[REV. MED. CLIN. CONDES - 2019; 30(1) 83-94]

REVISTA MÉDICA CLÍNICA LAS CONDES
https://www.journals.elsevier.com/revista-medica-clinica-las-condes

Una guía conceptual para usar y entender Big Data en la investigación clínica

A conceptual guide to use and understand Big Data in clinical research

Marcelo E. Andía MD/PhD[a], Cristóbal Arrieta MS PhD[b], Carlos A. Sing Long MS PhD[c] ✉,

[a] Departamento de Radiología, Escuela de Medicina, Pontificia Universidad Católica de Chile. Santiago, Chile.
[b] Centro de Imágenes Biomédicas, Pontificia Universidad Católica de Chile. Santiago, Chile.
[c] Instituto de Ingeniería Matemática y Computacional, Pontificia Universidad Católica de Chile. Santiago, Chile.

INFORMACIÓN DEL ARTÍCULO

Historia del Artículo:
Recibido: 10 08 2018.
Aceptado: 27 11 2018.

Palabras clave:
Análisis estadístico de datos, aprendizaje de máquina, inteligencia artificial, minería de datos.

Key words:
Statistical data analysis, machine learning, artificial intelligence, data mining.

RESUMEN

Hoy nos encontramos en medio de cambios profundos en nuestra economía y sociedad impulsados por el análisis de cantidades masivas de datos. La investigación y práctica clínica se encuentran prestas a ser revolucionadas por metodologías que extraen información útil de un gran volumen de registros clínicos y que puede no ser evidente al utilizar los métodos tradicionales de análisis. En los últimos años, la cantidad de artículos científicos que hacen uso de estos métodos en un contexto académico y que reportan resultados exitosos se ha incrementado. Junto con éstos, los artículos de prensa que advierten que médicos y radiólogos podrían ser reemplazados por estos métodos en el futuro se han incrementado. Sin embargo, ¿cómo evaluamos el impacto real de estas metodologías en la práctica? Este artículo presenta un marco conceptual que define las ideas principales tras Big Data y la Ciencia de Datos y permite identificar los criterios para evaluar el potencial impacto de estos métodos en la investigación y práctica clínica. Además, con este marco discutimos los resultados de algunos estudios importantes que han captado la atención en la prensa y finalizamos con los principales desafíos que presenta la adopción de estos métodos en medicina.

ABSTRACT

Today we find ourselves amidst profound changes in our economy and our society driven by the analysis of massive datasets. Clinical practice and research are poised to be revolutionized by methods that extract useful information from large volumes of medical records that might not be evident when using traditional medical analysis techniques. The number of scientific articles that report successful results when applying these methods of analysis, both in academic and clinical settings, has increased in recent years. Simultaneously, the number of articles in the media warning that medical doctors and radiologists might one day be replaced by these automated methods has also increased. However, how do we evaluate in practice the impact of these methods? This presents a conceptual framework that introduces the main ideas behind Big Data and Data Science and points out the main criteria to be used to assess the potential impact of these techniques in medical research and practice. In addition, we discuss within this framework the results of some of studies that have been reported in the media, and we end by laying out the main challenges that pose the adoption of these methods in practice.

✉ Autor para correspondencia
Correo electrónico: casinglo@uc.cl

https://doi.org/10.1016/j.rmclc.2018.11.003

1. INTRODUCCIÓN

En las últimas décadas nuestra capacidad de generar y almacenar datos se ha incrementado de manera exponencial[1,2]. Simultáneamente, nuestra capacidad de *procesamiento* a gran escala se ha incrementado, permitiendo *analizar* los datos generados por las actividades que realizamos día a día [3]. Si bien hasta hoy el análisis de datos ha sido una de las piedras angulares del avance científico, tecnológico y económico, es la *cantidad* de datos disponibles hoy la que ha abierto la puerta a nuevas oportunidades. Por ejemplo, existen patrones de comportamiento de consumidores imposibles de detectar con pocos datos, los cuales se hacen evidentes a gran escala; del mismo modo, los parámetros de ciertos modelos predictivos, que en ausencia de datos suficientes son escogidos gracias a la pericia de profesionales del área, pueden ser estimados de manera precisa cuando la cantidad de datos es masiva. Por lo tanto, es la combinación actual entre la capacidad de almacenar y procesar datos a escala masiva la que ha comenzado a revelar estructuras latentes en las actividades humanas que éstos reflejan.

Sin duda, la investigación clínica es una de las áreas en las que el análisis de datos a gran escala promete tener un mayor impacto[4-7]. Por ejemplo, revelar patrones en la expresión genética de pacientes permitiría elucidar los mecanismos a través de los cuales ciertas enfermedades actúan[8]. Determinar qué estructuras moleculares tienen correlaciones fuertes con efectos fisiológicos podría tener un gran impacto en el desarrollo de nuevos fármacos[9]. Y agrupar una cantidad masiva de casos y controles permitiría *validar* conclusiones clínicas obtenidas a partir de estudios con un número reducido de participantes[10]. Esta es una fracción ínfima de ejemplos: el número de artículos publicados que relacionan *inteligencia artificial* con *diagnóstico clínico* pasó de 110 artículos anuales en promedio, en la década de los 90, a más de 770 artículos el año 2017 de acuerdo a la base de datos Pubmed.

Como metodología, el análisis de datos a gran escala ha tenido un gran éxito comercial. Rápidamente podemos nombrar dos ejemplos emblemáticos: el desarrollo de perfiles de usuarios que Google y Facebook realizan a partir de datos, los cuales permiten distribuir avisos publicitarios dirigidos[11,12]; el otro es el desarrollo de sistemas de recomendación en Amazon y Netflix, los cuales permiten predecir las preferencias de un consumidor[13-15]. La empresa de *marketing intelligence* Tractiva proyecta que el mercado creado en torno a estas técnicas crecerá de USD $3 mil millones en 2016 a USD $60 mil millones en 2026[16]. En efecto, en el área médica se proyecta un mercado de USD $19 mil millones en 2025[17].

El éxito comercial ha llevado a la diseminación y divulgación general de los conceptos de *Big Data, Data Science y Machine Learning*, entre otros. Estos conceptos dan forma a un marco que permite describir y discutir el funcionamiento, desempeño e impacto del análisis de datos a gran escala. En ocasiones, su uso se encuentra rodeado de sensacionalismo e hipérbole, lo que ofusca los desafíos, riesgos y compromisos involucrados. Estos aspectos deben ser parte central de la discusión, para la implementación responsable de esta metodología en un contexto clínico. Mientras que la hipérbole puede llevar a una decepción prematura frente a resultados modestos[4,18], una cautela excesiva puede ralentizar la adopción de técnicas que objetivamente pueden mejorar el cuidado de pacientes y la práctica clínica. Por lo tanto, para que esta metodología pueda revolucionar la disciplina médica, es necesario disponer de un marco conceptual donde sea posible discutir de manera objetiva sus resultados.

Este artículo tiene tres objetivos. Primero, definir los conceptos esenciales que permitan dar perspectiva a la discusión del uso de estos métodos en aplicaciones clínicas. Segundo, discutir los principales desafíos técnicos y conceptuales del análisis de datos a gran escala. Tercero, discutir de manera crítica algunas aplicaciones clínicas relevantes, reportadas en artículos científicos y en la prensa donde el desempeño de estos métodos es prometedor.

La estructura del artículo es la siguiente. Primero discutiremos las ideas principales detrás del concepto de *Big Data*. Luego discutiremos los conceptos fundamentales detrás de las técnicas de análisis de datos e identificaremos los criterios que permiten evaluar en la práctica el desempeño de estas técnicas. Haciendo uso de estos conceptos, discutiremos algunas aplicaciones clínicas relevantes y finalizaremos con una breve discusión acerca de los desafíos que *Big Data* presenta y como abordarlos.

