Nutrición Clínica en Medicina | Vol. XIII - Número 3 - 2019
pp. 140-152

www.nutricionclinicaenmedicina.com

[artículo especial]

Big Data y Medicina

Eduardo de Teresa Galván

Cátedra de Terapias Avanzadas en Patología Cardiovascular. Universidad de Málaga. España.

Palabras clave

Big Data, causalidad, investigación

>>Resumen

El Big Data, entendiendo como tal el análisis de ingente cantidad de datos en un tiempo limitado por medio de técnicas basadas en inteligencia artificial, ha comenzado a aplicarse al campo de la Medicina. Este tipo de análisis va a inducir una serie de cambios en la Medicina, que incluyen un cambio conceptual desde las hipótesis basadas en relaciones de causalidad hasta las correlaciones por coincidencia. Además, pone de manifiesto el contraste entre el derecho a la privacidad de los datos y el bien común derivado de un mayor conocimiento de la enfermedad. Este conflicto está ya dando lugar a modificaciones en la legislación al respecto. Además, es posible que modifique la forma en que se valoran los currículos de los investigadores, otorgando un mayor reconocimiento a los que aportan sus datos, y no sólo a los que los analizan. Existe, por otra parte, una escasez de profesionales preparados para las técnicas de análisis necesarias, lo que supone una limitación para el campo de la investigación biomédica, que deberá competir de forma desfavorable con la demanda de dichos técnicos por parte de empresas y organizaciones gubernamentales. Todos estos factores aconsejan que los médicos se impliquen en el conocimiento de los mecanismos y posibilidades de Big Data con objeto de orientar, en su momento, su desarrollo para conseguir la aplicación al campo de la Medicina de modo que se obtenga el máximo beneficio para los pacientes y la mínima agresión para los profesionales de la sanidad.

Nutr Clin Med 2019; XIII (3): 140-152
DOI: 10.7400/NCD.2019.13.3.5079

Key words

Big Data, causality, research

>> Abstract

Big Data, understanding as such the analysis of huge amount of data in a limited period of time, has started to be applied to the field of Medicine. This type of analysis will induce a series of changes in Medicine, including a conceptual change from hypotheses based on causal relationships to correlations by coincidence. It also highlights the contrast between the right to data privacy and the benefits derived from a better knowledge of health and disease. Legislation modifications are already being introduced to deal with this conflict. On the other hand, Big Data may induce changes in the way in which researchers' curricula are evaluated, giving greater recognition to those who contribute their data, and not only to those who analyze them. However, there is a shortage of professionals prepared for the necessary analysis techniques, which implies a limitation for the field of biomedical research, which will have to compete unfavorably with the demand of such technicians by private corporations and government organizations. All these facts suggest that doctors should be involved in the knowledge of the mechanisms and

Correspondencia

Eduardo de Teresa Galván.
Email: eduardodeteresa@gmail.com

possibilities of Big Data, in order to guide in the future its development to assure that the application to the field of Medicine derives the maximum benefit.

Nutr Clin Med 2019; XIII (3): 140-152
DOI: 10.7400/NCD.2019.13.3.5079

>>Introducción

El siglo XVIII vivió una época de extraordinario optimismo en la ciencia. Después de Newton se podían por fin aplicar fórmulas matemáticas exactas al comportamiento de la realidad. El mundo parecía funcionar como un reloj. En ese ambiente, Pierre Simon Laplace (1749-1927) formuló la siguiente idea: "Si, en un momento dado, pudiésemos conocer la posición de todos los cuerpos y las fuerzas que actúan sobre ellos y, más aún, si dispusiésemos de un intelecto lo suficientemente potente para analizar todos estos datos en un corto período de tiempo, podríamos conocer con exactitud lo que sucederá en el futuro". En esta proposición se contienen gran parte de los componentes de lo que, a falta de una traducción apropiada y aceptada, denominaremos Big Data. Por desgracia para el determinista Laplace, el desarrollo posterior de la ciencia vino a demostrar que el mundo no se comporta como un reloj, y el principio de indeterminación de Heisenberg confirmó que no podemos conocer todo al mismo tiempo. La naturaleza de la realidad —al menos de la realidad cuántica— es probabilística, lo que impide hacer predicciones precisas.

Las predicciones del futuro son problemáticas, y ni siquiera un análisis correcto de la realidad actual garantiza que dichas predicciones se cumplan. Marx realizó un notable análisis de la situación de la economía capitalista en el siglo XIX, pero fracasó al aventurar lo que sucedería en el futuro. Entre las posibles razones para este fracaso hay una clara: Los agentes económicos, entre otros los propios capitalistas, conocieron los trabajos de Marx y, de una forma u otra, supieron adaptarse para modificar lo previsible. Quizá por eso Isaac Asimov introduce, en la saga de La Fundación, ciertas leyes para la Psicohistoria. Esta ciencia, supuestamente desarrollada por el matemático Hari Seldon en el siglo CXI de la Era Galáctica, permite predecir el futuro basándose en lo que hoy conocemos como Big Data, pero desde una perspectiva probabilística. Para ello debe seguir dos normas básicas: La cantidad mínima de datos —individuos— a considerar para hacer predicciones es de 75 millones; y la población debe permanecer ignorante de las predicciones de la Psicohistoria. Asimov tiene en cuenta el error de Marx; pero a pesar de ello, su ciencia también acaba fallando debido al factor imprevisible y recurrente: el factor humano.

