en este caso mortalidad). Aquí podemos ver una sección de unos cuantos registros de la base de datos que se utiliza como ejemplo en el material incluido en el repositorio de GitHub.

estamos tomando decisiones estratégicas o concretas en pacientes determinados, cuyas características, aunque complejas, se repiten con frecuencia.

Un ejemplo reciente, del que hablamos más adelante, es el seguimiento de las epidemias de gripe mediante el análisis de las búsquedas realizadas en internet. En este caso se utilizan bases de datos de muchos millones de registros para obtener conclusiones epidemiológicas rápidas y fiables[14].

La filosofía de código abierto ha irrumpido con fuerza en el mundo del Big Data, permitiendo su utilización sin que ello suponga una gran inversión económica. Proyectos como Apache Hadoop[15,16] incluyen toda una serie de recursos en su ecosistema (HDFS, Spark, MapReduce, Impala, HBase, Hive) que permiten construir a bajo coste un sistema de procesamiento de Big Data en *batch* (análisis de datos ya almacenados en grandes bases de datos relacionales o no relacionales) o en *streaming* (análisis y procesamiento de los datos según se van generando) de bajo coste.

¿Qué es Machine Learning?

El concepto de Machine Learning (ML) o «aprendizaje máquina» data de mediados del siglo XX, y se definió ya en un artículo de Samuel de 1959[17] como un apartado de la IA que usa técnicas estadísticas y algoritmos computacionales para proporcionar a los ordenadores la capacidad de «aprender», es decir, mejorar sus resultados en una tarea específica tras procesar datos en suficiente cantidad y sin unas instrucciones explícitas externas (y por tanto potencialmente sesgadas) proporcionadas por el programador.

El ámbito del ML está estrechamente relacionado con otros campos, como la simulación y el modelado, la optimización de sistemas y la estadística. En todos ellos se emplean de manera intensiva técnicas matemáticas comunes que requieren entrenamiento específico.

Es necesario definir algunos conceptos básicos en este momento. Debemos utilizar un formato conocido como «*dataframe*» para trabajar con los datos en ML. Un *dataframe* es una matriz donde cada fila corresponde a uno de los pacientes o registros y cada columna a una de las variables que registramos (fig. 2).

Supongamos, a modo de ejemplo, que queremos desarrollar un modelo de predicción de mortalidad en función de las siguientes variables: tipo de ingreso (médico, quirúrgico programado o quirúrgico urgente), edad, sexo, APACHE II al ingreso, SOFA al ingreso y niveles sanguíneos de ácido láctico. En este caso tendremos tantas filas como episodios de pacientes tengamos recogidos, y cada fila tendrá una columna que identifica el episodio (de manera anónima) y una columna por cada una de las variables explicativas que hemos definido más otra columna con la variable objetivo (en este caso es una variable dicotómica mortalidad [SÍ/NO]).

Para los lectores que quieran reproducir los ejemplos que se proporcionan en esta revisión hemos preparado un conjunto anonimizado de datos de nuestra base propia y la hemos subido en forma de tabla Excel a un repositorio en GitHub (https://github.com/anunezr/revision_medicina_intensiva).

Existen tres variantes principales de aprendizaje máquina:

- *Aprendizaje supervisado*. Cada registro está etiquetado con un valor de la variable objetivo, y se emplean distintas técnicas capaces de predecir en un registro nuevo el valor de esa variable. En nuestro ejemplo, la variable objetiva es la mortalidad (SÍ/NO). Una vez «entrenado» el sistema, este ha de ser capaz de predecir el valor de esa variable para un episodio que no haya sido presentado al sistema previamente. Por tanto, los datos de entrenamiento y los de prueba deben ser diferentes y estar almacenados por

separado. Se utiliza en tareas de clasificación, predicción y detección de similitud.
- *Aprendizaje no supervisado*. En este caso se trata de detectar patrones o tendencias en los datos sin utilizar una variable objetivo. Se utiliza por ejemplo para clasificar pacientes en grupos de manera automática y para reducir el número de variables y la complejidad de los modelos.
- *Aprendizaje por refuerzo*. En este caso el sistema debe perseguir un objetivo o recompensa, e irá aprendiendo a medida que se va explorando el entorno con el que interactúa, que no se conoce de antemano, evitando las acciones con recompensa negativa e intentando llevar a cabo acciones con recompensa positiva. Un ejemplo sería un sistema que va aprendiendo y ajustando las pautas antibióticas empíricas que realizan los clínicos según van siendo los resultados y las características de los pacientes sépticos que se van presentando.