2. EL DILUVIO DE DATOS

La capacidad de generar y almacenar datos se ha incrementado de manera exponencial en las últimas décadas y la medicina no es una excepción a este fenómeno. Este hecho considera los medios tradicionales de adquisición de datos, como imágenes radiológicas, fichas médicas y exámenes de laboratorios, pero también proyecta la adopción de *tecnologías vestibles*[19-21] que prometen adquirir señales fisiológicas, por ejemplo, cardíacas[22], en tiempo real. Es decir, la tasa de adquisición de datos clínicos se incrementará de forma considerable en un futuro cercano.

Esta cantidad masiva de datos, coloquialmente referida como *Big Data*, es parte de la metodología discutida previamente; los datos son la materia prima a partir de la que deseamos extraer información útil. Sin embargo, definir *Big Data* exclu-

sivamente en términos del *volumen* de los datos ofrece una visión parcial y limitada que no explica su potencial, ni evidencia los desafíos que presenta su manipulación. Por ello, es necesario considerar otras dimensiones al intentar caracterizar qué es *Big Data*. La *velocidad* y la *variedad* de los datos son dimensiones relevantes que complementan el *volumen*. La *velocidad* refiere tanto a la rapidez de generación de los datos, por ejemplo, señales fisiológicas adquiridas en *tiempo real* por sensores vestibles, como al tiempo en que el procesamiento de los datos debe ser realizado, por ejemplo, al correlacionar señales en tiempo real para determinar el riesgo de un paciente y así poder asignar recursos en una unidad de cuidado intensivo. La *variedad* refiere a la naturaleza diversa de los datos que se adquieren hoy en día, incluso de un mismo paciente, como por ejemplo imágenes radiológicas, pruebas de laboratorio, e información cualitativa presente en fichas médicas. En resumen, son el *volumen*, la *velocidad* y la *variedad*[23] de los datos que dan en parte origen a *Big Data* y elucidan los desafíos tecnológicos que presenta su manipulación y administración.

Sin embargo, es necesario complementar estas dimensiones para indicar aspectos de *Big Data* que van más allá de lo técnico. Usualmente se considera la veracidad de los datos como una dimensión de *Big Data,* que caracteriza en qué grado los datos reflejan una realidad objetiva, evitando errores sistemáticos o sesgos debido a factores humanos o técnicos. Las dimensiones de volumen, velocidad, variedad y veracidad dan origen a *"las cuatro Vs de Big Data"* propuestas por IBM[24] y proveen una heurística para determinar cuándo un régimen de generación y adquisición de datos constituye *Big Data*. Es necesario enfatizar que se trata de una heurística y no de una definición. Si bien el volumen, velocidad y variedad son dimensiones ampliamente aceptadas, también se han considerado como dimensiones adicionales el *valor*, esto es, la relevancia de la información que proveen los datos en el contexto en el que se generan, y la *variabilidad*, esto es, si los datos caducan y deben ser removidos o actualizados. Por lo tanto, los atributos denominados *"las 6 Vs de Big Data"*[25] son *volumen, velocidad* y *variedad*, que se refieren a la manipulación y administración de los datos, y *veracidad*, *variabilidad* y *valor*, que se refieren a la relevancia de los datos en el contexto del análisis que se quiere realizar. Estas 6 dimensiones proveen una completa caracterización del régimen de adquisición de datos que caracteriza *Big Data* en el contexto clínico y son las que consideraremos en este artículo.

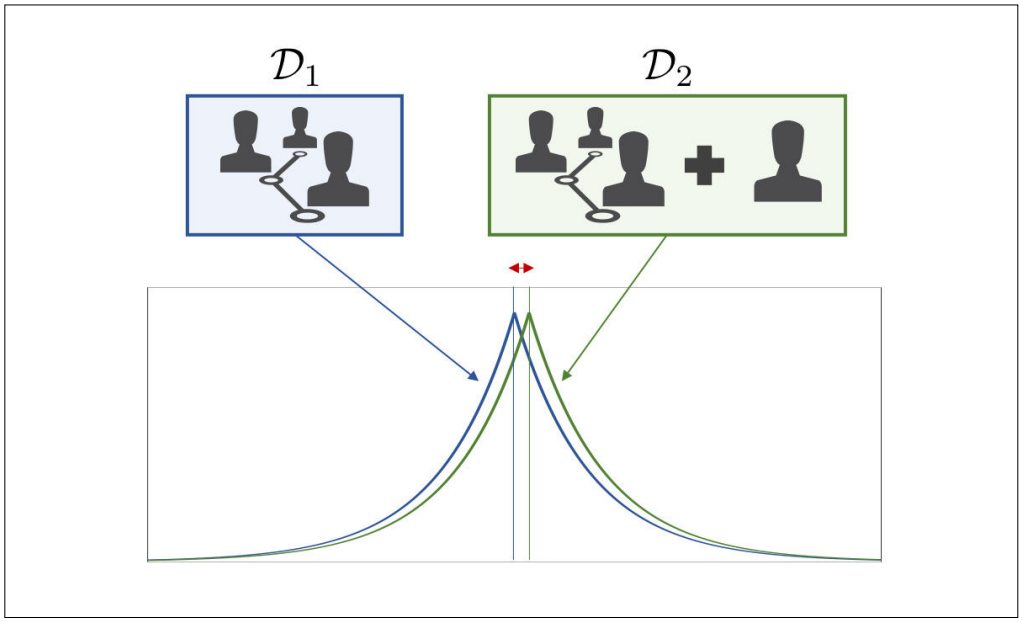
Estas dimensiones revelan los desafíos que Big Data presenta en el contexto clínico. Debido al *volumen* y la *velocidad* es necesario desarrollar una infraestructura computacional que permita almacenar y administrar los datos adquiridos de manera segura. Enfatizamos que esta infraestructura no sólo constituye un registro digital, sino que además debe garantizar el rápido acceso y procesamiento de los datos, permitiendo los múltiples análisis requeridos por la práctica e investigación clínica. Esto exige la creación de unidades informáticas asociadas a clínicas y hospitales, que velen por la mantención, organización y administración de estas bases de datos y diseñadas de acuerdo con los requerimientos específicos de *variedad* y *variabilidad* de los datos.

Además, es necesario velar por la privacidad de los pacientes y voluntarios involucrados. Este último punto es uno de los argumentos prevalentes en la discusión de los potenciales riesgos de *Big Data* en el contexto clínico. Un método de protección es anonimizar los datos, removiendo información que permita identificar individuos directamente; este es un método en uso hoy en día. Sin embargo, en el contexto de *Big Data, el volumen y la variedad* de datos de un mismo paciente podrían estar correlacionados con datos obtenidos de otras fuentes, como transacciones comerciales[26], revelando su identidad y dejándolo desprotegido. Afortunadamente, con una infraestructura de *Big Data* apropiada es posible analizar datos y proporcionar resultados agregados sin que el analista tenga acceso directo a la información de cada paciente; sólo el algoritmo computacional que implementa dicho análisis manipula los datos.

Aún en este caso, un individuo con intenciones nefastas, denominado *adversario*, podría tratar de inferir información personal a partir de múltiples análisis que entregan información agregada. La comunidad científica ha propuesto técnicas para evitar que esto ocurra: la *privacidad diferencial*[27,28] y las *bases de datos sintéticas*[29] son dos de ellas. Estos métodos utilizan técnicas matemáticas sofisticadas para reducir la probabilidad de éxito del *adversario*. La privacidad diferencial comprende el diseño de algoritmos en los que el efecto de los datos de sólo un individuo tiene un impacto pequeño; por tanto, nadie contribuye significativamente al resultado final, impidiendo al adversario inferir la identidad de un paciente dado (Figura 1). Las bases de datos sintéticas estiman la distribución estadística de los datos para reemplazar algunos de ellos por datos sintéticos o simulados; por tanto, *el adversario* es incapaz de distinguir si los datos son sintéticos o reales (Figura 2). Estas técnicas pertenecen a un área activa de investigación que, desafortunadamente, aún tienen una adopción limitada en comparación con técnicas tradicionales[30,31].