Pero volvamos a Laplace. En su formulación hace referencia a dos aspectos: conocer todos —o muchos de— los datos, y poder analizar esa ingente cantidad de información en un período corto de tiempo.

En cuanto a los datos, el panorama desde el cambio de siglo es sorprendente. A lo largo de los milenios de la historia humana, se han ido acumulando cantidades ingentes de información de manera más o menos constante y progresiva, aunque ese proceso se ha acelerado de forma espectacular en los últimos tiempos; por ejemplo, la cantidad total de información disponible se multiplicó por cien entre 1987 y 2007[1], y a partir del año 2000 la explosión de datos hace que dupliquemos nuestra información total en cuestión de meses (fig. 1); incluso algunos datos sugieren que el 90% de la información de que disponíamos en 2017 se había generado en los dos años previos[2].

Hasta el cambio de siglo, la mayor parte de los datos estaban almacenados en forma analógica; pero desde entonces la situación se ha invertido, de forma que en 2015 apenas un 1% se almacena de esa forma. La mayor parte de los datos se encuentran hoy en formato digital, puesto que gran parte procede del tráfico de internet, lo que facilita su almacenamiento y procesado. Disponemos hoy, pues, de una ingente cantidad de datos que se multiplica rápidamente y, lo que reviste gran importancia, en formato digital.

Pero ¿de dónde procede tal cantidad de datos? Es sencillo: en gran parte de nosotros mismos. Nos hemos acostumbrado a emplear instrumentos que nos son proporcionados de forma gra-

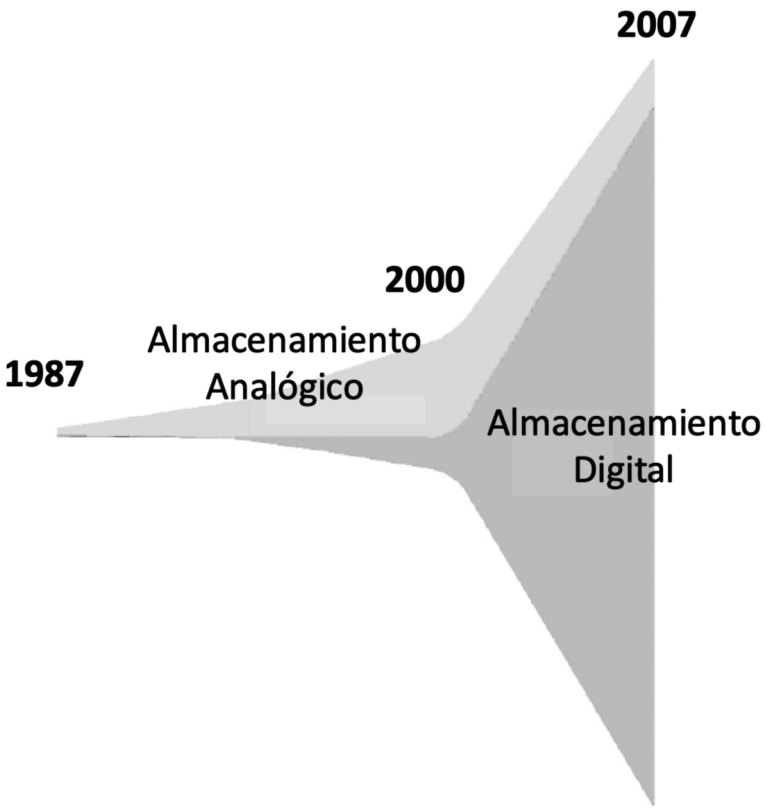

Figura 1. Crecimiento reciente de los datos totales acumulados por la Humanidad. Modificado de Ref. 12.

tuita, como el correo electrónico, Facebook, Instagram, Twitter…Ahora bien ¿gratuita? Realmente no. Estamos pagándolos con información sobre nosotros mismos, nuestros gustos, tendencias, amistades, etc. Y esa información es valiosa y útil para una serie de fines. Hace relativamente poco tiempo surgió un escándalo por el uso que una empresa —Cambridge Analytics— había dado a datos de 50 millones de usuarios de Facebook, a los que había accedido con fines supuestamente de investigación. Dichos datos fueron empleados para influir en la campaña presidencial americana, aprovechando los perfiles de preferencias de los usuarios para dirigir la información. Según las declaraciones de un empleado de la empresa en la investigación posterior, "*Explotamos Facebook para acceder a millones de perfiles de usuarios. Y construimos modelos para explotar lo que sabíamos de ellos y apuntar a sus demonios internos. Esa era la base sobre la cual la compañía se fundó*"[3].