En la tabla 1_Sup, que se incluye en el anexo A de material suplementario, se resumen las distintas técnicas de ML para cada una de estas variantes. El empleo de estas técnicas requiere tener unos conocimientos en programación y un dominio de los conceptos utilizados en el campo de la IA que normalmente no están al alcance del clínico. Por ello se va imponiendo un nuevo enfoque de trabajo colaborativo que incluye la aparición en las unidades de un nuevo tipo de profesional, con conocimientos clínicos básicos y dominio avanzado de la estadística y de las herramientas y métodos de BDA y ML. Estos profesionales, trabajando codo con codo con los clínicos, pueden ayudarnos a extraer valor de la gran cantidad de información clínica que ahora mismo se encuentra almacenada en nuestras bases de datos. Este *man in the middle*, que empieza a aparecer ya en centros punteros en desarrollo tecnológico, será en un futuro cercano una parte importante del equipo de nuestras unidades y ayudará a planificar las decisiones estratégicas y a optimizar las herramientas informáticas para obtener de ellas el máximo rendimiento.

Para los lectores con conocimientos de R o Python que quieran adquirir práctica básica en el campo del ML incluimos en el repositorio de GitHub asociado a esta revisión (https://github.com/anunezr/revision_medicina_intensiva) una serie de scripts que utilizan las distintas técnicas reflejadas en la tabla 1_Sup sobre la base de datos anonimizada que se proporciona como ejemplo.

En la tabla 1 se muestra un «mapa de ruta» para poner en práctica un sistema de ML adaptado a un problema concreto[18]. Cada uno de los aspectos que se contemplan daría para una revisión tan extensa o más que este manuscrito. El objeto de la tabla es hacer entender que el empleo de BDA y ML precisa de una metodología y una sistemática compleja que requiere conocimientos y experiencia específicos para conseguir resultados. Solo bajo estas premisas podremos extraer conocimiento de los datos eficazmente.

Tabla 1 Pasos a seguir para construir un sistema de Machine Learning

Limpieza, reorganización y preprocesamiento
Una vez extraídos los datos se procede a verificar su calidad y darles el formato y escala adecuados, resolviendo problemas de ausencias e inconsistencias para prepararlos para ser procesados por el modelo correspondiente
Son de hecho estas primeras dos tareas las que consumen la mayor parte del tiempo en un proyecto Big Data Analysis (BDA) y/o Machine Learning (ML)
Selección de atributos
Escoger qué variables se van a utilizar en el proceso de aprendizaje, reduciendo al mínimo posible su número y dimensionalidad. Sin embargo, una ventaja del ML frente a modelos de estadística convencional es que es capaz de asimilar mayor cantidad de variables y, basándose en ellas, hacer predicciones más potentes
Creación de los datasets de entrenamiento y validación
Decidir el tipo de muestreo y cómo vamos a dividir los conjuntos de datos entre entrenamiento, validación y prueba
Selección del modelo y ajuste hiperparamétrico
Escoger los algoritmos de aprendizaje que vamos a utilizar (selección de modelo, validación cruzada, métricas de resultados, optimización de hiperparámetros). Comparar los resultados entre los distintos algoritmos y modelos utilizados
Evaluación del funcionamiento del modelo elegido
Predicción
Empleo del modelo en nuevos casos y reevaluación de su funcionamiento

Herramientas

Para poder hacer uso de las técnicas de BDA y ML es necesario dominar al menos algunos de los lenguajes de computación estadística para análisis (R, Python o Java), dominar SQL como herramienta de consulta de bases de datos y saber utilizar las bibliotecas de código que se utilizan en este campo. En el apéndice de material suplementario puede verse una breve reseña del software más utilizado.