La protección de la privacidad de los pacientes ilustra un desafío importante, pero es sólo uno de los desafíos que *Big Data* presenta por sí mismo; como mencionamos, existen desafíos técnicos, relacionados con bases de datos, la ingeniería de software y la limpieza y mantención de datos,

Figura 1. Diagrama explicativo de la privacidad diferencial



En la ilustración se muestran dos bases de datos, denotadas D_1 y D_2, que se diferencian sólo por la presencia de los datos de un individuo adicional. El objetivo de la privacidad diferencial es diseñar de algoritmos cuyas conclusiones, representadas por las curvas en azul y verde, son similares en este caso. En otras palabras, la presencia del individuo en la base de datos D_2 no tiene un mayor impacto en las conclusiones obtenidas, lo que se ilustra a través del pequeño desplazamiento en la curva verde relativa a la azul.

Figura 2. Diagrama explicativo de las bases de datos sintéticas

En una primera etapa, se dispone de una base de datos con registros de individuos. En una segunda etapa, se selecciona al azar un subconjunto de estos registros para estimar las características estadísticas de ellos. En una tercera etapa, se sustituyen datos a través de simulaciones que preservan la estadística de la base de datos original. Este proceso se puede repetir varias veces para disminuir la probabilidad de identificación de cualquier individuo cuyos datos reales están en la base de datos.

además de desafíos organizacionales y culturales[32]. Presentar una lista exhaustiva esta fuera del alcance del presente artículo. A pesar de esto, las dimensiones de *Big Data* entregan al lector un sólido punto de partida para entender y discutir estos desafíos.

3. APRENDIENDO A PARTIR DE DATOS

Si *Big Data* es la materia prima, entonces su valor radica en la información y estructura que contiene. Las herramientas que permiten extraer esta información provienen de diversas disciplinas, tales como la *Ciencia de la Computación*, la Estadística y la Inteligencia Artificial, entre otras. La diversidad de fuentes de donde provienen estas herramientas ha dado origen a un nuevo campo científico interdisciplinario cuyo fin es desarrollar técnicas para extraer información a partir de datos. Esta disciplina, conocida como Ciencia de Datos (*Data Science* en inglés) es la segunda componente en la metodología de análisis de datos a gran escala.

Las dificultades encontradas al intentar definir *Big Data* persisten al intentar definir la Ciencia de Datos. Para efectos

del presente artículo, nos bastará concebir la Ciencia de Datos como la disciplina del "estudio científico de la creación, validación y transformación de datos para crear significado"(1)[33]. Para crear significado, lo más relevante es extraer información interpretable a partir de cantidades masivas de datos, por lo que nos enfocaremos en los métodos de *análisis* que usa la Ciencia de Datos, dejando de lado las dificultades, los desafíos, y las técnicas computacionales y algorítmicas asociadas a las tres primeras *Vs de Big Data*. La presentación y discusión tendrá un punto de vista estadístico, teniendo similitudes con la disciplina del *Aprendizaje Estadístico* de la que la Ciencia de Datos se nutre; ver, por ejemplo, (34-38) siendo referencias útiles para aprender más de esta disciplina[34-40].

Los métodos de la Ciencia de Datos ya son utilizados en algunas aplicaciones clínicas, que discutiremos más adelante, y es probable que su uso se incremente sustancialmente en un futuro cercano. Debido al entusiasmo, y muchas veces hipérbole, en torno a *Big Data* y la Ciencia de Datos, es necesario ser capaz de evaluar el desempeño de estos métodos en casos prácticos y en un contexto adecuado. Además, es importante que en un comienzo estos métodos no sean adoptados como *cajas negras* que simplemente reemplazan etapas del flujo de trabajo usual, sino como una fuente de información adicional que asista la toma de decisiones en cada etapa de dicho flujo. En otras palabras, se espera que en una primera etapa estos métodos no tomen decisiones por sí mismos, sino que asistan la toma de decisiones al proporcionar información que puede no ser evidente a partir de los datos disponibles y los análisis tradicionales.

En consecuencia, es fundamental que los profesionales de salud puedan generar una discusión acerca del rol que estos métodos tendrán en el futuro, de acuerdo con su desempeño y basada en criterios objetivos. Para este fin, es necesario hacer uso de un marco conceptual que dé forma a la discusión y permita identificar criterios adecuados de evaluación.

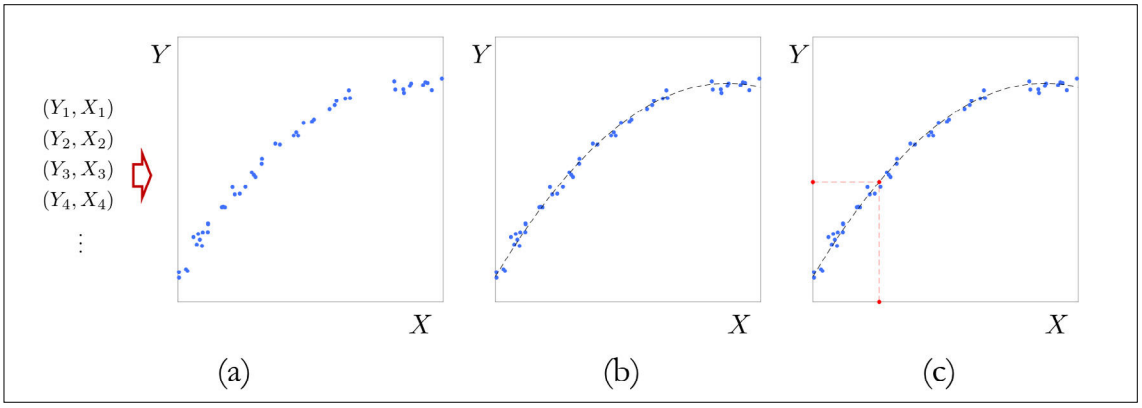
Este marco conceptual será guiado por las preguntas:

- ¿Cuál es el objetivo de analizar los datos?
- ¿Qué características tienen los métodos que alcanzan este objetivo?
- ¿Cómo evaluamos su desempeño?

3.1. Predicción e inferencia

El objetivo de analizar los datos depende fuertemente del contexto en el que se origina la *necesidad* del análisis. Estos objetivos suelen pertenecer a dos categorías [37]. En la primera, llamada *predicción*, los datos corresponden a registros históricos acerca del valor de una variable de interés, conocida como *variable de respuesta*, y los valores de múltiples variables que pueden predecir dicha respuesta, denominadas *variables predictoras*. El objetivo de la predicción es determinar el valor de la variable de respuesta para una colección de nuevos valores de las variables predictoras, distintos de aquellos que conocemos (Figura 3). Este es el caso, por ejemplo, cuando se dispone de registros de la sobrevida y el historial médico de pacientes que han sido sometidos a una intervención. Con estos datos, ¿es posible predecir la sobrevida de un paciente que es intervenido hoy dado su historial médico? Ejemplos concretos de estudios que constituyen objetivos de predicción