Otras veces las instancias que consultan esos datos son más elevadas. El Sínodo de Obispos, que se celebró en Octubre de 2018 en el Vaticano, encargó un estudio para conocer a quién seguían los jóvenes con inquietudes espirituales[4]. El estudio, que llevó a cabo un grupo de investigación de la Universidad Ramon Llull, analizó datos de 540 millones de perfiles de las redes sociales (parece ser que hasta Asimov y la ciencia ficción se han quedado cortos). No sabemos si los resultados fueron del agrado de los que lo encargaron, pues junto al Papa y el Dalai Lama, figuraban como "líderes espirituales" el cantante Justin Bieber, el futbolista Neymar o Kim Kardasian (Uno se pregunta ¿a quién seguirán los jóvenes *sin* inquietudes espirituales?).

El segundo aspecto es la capacidad y velocidad de procesado de esos datos. La conocida Ley de Moore establece que aproximadamente cada dos años se duplica el número de transistores en un microprocesador y, con ellos, la capacidad de computación. En el año en que se cumple el cincuenta aniversario de la llegada del hombre a la Luna, es bueno recordar que cualquiera de nuestros teléfonos móviles tiene significativamente más potencia que todos los ordenadores que se emplearon para aquélla gesta. Pero además no solo nuestra capacidad y velocidad para analizar

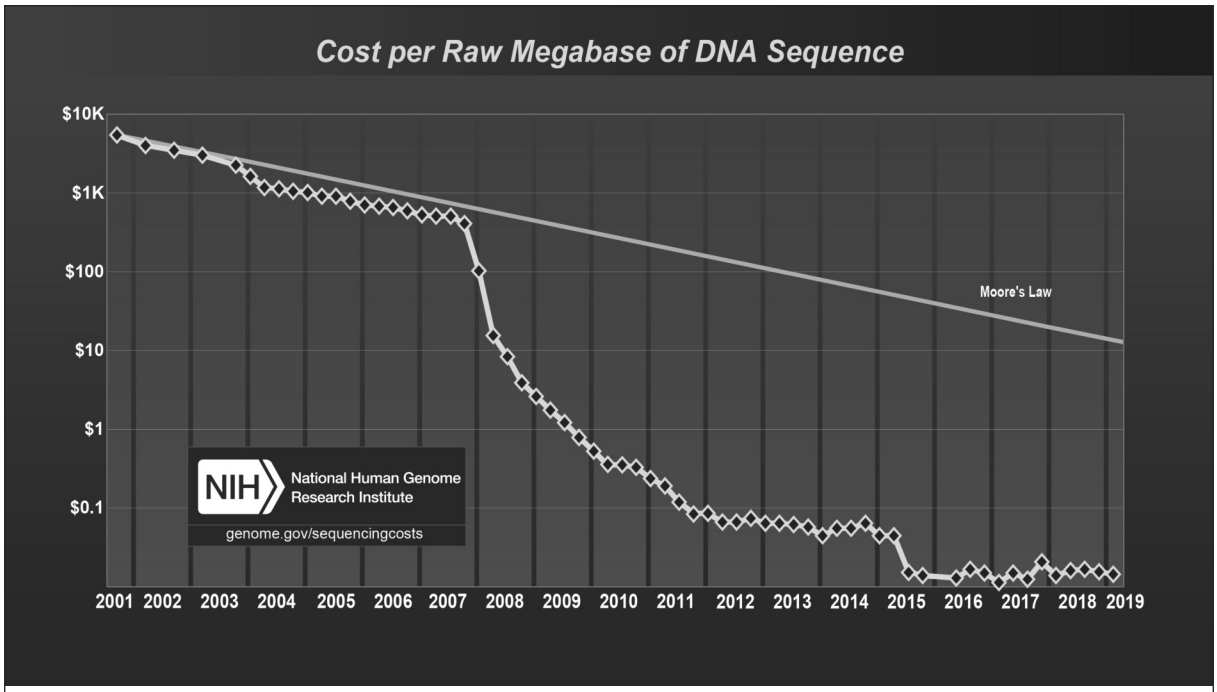

Figura 2. Evolución del coste de la secuenciación de un genoma humano. Tomado de https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

datos ha crecido: también se ha reducido su coste, como cualquier usuario de informática con cierta perspectiva histórica podría atestiguar. Un ejemplo aplicable a la medicina: Tras la finalización del Proyecto Genoma, la secuencia de un genoma humano completo costaba cien millones de dólares. En los años sucesivos dicho coste fue reduciéndose siguiendo, más o menos, la Ley de Moore; pero a partir de 2007 se produjo una reducción aún más abrupta de modo que durante el Black Friday se anunciaban secuenciaciones completas por poco más de 100 dólares[5] (fig. 2).