Dificultades y estrategias para aplicar estas técnicas en un servicio de medicina intensiva

Después de esta ensalada de acrónimos y tecnicismos, el intensivista de a pie tiende a preguntarse si realmente las técnicas de BDA y ML pueden aplicarse a su entorno de trabajo, y cómo hacerlo en la práctica de manera asequible. En realidad, el objetivo de esta revisión es concienciar al clínico de que entrar en el mundo de la inteligencia artificial aplicada a la medicina intensiva es posible, pero exige cambios estructurales e inversión en recursos humanos y tecnología, así como una visión ampliada del concepto de investigación clínica.

El primer requisito, lógicamente, es un sistema de adquisición y almacenamiento de los datos que permita su procesamiento analítico posterior. En España todavía estamos muy lejos de conseguirlo, ya que los sistemas de información clínica informatizados distan de tener una implantación universal en nuestras unidades, por no hablar de un estándar interhospitalario. Además, un gran número

de las que poseen estas aplicaciones departamentales carecen de un acceso eficiente a los datos almacenados (a menudo porque el contrato de adquisición de los sistemas no incluyó en su momento este aspecto, que suele tener un coste adicional). Estas situaciones paradójicas, que rayan en la falta de ética al quedar los datos clínicos «prisioneros» en manos de los proveedores del software comercial, deben ser solventadas con urgencia si queremos empezar a trabajar con las nuevas herramientas de BDA y ML por el bien de nuestros pacientes.

En segundo lugar, hace falta que los hospitales asignen personal especializado de los departamentos de tecnología de la información a las labores de custodia, procesamiento y análisis de los datos clínicos de nuestras unidades, trabajando de manera colaborativa con los médicos y enfermeras para conseguir obtener el máximo rendimiento de la información almacenada.

En tercer lugar, es necesario aliarse y trabajar conjuntamente con los comités de ética locales y regionales para conseguir garantizar por un lado la seguridad y la privacidad de los datos y, por otro lado, no ensombrecer las oportunidades de mejora de los procesos y la atención clínica que la explotación de los datos almacenados puede proporcionarnos.

Bases de datos compartidas

Se conoce como análisis secundario de las bases de datos de salud electrónicas (*Electronic Health Records* [EHR])[19] al procesamiento que se realiza sobre los datos clínicos de los pacientes generados como subproducto de su adquisición con fines asistenciales y que puede utilizarse como ayuda en la toma de decisiones estratégicas en una unidad o puntuales para un paciente concreto.

En un afán por potenciar la investigación clínica en el ámbito de la medicina intensiva, algunas de las instituciones que disponen de bases de datos clínicos muy extensas las han puesto a disposición de los investigadores de todo el mundo tras un proceso de anonimización para mantener la privacidad de los datos. La más popular en el momento actual es *Medical Information Mart in Intensive Care* (MIMIC-III)[20,21], que reúne como características principales el ser de acceso público y presentar una excelente calidad de los datos incluidos, que abarcan información clínica de todo tipo, extraída de más de 58.000 pacientes críticos, tanto adultos como neonatos.

Otro ejemplo de estas bases de datos sería *eICU Collaborative Research Database*[22], poblada con datos de una combinación de numerosas unidades de cuidados intensivos en todo el territorio continental de Estados Unidos. Los registros de la base de datos colaborativa recogen información de casi 200.000 pacientes que fueron admitidos en las unidades de cuidados críticos en 2014 y 2015.

Como podemos imaginar, aunque este análisis secundario puede ser de gran utilidad para un servicio de medicina intensiva o para un investigador clínico, su proyección se ve incrementada si los datos de varias unidades pueden combinarse para formar una gran base de datos multicéntrica. El efecto es similar al que sucede cuando consideramos las conclusiones que pueden obtenerse de un ensayo clínico realizado en un único centro con respecto a un ensayo multicéntrico.

Sin embargo, para poder llegar a este objetivo es necesario vencer una serie de dificultades.