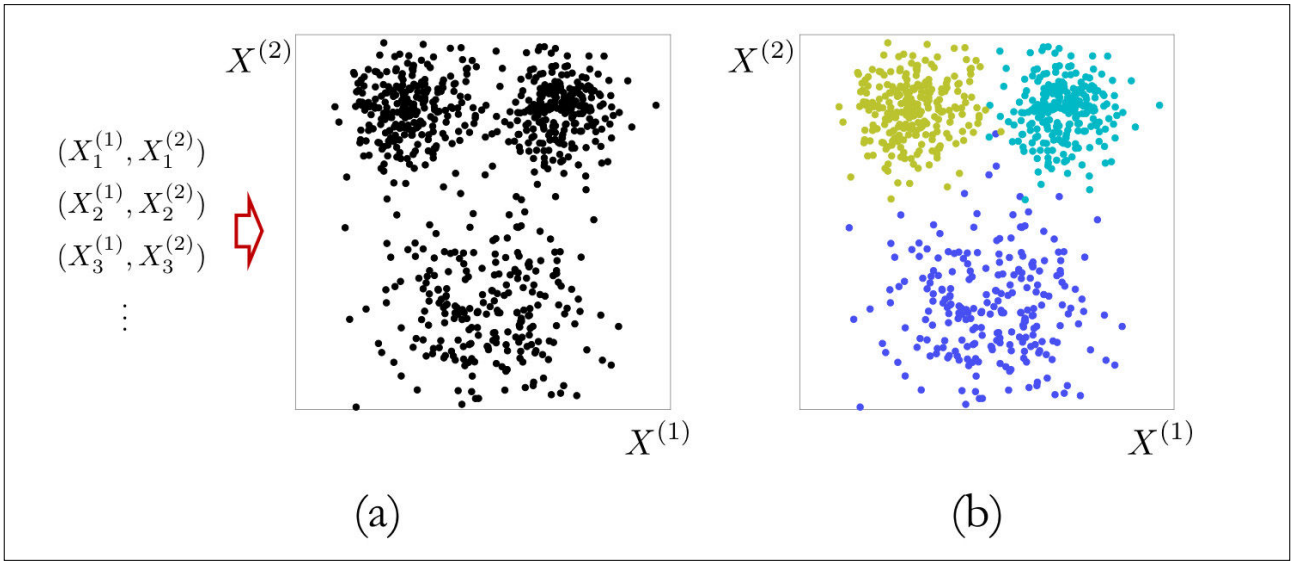
Figura 3. Diagrama explicativo de una tarea de predicción



Panel (a). Uno dispone de registros de variables predictoras, indicadas como *X*, y su correspondiente variable de respuesta, indicada como Y. El gráfico representa los datos, donde cada punto corresponde a un par ordenado (*X, Y*). Los datos sugieren una estructura que explica la relación entre la variable predictora y la variable de respuesta. Panel (b). Un método predictivo aproxima la relación entre la variable predictora y la variable de respuesta a partir de los datos disponibles. La función f obtenida por un método hipotético se grafica con una línea segmentada negra. Vemos que efectivamente $Y \approx f(X)$. Panel (c). Una vez que el método estima la relación entre la variable de respuesta y la variable predictora, podemos realizar una predicción para un valor nuevo, ilustrado en rojo en la abscisa, al evaluar la función f estimada en este valor, lo que entrega el valor en rojo en la ordenada.

(1) "The scientific study of the creation, validation and transformation of data to create meaning."

Figura 4. Diagrama explicativo de una tarea de inferencia



Panel (a). Uno dispone de registros de variables predictoras, indicadas como $X^{(1)}$ y $X^{(2)}$. En este caso no hay variable de respuesta. El gráfico representa los datos, donde cada punto corresponde a un par ordenado $(X^{(1)}, X^{(2)})$. En principio no pareciera haber una estructura evidente en los datos. Panel (b). Un método adecuado permite inferir que los datos se pueden agrupar en tres categorías, ilustradas con tres colores distintos. Notamos que en este caso no hay una predicción, el objetivo es simplemente extraer la estructura latente en los datos.

son estudios de diagnóstico clínico[38], genómica[39] y análisis de imágenes radiológicas[40, 41] entre otros.

En la segunda categoría, llamada *inferencia*, los datos corresponden a registros históricos de múltiples variables de interés, y el objetivo es determinar la relación que existe entre estas variables (Figura 4). Este es el caso, por ejemplo, cuando se dispone de registros históricos de pacientes con una misma patología, y se desea determinar si existen agrupaciones naturales de pacientes de acuerdo con dichas variables de interés. En la práctica, estudios de esta índole buscan mejorar la certeza tanto en diagnósticos clínicos[38,42,43] como radiológicos y de anatomía patológica[40,44,45], avanzar hacia una medicina personalizada, identificar poblaciones de riesgo para incluirlas en tamizajes poblaciones e identificar potenciales intervenciones de alto impacto en salud pública[46].

Estas categorías no son mutuamente excluyentes. Una vez propuesto un modelo predictivo para la sobrevida, es natural determinar qué variables del modelo tienen mayor poder predictivo, lo que constituye inferencia. De manera similar, luego de determinar grupos de pacientes con una cierta patología, puede ser de interés determinar modelos que predigan la progresión de ese paciente y a qué grupo pertenecería un paciente que ha sido diagnosticado hoy, lo que constituye predicción. Entonces, ¿por qué hacer una distinción?.

Los objetivos de predecir e inferir están íntimamente ligados con la *efectividad* y la *interpretabilidad* de un método. Existen métodos extremadamente efectivos para la predicción, como por ejemplo lo son las redes neuronales[51,52] o los bosques aleatorios[53]. Sin embargo, estos métodos son difíciles de interpretar; es difícil determinar cuáles son las variables que impactan la predicción, por lo que es complejo inferir relaciones entre las variables predictoras y la respuesta, o entre las variables predictoras. Por el contrario, métodos tradicionales para realizar inferencia, como la regresión multivariada[54,55], son fácilmente interpretables, pero suelen tener un peor desempeño que otras técnicas al realizar predicciones. Por tanto, determinar el objetivo como predicción o inferencia determina implícitamente el tipo de métodos que deseamos utilizar, y el compromiso entre desempeño e interpretabilidad que estamos dispuestos a asumir. Tener presente este efecto es fundamental al discutir lo idóneo de un método de la Ciencia de Datos aplicado al contexto clínico.

3.2. Aprendizaje supervisado y no supervisado

La primera etapa para determinar el tipo de metodología a utilizar involucra identificar si el objetivo corresponde a la predicción, inferencia o una combinación de ambos. La siguiente etapa involucra directamente los datos disponibles y los métodos seleccionados, por lo que requiere la profunda comprensión de la operación de estos métodos.

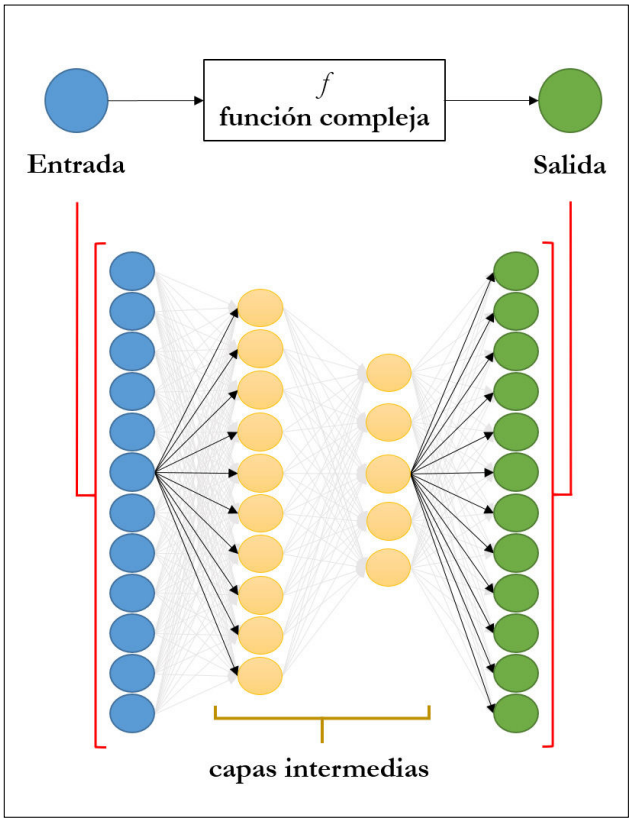
En el caso del problema de predicción, si denotamos Y a la variable de respuesta (variable dependiente) y X a las variables predictoras (variables independientes), entonces el método cuyo fin es realizar predicciones busca modelar matemáticamente la *información sistemática* que X propor-

ciona acerca del valor de Y. La relación entre la respuesta y los predictores puede ser descrita a través de una función f para la cual se tiene $Y \approx f(X)$; de esta forma, podemos realizar una predicción simplemente evaluando esta función para nuevos valores de las variables predictoras. Por lo tanto, el propósito de estos métodos es aprender a partir de los datos disponibles una buena aproximación de esta función f. Estos métodos se conocen como *supervisados* dado que los datos contienen tanto el valor de la respuesta como el valor de los predictores, y el proceso de aprendizaje se conoce también como *entrenamiento*[56].