En definitiva, disponemos hoy de muchos datos y podemos analizarlos de forma muy rápida y relativamente poco costosa ¿Hemos alcanzado, pues, lo que Laplace deseaba? Está claro que no, pues seguramente nunca podremos tener toda la información; pero nos hemos acercado bastante. Sobre esos presupuestos se basa el Big Data.

>>Big Data y medicina

En el año 2015 un grupo de investigadores daneses publicó en el European Heart Journal un interesante artículo[6]. En él se afirmaba que existía un mayor riesgo de padecer estenosis aórtica en los pacientes con psoriasis, según el análisis de una cohorte de población danesa. Lo llamativo era el tamaño de la cohorte: 5.107.604 individuos, esto es, toda la población de Dinamarca. Este tipo de estudios se han hecho populares en los últimos años en los países escandinavos, aprovechando la existencia de una única base de datos de salud nacional (En España, por cierto, mucho más avanzados, no nos limitamos a una única base de datos similar, sino que tenemos diecisiete, correspondientes a las 17 Comunidades Autónomas lo que, por supuesto, no facilita poder realizar estudios similares). Aparte del hecho de que, en lugar de un muestreo —que es lo habitual en la mayor parte de estudios médicos— se empleara a toda la población, y la capacidad de computación que ello supone, llama la atención la relación entre dos entidades tan diferentes. Los autores lo justifican sobre la base común de la inflamación; pero a uno le queda la sospecha de que lo que hicieron fue relacionar muchas variables hasta encontrar una correlación. Estos tres aspectos (toda la población, gran capacidad de análisis en corto tiempo y correlación en vez de causalidad) son los puntos clave en el cambio conceptual que supone la aplicación de Big Data a la Medicina.

El avance que supone la posibilidad de realizar análisis de muchos datos de forma rápida viene

ilustrado por la historia de Matthew Fontaine Maury, que no dispuso de esa capacidad[7]. Maury, un oficial de la Marina de EEUU, debió dejar de navegar a los 33 años, debido a un accidente. Destinado, como jefe del Depósito de Mapas, al Observatorio Naval en Washington, se encontró con una cantidad de datos sin usar que, durante años, los barcos americanos que rendían viaje habían proporcionado sobre vientos, corrientes y clima en diversos puntos del océano y en cada época del año. Decidió entonces recoger de forma sistemática y prospectiva todos estos datos, e incluso consiguió convencer a otros países de que le hicieran llegar los datos reportados por sus flotas. El resultado fue una obra, Physical Geography of the Seas, que proporcionó a los navegante información esencial que les permitió navegaciones más seguras y en menor tiempo que el hasta entonces empleado. Pero para conseguir estos resultados, Maury debió recoger millones de datos durante casi una década; y esos datos fueron trabajosamente analizados por un gran número de "computadores", que era la denominación de los individuos que computaban los datos. Este es un ejemplo de Big Data primitivo, que fue posible gracias a la movilización de ingentes recursos, que rara vez suelen estar disponibles. Posiblemente hoy día todos esos datos se podrían analizar en menos de un día.

>>Google Flu Trends

Google Flu Trends (GFT) fue un proyecto llevado a cabo por Google y que refleja quizá de forma paradigmática las posibilidades y flaquezas de aplicar el BigData a la sanidad. Entre 2003 y 2008, Google recogió datos de 50 millones de las consultas (queries) más frecuentes a su buscador y estableció la combinación de consultas que coincidía de forma más estrecha con la incidencia histórica de gripe en los distintos estados de Estados Unidos. Con objeto de entrever cómo funciona el BigData en estos casos, merece la pena que nos detengamos un momento en la forma en que Google abordó este problema. Google recibe cada día más de 3.000 millones de consultas, y las almacena todas. De estas eligieron las 50 millones más frecuentes, pero no predeterminadas; es decir, no se buscaban preguntas como "Me duele la cabeza y tengo fiebre", sino cualquier tipo de consulta que, a veces y de forma aparente, no guardaban ninguna relación con la gripe. Una vez seleccionadas estas consultas procesaron 450 millones de modelos matemáticos para encontrar el que más se ajustaba a la incidencia real retrospectiva de gripe estacional siguiendo datos históricos. Encontraron entonces una combinación de 45 términos de búsqueda que, al ser empleadas de forma conjunta en un modelo matemático, presentaba una fuerte correlación con los datos reales. Aplicándolo a los brotes de gripe siguientes, fueron capaces de rastrear la incidencia de la enfermedad prácticamente en tiempo real, lo que suponía una ventaja de al menos una semana sobre los informes del Center for Disease Control (CDC) de Atlanta. Y aunque los datos se manipulaban anonimizados y sin intervención humana por motivos de privacidad, por medio del IP pudieron establecer predicciones geográficas precisas. Los datos iniciales, publicados en la revista Nature[8] por seis autores, entre los cuales sólo había dos médicos, despertaron un gran interés, pues mostraban un 97% de coincidencia con los datos del CDC, pero anticipándose a estos en un intervalo variable que llegaba hasta las dos semanas. Este hecho reviste una gran importancia teniendo en cuenta que conocer los datos epidemiológicos en tiempo real, en vez de esperar a la compilación del CDC, podría ayudar a las estrategias de prevención o de utilización de recursos.