Cada unidad recoge los datos en una herramienta de software que puede no ser la misma en todas las unidades cuyos datos se quieren agrupar. Por tanto, hay que realizar un proceso de conversión a una estructura de datos común que permita su análisis conjunto. Esto es lo que antes hemos denominado «mapeo». Para ello han surgido ontologías que permiten expresar y mapear los conceptos almacenados de manera uniforme. Algunas de ellas, como *Observational Medical Outcomes Partnership* (OMOP)[10] o *Informatics for Integrating Biology and the Bedside* (i2b2)[23], ofrecen herramientas para facilitar la migración de la base de datos local a un estándar común de datos.

Si tenemos en cuenta únicamente los estudios publicados en los últimos años utilizando la base de datos pública MIMIC[20] y bibliotecas de software con algoritmos de última generación y acceso gratuito, nos encontramos con que se han obtenido resultados interesantes en campos muy diversos de la medicina intensiva, que se muestran en la tabla 2.

Además, una ventaja de esta estrategia es que los resultados son completamente reproducibles: podemos tomar el conjunto de datos que han utilizado los autores y reproducir los métodos estadísticos y de aprendizaje máquina que

Tabla 2 Aspectos de investigación clínica estudiados mediante técnicas de Big Data Analysis y Machine Learning

La utilidad del uso la ecocardiografía en la sepsis[35]
La utilidad de conocer la variabilidad de los niveles de glucosa en sangre en el pronóstico de los pacientes no diabéticos[36]
La mejora en la predicción de mortalidad en UCI en función de parámetros fisiológicos[37]
La detección precoz de la sepsis[38] y la posibilidad de generar una alerta precoz en los pacientes con alto riesgo de desarrollar shock séptico[39]
La diferencia entre los objetivos de balance hídrico a establecer entre las primeras y las segundas 24 h de resucitación de la sepsis[40]
El valor predictivo de las notas de enfermería en el pronóstico a largo plazo de los pacientes[41]
La utilidad del drenaje precoz de la vía biliar en la colangitis aguda[42]
La capacidad de las herramientas de procesamiento de lenguaje natural para detectar diagnósticos no codificados en notas de texto libre[29]
El valor pronóstico negativo de la taquicardia prolongada[43]
La paradójica relación entre obesidad y mortalidad en UCI en pacientes con patologías crónicas[44]
La predicción de los reingresos precoces no deseados[45]
El empleo de variables clínicas adicionales para ajustar mejor las dosis de heparina en perfusión[46]
El peor pronóstico que presentan tras ingresar en UCI pacientes que toman inhibidores de la recaptación de serotonina[47]

han utilizado, incluso con nuestros propios datos, ya que están disponibles habitualmente en repositorios públicos como GitHub[24,25]. Esto hace que los estudios estén sometidos a un control de calidad adicional y que cualquiera pueda revisar tanto métodos y como resultados.

Existen condicionantes de tipo legal y ético que pueden dificultar el proceso de compartición de los datos, y puede ser necesario que los comités de ética locales autoricen el acceso y verifiquen la salvaguarda de la privacidad y la seguridad de los datos.

La logística del proceso requiere una acción coordinada entre los centros participantes y un soporte tecnológico y de hardware en consonancia con el proyecto a través de equipos multidisciplinares que puedan hacer frente a los retos que las distintas facetas de tamaña empresa requieren.

Una vez superadas estas dificultades, las oportunidades de mejora de la gestión, de *benchmarking* y de avance en la investigación clínica son enormes, y sin duda compensarán el esfuerzo que la incorporación de un servicio de medicina intensiva en un proyecto de este tipo supone.

Procesamiento del lenguaje natural

Dado que la mayoría de la información que el profesional introduce en la historia clínica está en formato de texto libre (también conocido como lenguaje natural), un posible campo de aplicación de las técnicas de inteligencia artificial a la medicina es la creación de información estructurada a partir del texto libre de manera automática, así como la clasificación o fenotipado de los pacientes en distintos grupos en función del contenido de lo escrito por los profesionales. Estas técnicas, conocidas como procesamiento del lenguaje natural (*Natural Language Processing* [NLP]), están avanzando rápidamente[26]. Existen dos posibles enfoques:

- La estrategia de extracción de conceptos a partir del texto utilizando herramientas como cTAKES[27] está bastante desarrollada en el mundo anglosajón, pero existen pocas opciones prácticas para texto no escrito en inglés. El grupo de Investigación en Minería de Datos del Centro de Tecnología Biomédica de la Universidad Politécnica de Madrid está trabajando con una herramienta derivada de cTAKES llamada TIDA[28], y es posible que en el futuro este tipo de sistemas se incorporen a nuestro arsenal informático.
- Como alternativa está el empleo de redes neuronales convolucionales (CNN)[29], pero este tipo de estrategia tiene el inconveniente de que los algoritmos que se obtienen no son fácilmente interpretables por el médico (aunque si el modelo funciona correctamente esto puede ser irrelevante), y sus resultados son variables dependiendo de los fenotipos buscados.

Análisis retrospectivo de Big Data versus ensayos clínicos aleatorizados

Dado que los datos clínicos que se almacenan en nuestros sistemas de información clínica no se recogen con la misma exigencia que aplicamos a los ensayos clínicos, ¿cómo es posible que el análisis secundario de los datos asistenciales nos proporcione conclusiones fiables? La clave está en el volumen de datos y en la capacidad de las herramientas de ML de detectar estructura en medio del caos[19]. Podemos utilizar como ejemplo el seguimiento que se realizó de la incidencia de la gripe H1N1 durante la epidemia de 2008 mediante el contaje de las búsquedas en Google utilizando el término «gripe»[14], que se ha seguido validando posteriormente (fig. 3), y que se correlaciona perfectamente con los registros de casos obtenidos mediante métodos más ortodoxos.

Aunque hasta ahora no podemos decir que BDA y ML hayan desbancado a los métodos tradicionales de investigación clínica, sí es cierto que permiten la detección de tendencias o patrones en grandes conjuntos de datos que pueden pasar desapercibidos al investigador y que pueden guiar el diseño de nuevos estudios que se atengan a la metodología convencional[30]. Además, son la única alternativa en aquellas situaciones en las que no es posible logísticamente plantearse un ensayo clínico ortodoxo con suficiente poder estadístico para resolver una pregunta relevante.

Cuestiones éticas y legales

En un reciente artículo en *Intensive Care Medicine*, McLennan et al.[31] hacen un excelente análisis de los problemas que la compartición de datos clínicos incluidos en bases de datos asistenciales puede generar. Desde la puesta en vigor de la directiva general europea sobre protección de datos (GDPR)[32] en mayo de 2018, las instituciones han de cumplir una normativa con respecto al manejo de los datos personales de pacientes que residen en la Unión Europea. El equivalente que impera en Estados Unidos de América es conocida como *Health Insurance Portability and Accountability Act* (HIPAA), de 1996[33].

El aspecto clave es la anonimización de los datos: si los datos son completamente anónimos, no hay problema legal en su compartición desde el punto de vista de los derechos de privacidad y seguridad. Pero ¿cuándo podemos considerar nuestros datos completamente anónimos? Únicamente cuando no lleven a la identificación unívoca del paciente cuando se complementan con otros datos (por ejemplo, aunque no sepamos el nombre del paciente o su número de historia, si sabemos la fecha de ingreso y la edad podemos llegar a identificar de qué paciente se trata). También debe evitarse la identificación unívoca del personal sanitario que aparece en la historia clínica del paciente. En todos los demás casos se considera que los datos están pseudoanonimizados, y en ese caso estaríamos obligados, según la directiva GDPR, a pedir consentimiento al paciente para su utilización.

Por ello resulta crucial el paso de anonimización de las bases de datos y es necesario estandarizar un proceso que permita cumplir la normativa europea perdiendo el mínimo valor posible de la información. Los investigadores deben trabajar juntamente con los comités de ética locales para armonizar la salvaguarda de la privacidad y seguridad de los datos con los avances en la investigación clínica que las nuevas herramientas de BDA nos pueden proporcionar en el futuro. Deben desarrollarse mecanismos que faciliten el consentimiento del paciente (o su retirada) a la utilización de los datos generados en su proceso asistencial y convenientemente anonimizados para la investigación clínica.