Un caso concreto es el diagnóstico de melanoma a partir de imágenes dermatoscópicas[48]. En este caso las variables predictoras corresponden a las imágenes ya adquiridas, mientras que la variable de respuesta corresponde al diagnóstico, por ejemplo, 1 si es melanoma o 0 si no lo es. Dado que el objetivo es predicción, un método factible podría ser una red neuronal (ver Figura 5), pues determina de forma aproximada la relación entre las imágenes y el diagnóstico haciendo uso de datos históricos. Cuando queremos predecir el diagnóstico de un nuevo paciente, conceptualmente estamos evaluando la función f en esta nueva imagen dermatoscópica para predecir la presencia o ausencia de esta patología. Por lo tanto, lo relevante para evaluar estos métodos es comprender cómo la estructura de los datos históricos refleja de una buena forma la estructura de la relación que deseamos aprender. Volveremos a este punto en más detalle en la siguiente sección.

En el caso del problema de inferencia, en general no se dispone de variables de respuesta: el propósito es determinar si existe estructura en las variables que registran los datos. La clase de métodos que son utilizados en este caso se denominan *no supervisados*[58,59]. Un ejemplo clásico es el agrupamiento (*clustering* en inglés) (ver sección 3 en (58) o sección 3 en (59))[58,59], mediante el cual, dadas múltiples mediciones de variables continuas para múltiples pacientes, por ejemplo, la edad, el peso, y la presión arterial, se desea encontrar sugbrupos naturales de pacientes. Estos grupos pueden ser analizados posteriormente para verificar si coinciden, por ejemplo, con pacientes obesos o hipertensos, dándole así una interpreta-

Figura 5. Esquema de una red neuronal profunda



El objetivo es modelar la relación, que puede ser muy compleja, entre variables de entrada y de salida. Para ello, las componentes de la variable de entrada, que pueden ser las intensidades de los pixeles de una imagen, son utilizados como nodos de entrada. Cada nodo utiliza como entrada una combinación lineal del resto de los nodos de la capa anterior. La información se propaga a través de las capas intermedias hasta llegar a los nodos de salida. El número de capas y de nodos es determinado por el usuario. Los coeficientes de las combinaciones lineales son estimados a partir de datos.

ción clínica a *posteriori* a estos grupos. En general los métodos no supervisados son más complejos de analizar en la práctica, por lo que no los discutiremos con mayor detalle.

3.3. Muestreo, validación y error

La metodología que ha sido el foco de nuestra discusión se basa en el análisis de registros de datos existentes; en otras palabras, en el resultado de analizar lo que hemos observado hasta el presente. Esto tiene implicaciones concretas al evaluar los resultados de un método. Si nuevamente nos enfocamos en predicción, ¿cómo aseguramos que un método efectivamente puede predecir una respuesta a una observación futura en vez de limitarse a modelar las relaciones que existen sólo en los datos existentes?.

La promesa de *Big Data* es que el volumen de datos es tal, que contiene la visión completa del fenómeno en consideración. Esta es una proposición ambiciosa y difícil de verificar. Una técnica comúnmente utilizada para verificar que la información extraída es *generalizable*, es la *validación cruzada* (ver sección 1 en (34))[34] ésta también es útil para entender las limitaciones que existen al aprender a partir de datos. En la validación cruzada, los datos se separan en dos grupos al azar. El primero, llamado *entrenamiento*, se utiliza para aprender la relación que existe entre respuestas y predictores. Para verificar que es posible generalizar esta relación, se utiliza el segundo grupo de datos, llamado *prueba*, como un sustituto de los datos que observaremos en el futuro. De este modo, podemos determinar si la información extraída a partir del grupo de entrenamiento es generalizable al contrastarla con el grupo de prueba, proporcionando una métrica cuantitativa para evaluar el resultado del método propuesto.

Un ejemplo en el cual la adopción general de un modelo predictivo puede conducir a errores de predicción es el caso del estimador del *peak* estacional de influenza. En el año 2013 un equipo de investigadores de Google demostró que es posible predecir el *peak* estacional de influenza mediante el análisis de la frecuencia con que ciertos términos claves son buscados por los usuarios en la web[49]. El interés en el hallazgo reportado en esta publicación fue inmediato, pues los países destinan importantes recursos en centros centinelas con el objeto de predecir el *peak* estacional de influenza, de manera de asignar los recursos físicos y humanos a tiempo para dar respuesta a este *peak*. Sin embargo, el uso regular de este modelo en varios estados de EE.UU. y Europa ha demostrado que el modelo no es suficientemente preciso en su predicción[50]. Este "fracaso" ha abierto el debate respecto de la capacidad de generalizar modelos predictivos, el rol de la validación de estos modelos y la necesidad de calibrarlos continuamente para incluir nuevas fuentes de información.

Recapitulando, al evaluar si un método efectivamente extrae información *generalizable* a nuevos casos, es necesario ser cuidadoso al seleccionar e implementar técnicas de análisis adecuadas cuyos resultados hayan sido validados usando metodologías estadísticamente robustas. Cabe destacar que el resultado obtenido por estos métodos es siempre un reflejo de la cantidad y la calidad de la información que pudo ser extraída de los datos.

4. APLICACIONES

El marco conceptual presentado nos permite poner casos prácticos en perspectiva, y analizar de manera crítica algunos de los usos de *Big Data* y la Ciencia de Datos que han mostrado ser efectivos en aplicaciones clínicas.

Uno de los ejemplos que recientemente ha recibido atención, es el uso de una Red Neuronal Convolucional (RNC)[62] para diagnosticar melanoma a partir de imágenes dermatoscópicas[48]. Una de las causas de su impacto es que los dermatólogos fueron capaces de predecir correctamente el 86.6% de los melanomas, y el 71.3% de las lesiones benignas; a una misma sensibilidad, la especificidad de la RNC es mayor, con un 82.5% de las lesiones benignas correctamente diagnosticadas. Un artículo de prensa, titulado "La IA [N.T.: Inteligencia Artificial] supera a doctores en diagnóstico de cáncer"(2) nos indica que una RNC está "inspirada por los procesos biológicos que actúan cuando las células neurales (neuronas) en el cerebro se conectan unas con otras y responden a lo que el ojo ve."(3)[51] ¿Cómo un marco conceptual nos permite interpretar estos resultados?

Primero, el problema de diagnóstico es un problema de predicción; a partir de una imagen dermatoscópica, que constituye la variable predictora, se desea determinar si la lesión es maligna o no, lo que constituye la variable de respuesta. Nuestro marco conceptual nos indica que la RNC es, por tanto, un método que intenta aproximar la relación que existe entre la imagen y el estado de la lesión, maligno o benigno, a partir de diagnósticos efectuados en el pasado. En general, es sabido que una RNC es un método particularmente efectivo para problemas de predicción a partir de imágenes[52]. Al examinar el artículo, vemos que esta red fue *adaptada* para la detección de melanomas a partir de una red existente, entrenada para otras tareas, utilizando 100 mil imágenes digitales con su respectivo diagnóstico. Los resultados fueron evaluados utilizando una base de 100 dermatoscopías clasificadas por el método

(2) *"AI Beats Doctors at Cancer Diagnoses"*
(3) *"Inspired by the biological processes at work when nerve cells (neurons) in the brain are connected to each other and respond to what the eye sees"*

entrenado y por 57 dermatólogos, 17 de los cuales declaraban 2 años o menos de experiencia, 11 declaraban entre 2 y 5 años de experiencia, y 30 declaraban más de 5 años de experiencia.