Algunos conceptos empiezan a emerger de esta historia. En primer lugar, la tremenda cantidad de datos empleados; en segundo lugar, la impresionante capacidad de procesarlos a alta velocidad; y en tercero, la ausencia de una hipótesis de causalidad. La predicción se basó en coincidencias y correlaciones. Volveremos sobre esto.

El modelo predictivo tuvo que ser ajustado tras subestimar los casos de gripe durante la pandemia no estacional de gripe A (H1N1) en 2009 pero, una vez ajustado, siguió comportándose con gran exactitud[9].

Pero en 2013 GFT sobreestimó de forma grosera en un 140% el impacto real de la enfermedad[10], lo que condujo a que Google diera por finalizado el proyecto (fig. 3). Entre las causas barajadas para este fracaso los diversos autores coinciden en —de nuevo— el factor humano: los patrones de consulta cambian, a medida que lo hace la sociedad y los propios individuos. Por ello se ha postulado que modelos predictivos del

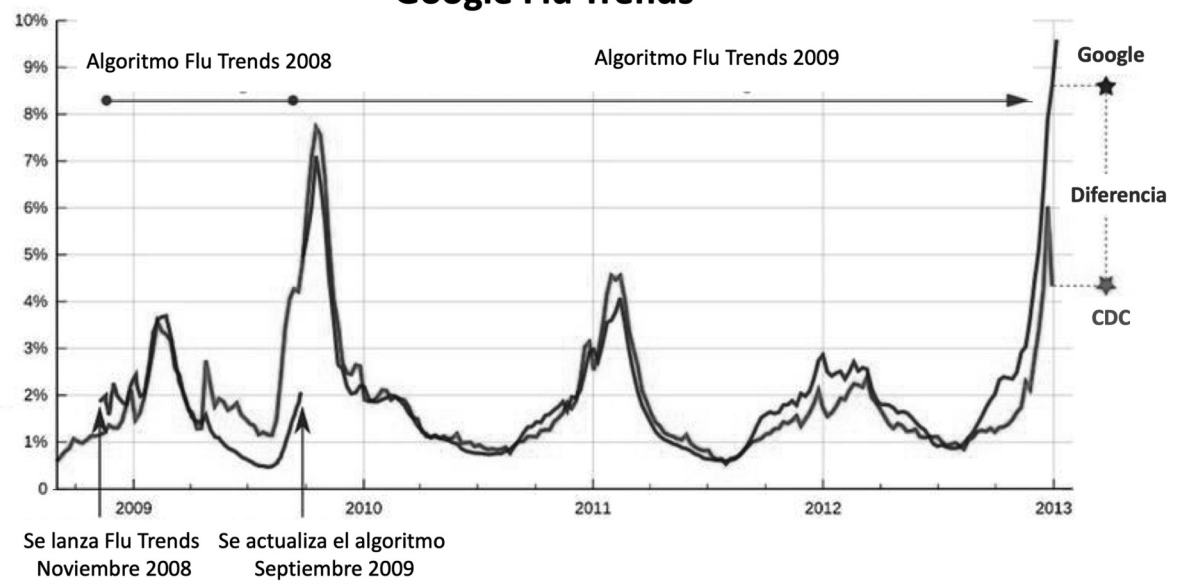

Figura 3. Predicciones de Google Flu Trends. Tras una concordancia inicial muy exacta con los datos del Center for Disease Control (CDC) el brote no estacional de gripe A originó una subestimación grosera de la incidencia. Tras "entrenar" al algoritmo (Septiembre de 2009) las estimaciones volvieron a coincidir hasta que en 2013 se produjo una sobreestimación de los casos que condujo a que Google diera por terminado GFT. Modificado de https://www.wbur.org/commonhealth/2013/01/13/google-flu-trends-cdchttps://www.wbur.org/commonhealth/2013/01/13/google-flu-trends-cdc

tipo de GFT pueden ser válidos durante un periodo corto de tiempo, no superior a los dos-tres años, debiendo ser periódica y permanentemente actualizados[11].

>>¿Qué es Big Data?

No existe una definición clara de lo que entendemos por BigData. En muchos lugares se intenta definir empleando las tres Vs: Volumen, velocidad y variedad de datos, aunque quizá una definición más amplia, aunque poco concreta, sea un cambio masivo en nuestra capacidad de almacenar y analizar datos[12]. En todo caso, y desde el punto de vista de la búsqueda de conocimiento empleando este modelo, lo que sí que supone es un cambio conceptual.