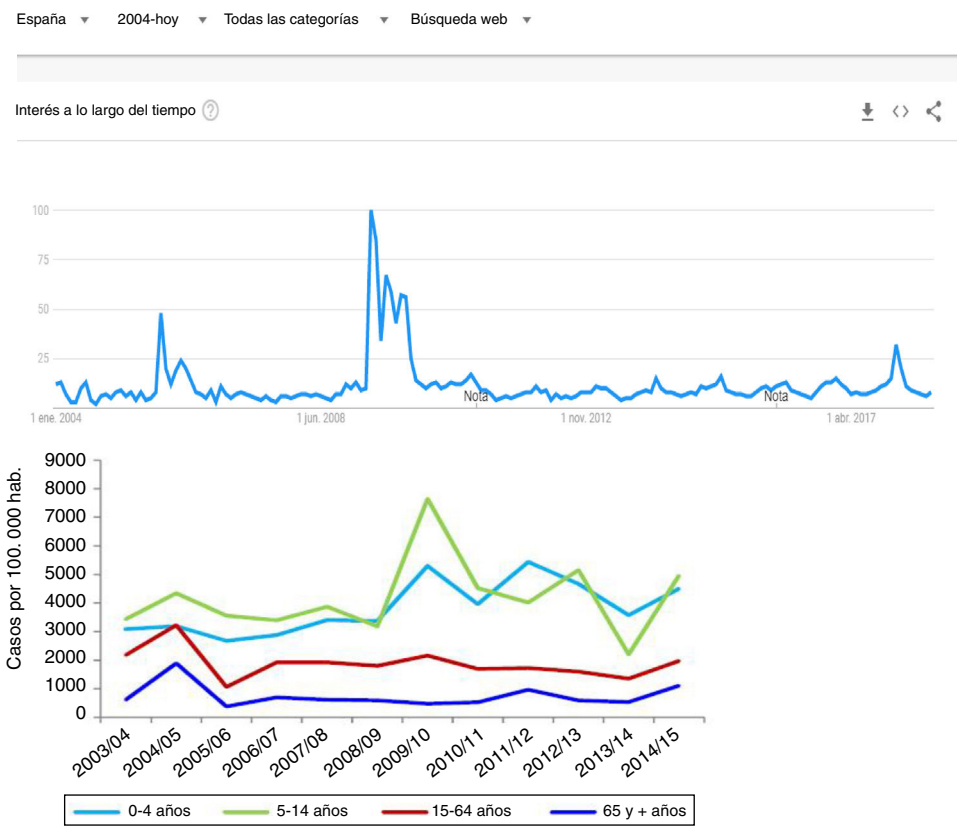

Figura 3 Uno de los primeros ejemplos de la aplicación del análisis de Big Data a la investigación clínica. Durante la epidemia de gripe N1H1 de 2008, las búsquedas en Google del término «gripe» o sus síntomas predijeron con gran precisión el aumento de casos y la severidad del cuadro. En la imagen superior puede visualizarse la cantidad de búsquedas y en la imagen inferior la evolución de los casos según el Sistema de Vigilancia Oficial de Gripe en España (las escalas de tiempo están ajustadas para que se correspondan verticalmente ambos gráficos). Este hallazgo ha sido utilizado posteriormente como sistema de vigilancia epidemiológico con gran éxito.

Otro aspecto importante que tratar es el tema de la propiedad de los datos. En este caso estamos enfocando ya no aspectos de seguridad o privacidad, sino más bien su valor intrínseco (incluso como bien comercial). ¿Tienen el paciente (origen de los datos) o el clínico (actor del proceso que introduce la información en el sistema) derecho a exigir una compensación económica o a prohibir el empleo de los datos para investigación clínica o cualquier otro uso? Aunque la realidad actual es que los datos están frecuentemente en manos de las empresas desarrolladoras del software sanitario, que son las que pretenden obtener un beneficio comercial por su explotación[34], este aspecto está hoy en día sometido a una gran controversia y sería deseable una clarificación por parte del legislador para resolver los problemas que ahora mismo se plantean.