Podemos detectar varios elementos del diseño experimental que pueden influenciar el desempeño del método y de los dermatólogos, y que nos permitan interpretar este resultado. Es importante evaluar siempre la relación que los datos de entrenamiento pueden tener con los datos de prueba. Por ejemplo, los datos de entrenamiento pueden tener sesgos, en términos de la presencia desbalanceada de las distintas lesiones en estos datos, que expliquen la alta especificidad de la RNC en comparación con los dermatólogos. Los autores reconocen que los datos de prueba "no muestran un rango completo de lesiones"(4) y que existe una "carencia de lesiones melanocíticas de otros tipos de piel y origen genético"(5) ¿Es posible que los resultados estén influenciados por la exposición de los dermatólogos a una mayor variedad de lesiones? Por otra parte, los dermatólogos pertenecen a 17 países distintos, y su entrenamiento clínico puede generar una exposición dispar a distintos tipos de lesiones; el grupo es además heterogéneo respecto a la experiencia profesional de sus integrantes. Sin mayores detalles acerca de las características de los datos de entrenamiento utilizados, es difícil determinar que otros efectos pueden jugar un rol. Si bien no podemos dar respuesta a estas preguntas, este ejemplo ilustra cómo el marco conceptual presentado nos permite identificar elementos que se deben considerar al interpretar las notas de prensa tras un estudio de este tipo.

Un estudio realizado por Weng *et al* en 2017[53], hace uso de técnicas de análisis de datos para predecir riesgo cardiovascular a partir de fichas médicas de pacientes. La nota de prensa indica que "computadores que se pueden enseñar a sí mismos pueden desempeñarse aún mejor que guías médicas estándar, incrementando significativamente las tasas de predicción"(6) . Nuestro marco permite discutir este estudio en un contexto adecuado. Al revisar el artículo, verificamos que 8 variables de riesgo medidas en 378256 individuos en el Reino Unido fueron utilizadas para predecir el diagnóstico del primer evento cardiovascular. Vemos además la validación efectuada: el 75% de los datos fueron utilizados para entrenar los métodos considerados en el estudio, mientras que el 25% de los datos fueron utilizados como prueba. Estos métodos fueron comparados con la predicción obtenida a partir de las guías del Colegio Americano de Cardiología (ACC por sus siglas en inglés, *American College of Cardiology*) y la Sociedad Americana del Corazón (AHA por sus siglas en inglés, *American Heart Association*) que hacen uso de estas mismas 8 variables para predecir el riesgo de un evento cardiovascular. En general, los métodos de análisis de datos utilizados presentan un incremento en la tasa de predicción de eventos cardiovasculares por sobre las guías médicas utilizadas hoy en día. Este estudio ilustra algunos elementos importantes. Si bien los métodos fueron entrenados en datos, las variables de riesgo son las mismas determinadas por las guías de la ACC y AHA; el desempeño se puede deber a que los métodos capturan correlaciones entre estas variables que no son obvias, algo que los autores discuten en su artículo. También, mientras que las guías médicas son definidas a través de un comité especializado y tienen por objetivos ser fáciles de implementar en la clínica diaria y persistir en el tiempo, los métodos de análisis de datos pueden ser actualizados día a día, en la medida que nuevos datos estén disponibles y procedimientos de validación adecuados hayan sido efectuados.

Esto nos lleva a discutir el grado de especialización de estos algoritmos a los datos utilizados como entrenamiento. ¿Tienen estas bases de datos algo *particular*? ¿Se aplican las relaciones aprendidas a partir de estos datos a la población general? Como mencionamos, una primera medida para evitar una sobre-especialización es el uso de validación cruzada. Sin embargo, ¿qué tan precisa es esta técnica para estimar el error de generalización?.

Un análisis importante en esta dirección lo constituye el trabajo de Bernau *et al*[10]. En él, los autores diseñan un método de validación entre diferentes bases de datos; en otras palabras, un método que utiliza datos adquiridos por varios grupos de investigación para ser utilizados en distintos estudios. Esta técnica, que denominan *validación cruzada entre estudios* (7) permitiría no sólo una evaluación más efectiva de los métodos reportados por la comunidad científica, sino que una validación *continua* de los mismos en la medida que más datos se encuentren a disposición del público. Además, la validación que consideraría bases de datos adquiridas por distintos grupos de investigación debería reflejar de mejor forma la variabilidad natural que ocurre cuando estos métodos son adoptados en la práctica clínica. El resultado de este trabajo es claro. Estos "...sugieren que la validación cruzada estándar produce una sobreestimación de la precisión de discriminación para todos los algoritmos considerados, en comparación con la validación cruzada entre estudios"(8).

(4) "...the test-sets of our study did not display the full range of lesions."
(5) "...the poor availability of validated images led to a shortage of melanocytic lesions from other skin types and genetic backgrounds."
(6) "...computers capable of teaching themselves can perform even better than standard medical guidelines, significantly increasing prediction rate."
(7) *Cross-study validation en inglés.*
(8) "...suggest that standard cross-validation produces inflated discrimination accuracy for all algorithms considered, when compared to cross-study validation."

En resumen, es importante tener presente la doble responsabilidad que tendrán las instituciones clínicas que utilicen métodos basados en el análisis de datos: por un lado, adoptar métodos estrictamente validados y, por otra parte, validarlos debidamente en sus propias bases de datos, para asegurar una alta calidad, precisión y reproducibilidad de los resultados.

5. DISCUSIÓN

Probablemente la mejor manera de dimensionar el potencial impacto futuro de *Big Data* en medicina consiste en reflexionar sobre el impacto que ha tenido, y sigue teniendo, el estudio Framingham [54]. Este proyecto estableció el seguimiento de una cohorte de 5209 hombres y mujeres sanos entre 30 y 62 años en la ciudad de Framingham en Massachusetts, Estados Unidos. El seguimiento se inició en 1948 y hoy continúa siguiendo a la tercera generación de los participantes originales. Las conclusiones que se han obtenido de este estudio han sentado las bases fisiopatológicas y terapéuticas de muchas enfermedades cardiovasculares y nutricionales entre otras [55,56]. Inspirados por este estudio, consideremos el volumen de datos que se obtendrían a partir de la información clínica de los egresos hospitalarios chilenos: 1.7 millones de egresos al año (DEIS 2015) más toda la información clínica de las consultas ambulatorias, que sólo en el sector público suman 10.8 millones al año (DEIS 2014) y todos los registros de defunciones y causas de muerte (103327 casos anuales, DEIS 2015). La información que se podría extraer supera en órdenes de magnitud aquella obtenida de un estudio de la escala del estudio Framingham. En la literatura se ha sugerido además que sistematizar toda esta información y contar con registros clínicos electrónicos permitiría extraer información similar a la que se obtiene desde estudios aleatorizados y meta análisis [57].

Sin duda que el futuro de la unión de la Ciencias de Datos y la medicina es promisorio, pero ¿cuánto de esto resultará en una mejor salud para los pacientes o en una drástica transformación de la profesión médica? Diversos autores han reflexionado sobre los cambios que se avecinan[38,58-60] y el consenso es que éstos serán profundos y significativos. Para estimar su impacto, podemos utilizar como referente la revolución que constituyó en las últimas décadas la tecnificación de la medicina por sobre el *arte de la medicina*[59]. Esta última revolución sin duda modificó el rol de los equipos médicos, impactó la forma en que se educan y entrenan los profesionales de la salud, generó nuevas necesidades, y mejoró significativamente la calidad de vida de la población. Del mismo modo, hoy nos encontramos en una etapa en la que es necesario responder a nuevas necesidades de cómo realizar investigación y cómo educar a los profesionales médicos del futuro.