>>Cambios conceptuales: del muestreo al universo

Durante gran parte de la historia de la Medicina el conocimiento ha procedido del análisis de datos; pero la obtención de datos, en Medicina como en cualquier otra actividad, no es fácil, es cara y consume recursos. Por ello se ha recurrido a limitarse a un conjunto reducido de datos que pudiesen suponer una muestra representativa de todo el conjunto. Una serie de métodos estadísticos nos permiten aproximarnos a la verdadera representatividad de esos datos y a la probabilidad de que los hallazgos encontrados sean más o menos fiables. Cualquier investigador familiarizado con el diseño de estudios médicos sabe lo importante que es la fase previa, cuando se calcula el tamaño muestral en función de los resultados esperados, de la variabilidad prevista y de la exactitud que se quiere alcanzar (la famosa p). Pero este abordaje tiene limitaciones, que puede ilustrarse con la historia de los amaños de los combates de sumo en Japón. Durante mucho tiempo se había sospechado que existían combates amañados en este popular deporte japonés, pero los abordajes convencionales para demostrarlo se habían mostrado infructuosos. Se había recurrido al muestreo de una serie de combates al azar para estudiar, empleando la estadística convencional, si había resultados que no encajaban con la calidad e historial relativos de los contendientes; también se habían hecho estudios centrados en los combates finales del campeonato entre los mejores luchadores, con el

mismo resultado. Pero hizo falta aplicar un análisis más completo para descubrir dónde estaba el problema. El campeonato de sumo incluye a 66 luchadores que combaten 15 veces durante la temporada. Al final, los que han obtenido ocho victorias mantienen su rango, mientras que los que no han llegado a esa cifra descienden. Así, si en el último combate se enfrenta un luchador con un record victorias-derrotas de 7-7 con otro cuyo record es 8-6, el primero está necesitado de esa victoria adicional mientras que al otro le da lo mismo. Duggan y Levitt, dos economistas de Chicago, analizaron los 64.000 combates llevados a cabo entre 1989 y 2000 y encontraron que en una situación como la descrita, el luchador con ese record a falta de una victoria ganaba un 25% más de veces que lo normal; es más, la probabilidad de que un luchador terminara exactamente con 8 victorias era el doble (26%) que la de que terminara con 7 (12%), cuando deberían ser iguales (19%)[13]. De esta historia se pueden extraer varias conclusiones: El muestreo aleatorizado puede no descubrir lo que andamos buscando, cuando la anomalía está concentrada en un segmento no representativo. Hace falta entonces recurrir al análisis de *todos* los datos. En segundo lugar, un análisis de este tipo no tiene por qué abarcar una cantidad desmesurada de datos. En este caso el carácter de BigData viene dado no por lo apabullante del número, sino por el hecho de que se incluyen todos los datos. Por el contrario, el incluir una cantidad desmesurada de datos no necesariamente constituye lo que entendemos por Big Data. Algunas webs de noticias sobre nutrición saludaron como Big Data[14] un estudio que analizaba de forma conjunta los resultados de 304 estudios previos sobre alimentación y salud; los autores, más prudentes, lo incluyeron dentro de las revisiones sistemáticas[15]. Una de las diferencias básicas entre un concepto y otro es que los datos de los estudios clínicos tradicionales son, por su propia naturaleza, estructurados, mientras que los que se emplean en Big data no lo son.

Por otra parte, si estudiamos todos los datos, la probabilidad de que los resultados encontrados sean debidos al azar se reducen o desaparecen. Si quisiéramos saber si los españoles son más altos o más bajos que los japoneses, podríamos recurrir a seleccionar una muestra suficiente de ambas poblaciones, medir sus estaturas y saber las diferencias, junto con la probabilidad de que el resultado obtenido fuese debido al azar (la p). En este tipo de estudios la calidad de las muestras adquiere gran importancia, no sólo en cuanto al número (la N) sino también en que sean realmente representativos y a que las mediciones sean lo más precisas posibles. Pero si lo que hacemos es medir a toda la población de España y Japón el planteamiento es diferente, pues la gran cantidad de datos hace que su calidad sea menos importante y no es necesario extrapolar los resultados obtenidos a toda la población, puesto que disponemos de los datos de toda ella. Además, las diferencias encontradas no son una aproximación, sino la realidad. Por otra parte, la gran cantidad de información permite extraer conclusiones sobre subgrupos (por ejemplo, si las españolas son más altas o bajas que las japonesas; si las personas de edad de cada país son diferentes…).

Llevando más allá este aspecto, podríamos incluso establecer relaciones insospechadas. Es sabido que el número de variables que pueden incorporarse en un análisis multivariado depende del tamaño muestral; por ello, cuando se hace un análisis de este tipo se escogen aquéllas variables que razonablemente pueden guardar una relación lógica con lo estudiado; y, por lo general, buscando una relación causa-efecto, basándose en hipótesis apriorísticas. Pero con tamaños enormes se pueden hallar relaciones insospechadas que, aunque pueden ser debidas al azar, sí constituyen la base para generar hipótesis posteriores.