Conclusiones

Las herramientas de Big Data Analysis y Machine Learning suponen una gran oportunidad para mejorar la gestión estratégica de las unidades, el manejo de casos clínicos concretos y la investigación clínica. Sin embargo, para poder aprovechar esta nueva metodología necesitamos evolucionar incorporando nuevos recursos humanos (personal especializado con conocimientos clínicos y entrenamiento en inteligencia artificial) y tecnológicos. Además, debemos ser capaces de armonizar los condicionantes de privacidad y seguridad de los datos de nuestros pacientes con la posibilidad de utilizar grandes bases de datos clínicas de manera eficiente.

Los autores estamos convencidos de que un equipo internacional y multidisciplinar, que trabaje colectivamente para extraer conocimiento de grandes conjuntos de datos clínicos según se ha explicado a lo largo de este manuscrito, puede traer una revolución a la práctica moderna de la atención al paciente crítico.

Financiación

Los autores no han recibido ningún soporte económico para la realización del manuscrito.

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses en relación con el contenido del artículo.

Anexo. Material adicional

Se puede consultar material adicional a este artículo en su versión electrónica disponible en doi:10.1016/j.medin.2018.10.007.

Bibliografía

1. Sevenster M, Buurman J, Liu P, Peters JF, Chang PJ. Natural language processing techniques for extracting and categorizing finding measurements in narrative radiology reports. Appl Clin Inform. 2015;6:600–10.
2. Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan WJ, Sinsky CA, et al. Tethered to the EHR: Primary care physician workload assessment using EHR event log data and time-motion observations. Ann Fam Med. 2017;15:419–26.
3. SNOMED International. SNOMED 2018. Disponible en: https://www.snomed.org/.
4. HL7. HL7 Standards 2018. Disponible en: http://www.hl7.org/.
5. NIH. UMLS 2018. Disponible en: https://www.nlm.nih.gov/research/umls/.
6. DICOM. DICOM 2018. Disponible en: https://www.dicomstandard.org/.
7. Institute R. LOINC 2018. Disponible en: https://loinc.org/.
8. NIH. Medical Subject Headings 2018. Disponible en: https://www.ncbi.nlm.nih.gov/mesh.
9. Diseases ICo. CIE 10 2018. Disponible en: https://eciemaps.msssi.gob.es/.
10. OHDSI. OMOP Data Model 2018. Disponible en: https://www.ohdsi.org/data-standardization/the-common-data-model/.
11. XML. Extensible Markup Language 2018. Disponible en: https://es.wikipedia.org/wiki/Extensible_Markup_Language.
12. Celi LA, Mark RG, Stone DJ, Montgomery RA. ''Big Data'' in the intensive care unit. Closing the data loop. Am J Respir Crit Care Med. 2013;187:1157–60.
13. Andrea DM, Marco G, Michele G. A formal definition of Big Data based on its essential features. Library Review. 2016;65:122–35.
14. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2008;457:1012.
15. Foundation AS. Apache Hadoop 2014. Disponible en: http://hadoop.apache.org/.
16. Mehta R. 1. Big Data Analytics with Java. Big Data Analytics with Java: Data analysis, visualization & machine learning techniques. Birmingham, UK: Packt; 2018.
17. Samuel AL. Some studies in machine learning using the game of checkers. IBM J Res Dev. 1959;3:210–29.
18. Raschka S MV. Chapter 1. Giving computers the ability to learn from data. Python Machine Learning Machine learning and deep learning with Python, scikit-learn and Tensorflow. Birmingham, UK: Packt; 2017, September.
19. Data MC. Secondary Analysis of Electronic Health Records 2016.
20. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035.
21. MIT Lab for Computational Physiology. MIMIC 2018. Disponible en: http://mimic.physionet.org.
22. Database eCR. eICU Collaborative Research Database 2018. Disponible en: https://eicu-crd.mit.edu/.
23. Computing NCfB. i2b2 (Informatics for Integrating Biology and the Bedside) 2018. Disponible en: https://www.i2b2.org.
24. Johnson AE, Stone DJ, Celi LA, Pollard TJ. The MIMIC Code Repository: Enabling reproducibility in critical care research. J Am Med Inform Assoc. 2018;25:32–9.