Un ejemplo de iniciativas que responden a esta necesidad en investigación en Chile es el *Centro de Imágenes Biomédicas* de la Pontificia Universidad Católica de Chile. En este centro de investigación interdisciplinario, que depende del Departamento de Radiología, el Departamento de Ingeniería Eléctrica y del Instituto de Ingeniería Biológica y Médica, se desarrollan técnicas de radiología cuantitativa, cuyo fin es transformar la información contenida en imágenes radiológicas en *métricas* precisas y reproducibles. Por ejemplo, a partir de imágenes de flujo adquiridas por resonancia magnética es posible desarrollar modelos físicos que permiten caracterizar el comportamiento hemodinámico a alta resolución. Estas métricas pueden ser analizadas utilizando las técnicas discutidas a lo largo de este artículo con el fin encontrar biomarcadores que detecten de forma temprana situaciones de riesgo. Gracias al carácter interdisciplinario de este Centro, la visión de médicos, ingenieros, matemáticos y estadísticos forman parte de la motivación, desarrollo, implementación y análisis de estas nuevas metodologías.

6. CONCLUSIÓN

El marco conceptual delineado en este artículo permite generar una discusión en torno a aplicaciones del análisis de datos masivos en datos clínicos reportados en la prensa y en la literatura científica. Esta discusión informada es un primer paso para facilitar la amplia adopción de estas técnicas en la práctica clínica y para que los profesionales de la salud no sólo sean *generadores de datos*, sino que, con el apoyo de las tecnologías emergentes, los datos, con su correcto análisis e interpretación, apoyen a los equipos hacia mejores decisiones médicas. Esto no implica que los profesionales de la salud se conviertan en *cientistas de datos*. Por el contrario, es un llamado a crear grupos interdisciplinarios al interior de hospitales, clínicas y escuelas de medicina, que permitan a los profesionales de la salud familiarizarse con las nuevas técnicas de análisis desarrolladas, y a su vez permitan a los profesionales que desarrollan dichas técnicas familiarizarse con las inquietudes y desafíos que enfrentan los profesionales clínicos. Este es el único camino que garantiza que el desarrollo de estas técnicas computacionales en medicina evolucione en una dirección que beneficie a los pacientes y a los usuarios de los sistemas de salud. El futuro del uso de *Big Data* en la medicina es brillante, pero no exento de desafíos a los que debemos responder hoy.

Declaración Conflicto de Interés
Financiamiento: C.A.S.L. fue parcialmente financiado por un Fondecyt de Iniciación # 11160728. M.A. fue parcialmente financiado por el proyecto Fondecyt #1180525.

REFERENCIAS BIBLIOGRÁFICAS

1. *The Data Deluge. The Economist [Internet]. 25 de febrero del 2010 [citado el 1 de agosto del 2018]. Recuperado de: https://www.economist.com/leaders/2010/02/25/the-data-deluge*
2. *The data deluge. Nature Cell Biology [Internet]. Nature Publishing Group. 1 de agosto del 2012 [citado el 1 de agosto del 2018]; 14:775. DOI: http://dx.doi.org/10.1038/ncb2558*
3. *Reed DA, Dongarra J. Exascale Computing and Big Data. Communications of the ACM. New York, NY, USA: ACM. Julio 2015; 58(7):56-68.*
4. *Mehta N, Pandit A. Concurrence of Big Data analytics and healthcare: A systematic review. International Journal of Medical Informatics. 2018; 114:57-65.*
5. *Jaret P. Mining Electronic Records for Revealing Health Data. The New York Times [Internet]. 14 de enero del 2013 [citado el 1 de agosto del 2018]; Recuperado de: https://www.nytimes.com/2013/01/15/health/mining-electronic-records-for-revealing-health-data.html*
6. *Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. New England Journal of Medicine. 2016; 375(13):1216-1219.*
7. *Becker RA. Dealing with the Health Data Deluge. PBS [Internet]. 20 de mayo del 2015 [citado el 1 de agosto del 2018]. Recuperado de: http://www.pbs.org/wgbh/nova/next/body/health-data/*
8. *Quinodoz M, Royer-Bertrand B, Cisarova K, Di Gioia SA, Superti-Furga A, Rivolta C. DOMINO: Using Machine Learning to Predict Genes Associated with Dominant Disorders. The American Journal of Human Genetics. 2017; 101(4):623-629.*
9. *Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discovery Today. 2018; 23(6):1241-1250.*
10. *Bernau C, Riester M, Boulesteix A-L, Parmigiani G, Huttenhower C, Waldron L, et al. Cross-study validation for the assessment of prediction algorithms. Bioinformatics. 15 de junio del 2014; 30(12):i105-i112.*
11. *Cakebread C. Google and Facebook dominate digital advertising - and they now account for 25% of all ad sales, online or off. Business Insider [Internet]. 7 de diciembre del 2017 [citado el 1 de agosto del 2018]; Recuperado de: https://www.businessinsider.com/google-and-facebook-dominate-the-world-of-online-advertising-charts-2017-12*
12. *Tran TP. Personalized ads on Facebook: An effective marketing tool for online marketers. Journal of Retailing and Consumer Services. 2017; 39:230-42.*
13. *Plummer L. This is how Netflix's top-secret recommendation system works. Wired UK [Internet]. 22 de agosto del 2017 [citado el 1 de agosto del 2018]; Recuperado de: https://www.wired.co.uk/article/how-do-netflixs-algorithms-work-machine-learning-helps-to-predict-what-viewers-will-like*
14. *Gomez-Uribe CA, Hunt N. The Netflix Recommender System: Algorithms, Business Value, and Innovation. ACM Transactions on Management Information Systems. New York, NY, USA: ACM; Diciembre del 2015; 6(4):13:1-13:19.*
15. *Smith B, Linden G. Two Decades of Recommender Systems at Amazon.com. IEEE Internet Computing. 2017; 21(3):12-8.*
16. *Artificial Intelligence Software Revenue to Reach $59.8 Billion Worldwide by 2025. Tractica [Internet]. 2 de mayo del 2017 [citado el 1 de agosto del 2018]; Recuperado de: https://www.tractica.com/newsroom/press-releases/artificial-intelligence-software-revenue-to-reach-59-8-billion-worldwide-by-2025/*
17. *Healthcare Artificial Intelligence, Software, Hardware, and Services Market to Reach $19.3 Billion Worldwide by 2025. Tractica [Internet]. 25 de septiembre del 2017 [citado el 1 de agosto del 2018]; Recuperado de: https://www.tractica.com/newsroom/press-releases/healthcare-artificial-intelligence-software-hardware-and-services-market-to-reach-19-3-billion-worldwide-by-2025/*
18. *Peniewski P. The AI winter is well on its way. VentureBeat [Internet]. 4 de junio del 2018 [citado el 1 de agosto del 2018]; Recuperado de: https://venturebeat.com/2018/06/04/the-ai-winter-is-well-on-its-way/*
19. *Bhelkar V, Shedge DK. Different types of wearable sensors and health monitoring systems: A survey. In: 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT). 2016. p. 43-8.*
20. *Miyamoto A, Lee S, Cooray NF, Lee S, Mori M, Matsuhisa N, et al. Inflammation-free, gas-permeable, lightweight, stretchable on-skin electronics with nanomeshes. Nature Nanotechnology. 17 de julio del 2017; 12:907.*
21. *Pattani A. A Sensor on Your Skin That Looks and Feels Like a Temporary Tattoo. The New York Times [Internet]. 20 de julio del 2017 [citado el 1 de agosto del 2018]; Recuperado de: https://www.nytimes.com/2017/07/20/health/breathable-wearable-sensor-temporary-tattoo.html*
22. *Lee SP, Ha G, Wright DE, Ma Y, Sen-Gupta E, Haubrich NR, et al. Highly flexible, wearable, and disposable cardiac biosensors for remote and ambulatory monitoring. npj Digital Medicine. 2018; 1(1):2.*
23. *Laney D. 3-D Data Management: Controlling Data Volume, Velocity and Variety. META Group [Internet]. 6 de febrero del 2001 [citado el 1 de agosto del 2018]. Recuperado de: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf*
24. *The Four V's of Big Data. IBM Big Data & Analytics Hub [Internet]. [citado el 1 de agosto del 2018]; Recuperado de: http://www.ibmbigdatahub.com/infographic/four-vs-big-data*
25. *Eapen BA. The 6 V's of Big Data [online]. 7 de abril del 2017 [citado el 1 de agosto del 2018]; Recuperado de: https://community.mis.temple.edu/mis520817/2017/04/07/the-6-vs-of-big-data/*
26. *De Montjoye Y-A, Radaelli L, Singh VK, Pentland AS. Unique in the shopping mall: on the reidentifieability of credit card metadata. Science. 205; 347(6221):536-539.*
27. *Dalenius T. Towards a methodology for statistical disclosure control. Statistik Tidskrift. 1977; 15:2-1.*
28. *Dwork C. Differential Privacy. In: Bugliesi M, Preneel B, Sassone V, Wegener I, editors. ICALP 2006: Automata, Languages and Programming. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. p. 1-12.*
29. *Reiter JP. New Approaches to Data Dissemination: A Glimpse into the Future (?). CHANCE. Taylor & Francis; 1 de junio del 2004; 17(3):11-5.*
30. *Dankar FD, El Emam K. Practicing Differential Privacy in Health Care: A Review. Transactions in Data Privacy. 2013; 6(1):35-67.*
31. *Abowd JM, Vilhuber L. How Protective Are Synthetic Data? In: Domingo-Ferrer J., Saygın Y. (eds) Privacy in Statistical Databases. PSD 2008. Lecture Notes in Computer Science, vol 5262. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008. p. 239-246.*
32. *Barlow M. The Culture of Big Data. O'Reilly Media Inc. 2013.*
33. *Code of Conduct. Data Science Association. Recuperado de: http://www.datascienceassn.org/code-of-conduct.html [citado el 1 de agosto del 2018].*
34. *James G, Witten D, Hastie T, Tibshirani R. Introduction. En: James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning - with applications in R. Springer Texts in Statistics. New York: Springer-Verlag New York; 2013. p. 1-14.*
35. *James G, Witten D, Hastie T, Tibshirani R. Statistical Learning. En: James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning - with applications in R. Springer Texts in Statistics. New York: Springer-Verlag New York; 2013. p. 15-58.*
36. *Ethem A. Why We Are Interested in Machine Learning. En: Ethem A. Machine Learning: The New AI. The MIT Press Essential Knowledge. Massachussets: MIT Press; 2016. p. 1-28.*