>>Cambios que Big Data va a inducir en la Medicina

Tener una idea de lo que es Big Data no es suficiente, pues no cabe duda que su entrada dentro del campo de la Medicina va a inducir una serie de cambios para los cuales debemos estar preparados. Y, como clínicos, no podemos renunciar a, tras establecer un análisis de la situación —el diagnóstico— aventurar las medidas apropiadas que deberíamos tomar si no queremos que, como ha pasado en otras situaciones, otras las tomen por nosotros.

A primeros de Septiembre de 2018, y durante el Congreso de la European Society of Cardiology, se presentaron los resultados de una macroencuesta realizada a más de 15.000 médicos ameri-

canos. Entre la multitud de datos, destacaba uno: casi la mitad de los encuestados confesaba sentirse "quemado"[16]. Esta cifra, que es similar a la que arrojan estudios entre médicos españoles, revela que la razón de este síndrome de burn-out no es económica, pues los ingresos de los médicos estadounidenses son sustancialmente más elevados que los de los españoles. Podemos entrever las razones del malestar de los médicos a la vista de lo que sucedió tras comunicar los datos. Al finalizar la exposición, y al comienzo del turno de intervenciones, un asistente se puso en pie y pronunció una sola palabra: EPIC, que fue recibida con evidentes muestras de aprobación por parte de muchos asistentes. EPIC es el sistema de historia clínica electrónica y de gestión que emplean un número elevado de hospitales americanos, y que es similar a los sistemas equivalentes que emplean los servicios de salud de las Comunidades Autónomas de España (en este caso no 17 sino más, pues en alguna Comunidad existen dos que, por supuesto, son incompatibles entre ellos). Y la pregunta es ¿por qué?¿Por qué un sistema que, en teoría, debería facilitar las tareas al médico es percibido por éste como motivo de frustración? A cualquier usuario habitual de estos sistemas se le ocurren diversas explicaciones, pues, por lo general, estos sistemas son poco "amigables" para el que los emplea ("user friendly"). A uno le gustaría poder emplear órdenes por voz, que el sistema te reconociera y saludara, que no hubiera que dar tantos clicks, que en vez de repetir innumerables veces una contraseña se pudiese recurrir a la identificación facial o por huella dactilar, que el sistema reconociera la similitud de términos diferentes para referirse a la misma entidad patológica… Pero, claro, eso debe ser pedirle demasiado a un sistema de historia clínica. ¿Lo es? Parece ser que sí, si es que lo que se pretende es incorporar estas mejoras a un sistema ya funcionando; pero no lo es si se tienen en cuenta desde el diseño original. Y es que en este caso lo que los diseñadores originales tuvieron en cuenta fueron las demandas y requerimientos de los que encargaron los sistemas: y esos no fueron los médicos. En Estados Unidos se primó el diseño de una potente herramienta que facilitara la gestión económica; en España quizá lo que se requirió fue otra cosa, pero en todo caso sin tener en cuenta las necesidades del usuario diario, que es el médico. La consecuencia a extraer está clara: Si no queremos que en el futuro las tecnologías que se vayan incorporando a la Medicina nos sean hostiles, habrá que familiarizarse con ellas y conocer sus posibilidades para estar presente cuando se diseñen sus aplicaciones específicas. Si no lo hacemos alguien lo hará por nosotros, con consecuencias negativas para médicos y pacientes.

>>Causalidad frente a correlación

La investigación biomédica se basa en la comprobación de hipótesis y, por tanto, su éxito depende de lo buenas que sean estas hipótesis y lo precisos que sean los métodos de confirmarlas o refutarlas. Por lo general las hipótesis se basan en relaciones de causalidad; pero en ocasiones las verdaderas causas son poco plausibles y, por tanto, no son tenidas en cuenta. Por otra parte, la naturaleza muestral de muchos estudios, sobre un número limitado de observaciones —el mínimo necesario para evaluar la hipótesis en una población determinada— hace imposible realizar análisis de subgrupos, cuyos resultados poco fiables sirven en el mejor de los casos para establecer nuevas hipótesis. El análisis de gran número de datos, o incluso de todos los datos, mediante Big Data es capaz de descubrir relaciones de coincidencia, que no implican causalidad pero que pueden acercarnos a desentrañar mecanismos no sospechados.

La causalidad está tan anclada en la mentalidad biomédica, incluso entre legos, que es difícil que se acepte algún descubrimiento que no ofrezca, al menos aparentemente, una explicación causal. Cuando, en el transcurso de charlas de divulgación, se menciona el hecho de la baja mortalidad cardiovascular en los países mediterráneos a pesar de que, en el caso concreto de Francia —al menos en el centro y norte del país— la dieta no es precisamente mediterránea, la primera pregunta del oyente no informado es inevitablemente "Y eso ¿por qué?".