25. MIT-LCP. MIMIC Code Repository: Code shared by the research community for the MIMIC-III database 2018. Disponible en: https://github.com/MIT-LCP/mimic-code.
26. Pai VM, Rodgers M, Conroy R, Luo J, Zhou R, Seto B. Workshop on using natural language processing applications for enhancing clinical decision making: An executive summary. J Am Med Inform Assoc. 2014;21:e2–5.
27. Apache Software Foundation. Apache cTAKES™ 2018. Disponible en: http://ctakes.apache.org/.
28. Costumero R, Gonzalo C, Menasalvas E. TIDA: A Spanish EHR Semantic Search Engine. 8 th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2014). 1 st edition. New York: Springer; 2014: pp. 235-242.
29. Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, Welt J, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PLoS One. 2018;13, e0192360.
30. Mayo CS, Matuszak MM, Schipper MJ, Jolly S, Hayman JA, Ten Haken RK. Big Data in designing clinical trials: Opportunities and challenges. Front Oncol. 2017;7:187.
31. McLennan S, Shaw D, Celi LA. The challenge of local consent requirements for global critical care databases. Intensive Care Med. 2018, http://dx.doi.org/10.1007/s00134-018-5257.
32. EU GDPR Portal. EU General Data Protection Regulation (GDPR) 2018. Disponible en: https://www.eugdpr.org/.
33. HHS.gov. Health information privacy. 2018.
34. Tanner A. Our Bodies, Our Data: How Companies Make Billions Selling Our Medical Records. Boston: Beacon Press; 2016. p. 218.
35. Feng M, McSparron JI, Kien DT, Stone DJ, Roberts DH, Schwartzstein RM, et al. Transthoracic echocardiography and mortality in sepsis: Analysis of the MIMIC-III database. Intensive Care Med. 2018;44:884–92.
36. Liu WY, Lin SG, Zhu GQ, Poucke SV, Braddock M, Zhang Z, et al. Establishment and validation of GV-SAPS II scoring system for non-diabetic critically ill patients. PLoS One. 2016;11:e0166085.
37. Calvert J, Mao Q, Hoffman JL, Jay M, Desautels T, Mohamadlou H, et al. Using electronic health record collected clinical variables to predict medical intensive care unit mortality. Ann Med Surg (Lond). 2016;11:52–7.
38. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: A Machine Learning approach. JMIR Med Inform. 2016;4:e28.
39. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Sci Transl Med. 2015;7, 299ra122.
40. Shen Y, Ru W, Huang X, Zhang W. Time-related association between fluid balance and mortality in sepsis patients: Interaction between fluid balance and haemodynamics. Sci Rep. 2018;8, 10390.
41. Waudby-Smith IER, Tran N, Dubin JA, Lee J. Sentiment in nursing notes as an indicator of out-of-hospital mortality in intensive care patients. PLoS One. 2018;13:e0198687.
42. Aboelsoud M, Siddique O, Morales A, Seol Y, al-Qadi M. Early biliary drainage is associated with favourable outcomes in critically-ill patients with acute cholangitis. Prz Gastroenterol. 2018;13:16–21.
43. Sandfort V, Johnson AEW, Kunz LM, Vargas JD, Rosing DR. Prolonged elevated heart rate and 90-day survival in acutely ill patients: Data from the MIMIC-III database. J Intensive Care Med. 2018, http://dx.doi.org/10.1177/0885066618756828, 885066618756828.
44. Janice P, Shaffer R, Sinno Z, Tyler M, Ghosh J. The obesity paradox in ICU patients. Conf Proc IEEE Eng Med Biol Soc. 2017;2017:3360–4.

45. Desautels T, Das R, Calvert J, Trivedi M, Summers C, Wales DJ, et al. Prediction of early unplanned intensive care unit readmission in a UK tertiary care hospital: A cross-sectional machine learning approach. BMJ Open. 2017;7:e017199.
46. Ghassemi MM, Richter SE, Eche IM, Chen TW, Danziger J, Celi LA. A data-driven approach to optimized medication dosing: A focus on heparin. Intensive Care Med. 2014;40:1332–9.
47. Ghassemi M, Marshall J, Singh N, Stone DJ, Celi LA. Leveraging a critical care database: Selective serotonin reuptake inhibitor use prior to ICU admission is associated with increased hospital mortality. Chest. 2014;145:745–52.