37. *Ethem A. Machine Learning, Statistics, and Data Analytics. En: Ethem A. Machine Learning: The New AI. The MIT Press Essential Knowledge. Massachussets: MIT Press; 2016. p. 29-54.*
38. *Hastie T, Tibshirani R, Friedman J. Introduction. En: Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning - Data Mining, Inference, and Prediction. Springer Texts in Statistics. 2da edición. New York: Springer-Verlag New York; 2009. p. 1-8.*
39. *James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning - with applications in R. Springer Texts in Statistics. New York: Springer-Verlag New York. 2013.*
40. *Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning - Data Mining, Inference, and Prediction. Springer Texts in Statistics. 2da edición. New York: Springer-Verlag New York. 2009.*
41. *Breiman L. Statistical Modeling: The Two Cultures. Statistical Science. 2001; 16(3):199-231.*
42. *Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. The New England Journal of Medicine. 2016; 375(13):1216-1219.*
43. *He KY, Ge D, He MM. Big Data Analytics for Genomic Medicine. International Journal of Molecular Sciences. 2017; 18(2):412.*
44. *Jha S, Topol EJ. Adapting to Artificial Intelligence: Radiologists and Pathologists as Information Specialists. The Journal of the American Medical Association. 2016; 316(22):2353-2354.*
45. *Kansagra A P, Yu J-P J, Chatterjee A R, Lenchik L, Chow D S, Prater A B, et al. Big Data and the Future of Radiology Informatics. Academic Radiology. 2016; 23(1):30-42.*
46. *Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial Intelligence in Precision Cardiovascular Medicine. Journal of the American College of Cardiology. 2017; 69(21):2657-2664.*
47. *Lisboa PJ, Taktak AFG. The use of artificial neural networks in decision support in cancer: a systematic review. Neural Networks: The Official Journal of the International Neural Network Society. 2006; 19(4):408-415.*
48. *Kaymak S, Helwan A, Uzun D. Breast cancer image classification using artificial neural networks. Procedia Computer Science. 2017; 120, 126-131.*
49. *Shiraishi J, Li Q, Appelbaum D, Doi K. Computer-aided diagnosis and artificial intelligence in clinical imaging. Seminars in Nuclear Medicine. 2011; 41(6):449-462.*
50. *Khoury MJ, Ioannidis JPA. Medicine. Big data meets public health. Science. 2014; 346(6213):1054-1055.*
51. *Hastie T, Tibshirani R, Friedman J. Neural Networks. En: Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning - Data Mining, Inference, and Prediction. Springer Texts in Statistics. 2da edición. New York: Springer-Verlag New York; 2009. p. 389-416.*
52. *Goodfellow I, Bengio Y, Courville A. Deep Networks: Modern Practices. En: Goodfellow I, Bengio Y, Courville A. Deep Learning. Adaptive Computation and Machine Learning. MIT Press; 2016. p. 161-474.*
53. *Hastie T, Tibshirani R, Friedman J. Random Forests. En: Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning - Data Mining, Inference, and Prediction. Springer Texts in Statistics. 2da edición. New York: Springer-Verlag New York; 2009. p. 587-604.*
54. *James G, Witten D, Hastie T, Tibshirani R. Linear Regression. En: James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning - with applications in R. Springer Texts in Statistics. New York: Springer-Verlag New York; 2013. p. 59-126.*
55. *Hastie T, Tibshirani R, Friedman J. Linear Methods for Regression. En: Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning - Data Mining, Inference, and Prediction. Springer Texts in Statistics. 2da edición. New York: Springer-Verlag New York; 2009. p. 43-100.*
56. *Hastie T, Tibshirani R, Friedman J. Overview of Supervised Learning. En: Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning - Data Mining, Inference, and Prediction. Springer Texts in Statistics. 2da edición. New York: Springer-Verlag New York; 2009. p. 9-42.*
57. *Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Annals of Oncology [Internet]. 28 de mayo del 2018; mdy166. 58. James G, Witten D, Hastie T, Tibshirani R. Unsupervised Learning. En: James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning - with applications in R. Springer Texts in Statistics. New York: Springer-Verlag New York; 2013. p. 373-418.*
59. *Hastie T, Tibshirani R, Friedman J. Unsupervised Learning. En: Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning - Data Mining, Inference, and Prediction. Springer Texts in Statistics. 2da edición. New York: Springer-Verlag New York; 2009. p. 485-586.*
60. *Butler D. When Google got flu wrong [Internet]. Nature. 13 de febrero del 2014; 494(7436):155-156. Recuperado de: https://www.nature.com/news/when-google-got-flu-wrong-1.12413*
61. *Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009; 457(7232):1012-1014.*
62. *Goodfellow I, Bengio Y, Courville A. Convolutional Networks. En: Goodfellow I, Bengio Y, Courville A. Deep Learning. Adaptive Computation and Machine Learning. MIT Press; 2016. p. 321-362.*