Un caso paradigmático es la historia de Semmelweiss. Ignaz Philippe Semmelweiss (1818-1865) fue un médico húngaro que trabajaba como obstetra en Viena. Semmelweiss observó que la mortalidad por fiebre puerperal era mucho más alta en una de las dos salas que atendían parturientas que en la otra, e incluso que entre las parturien-

tas que daban a luz fuera del hospital. La primera sala era atendida por médicos y estudiantes de Medicina, mientras que la segunda lo era por matronas y estudiantes de matrona. Tras la muerte de un amigo suyo al realizarse un corte con el mismo bisturí con el que realizaba una autopsia, Semmelweiss pensó que la causa era una "contaminación cadavérica", transmitida a las parturientas por los estudiantes de Medicina, que realizaban autopsias y disecciones de cadáveres, pero no por las matronas que no tenían relación con cadáveres. Aunque los médicos se lavaban las manos con agua y jabón —si bien no siempre— antes de atender a las pacientes, esa práctica no eliminaba el olor cadavérico. Semmelweiss hizo pruebas con distintas soluciones hasta que halló que lavándose las manos con una solución de hipoclorito sódico se eliminaba por completo el olor a cadáver. Instituyó esa práctica en la primera sala de obstetricia, consiguiendo reducir la incidencia de fiebre puerperal casi a cero[17]. Pero a pesar del éxito evidente, su técnica no fue aceptada ni copiada por otras instituciones, porque, en una época en que se ignoraba aún el papel de los gérmenes en la transmisión de las enfermedades infecciosas, no ofrecía una relación de causalidad plausible para los conocimientos de entonces. Es más, fue expulsado del hospital de Viena y retornó a su Hungría natal donde, en el hospital local, repitió con éxito su técnica de lavado reduciendo la mortalidad por fiebre puerperal por debajo del 1%, de nuevo ante el escepticismo y rechazo de sus colegas[18]. No fue hasta después de su muerte, cuando los trabajos de Pasteur y Lister establecieron la teoría de la transmisión de la infección por gérmenes y las prácticas de asepsia y antisepsia en cirugía, que los trabajos de Semmelweiss adquirieron un justo, aunque tardío, reconocimiento.

El tipo de investigación basado en razonamientos deductivos, sobre una idea previa de relación causal, es sustituido en Big Data por una investigación inductiva, basada en hallazgos de coincidencia que, posteriormente, pueden ser investigados por métodos más convencionales. Un ejemplo de éxito comparativo de esta segunda aproximación sobre la primera lo ofrece la búsqueda de correlaciones genéticas de distintas enfermedades. Hasta los primeros años de este siglo, la investigación sobre la base genética de diversas patologías se fundaba en la búsqueda de mutaciones o polimorfismos en uno o varios genes candidatos, basándose en probables relaciones de causalidad según el conocimiento previo. Aunque se publicaron numerosos estudios al respecto, apenas alguno de los hallazgos fue consistente o replicable. No fue hasta la popularización de las técnicas de análisis genómico amplio (genome wide analisys studies, GWAS), no necesitadas de hipótesis inicial, que se empezaron a desentrañar numerosas asociaciones genotipo-fenotipo con implicaciones fisiopatológicas, epidemiológicas e incluso terapéuticas[19].

Pero sería ingenuo pensar, al igual que los inductivistas ingenuos primitivos, que los análisis basados en Big Data están siempre desprovistos de hipótesis. La cantidad de datos que pueden relacionarse es tal, y el nivel de "ruido" que los contamina, que crece con el volumen de los datos, tan alto, que hace que en muchas ocasiones los campos de análisis se centren y circunscriban a áreas determinadas por hipótesis previas.

>>Herramientas de análisis

La obtención de conocimiento basada en hipótesis causales y muestras representativas han llevado al desarrollo de complejas herramientas estadísticas que nos permiten precisar el grado de aproximación a la realidad, o de error, que podemos asumir. Pero esas herramientas son poco adecuadas para manejar conjuntos inmensos y completos de datos. Por ejemplo, la estadística convencional puede determinar que alguna variable en concreto es significativa cuando no se debe más que a lo que en el campo del Big Data se conoce como ruido. Por todo ello no es sorprendente que los análisis de este tipo de datos no sean el campo de los estadísticos sino el de matemáticos y otros expertos. Big Data emplea herramientas diferentes en el ámbito de la inteligencia artificial, como el *machine learning* (aprendizaje por las máquinas) o las redes neuronales. Un ejemplo de lo primero nos lo ofrecen los distintos métodos de traducción automática. Los que se acercaron, por ejemplo, a Google Translator, hace unos años pudieron quedar decepcionados por su funcionamiento. Pero al compararlo con el funcionamiento actual se tiene la misma impresión que cuando, al cabo del tiempo, encontramos al niño de pocos años que comenzaba a balbucear convertido ahora en un correcto dominador del lenguaje. Y, al igual que la forma natural

de aprender a hablar, esa evolución no se debe al estudio previo de vocabularios o reglas gramaticales, sino al aprendizaje por reiteración y repetición en distintos contextos. Google Translator se basa precisamente en eso; al alimentarlo con una tremenda cantidad de textos y su traducción a otros idiomas va mejorando progresivamente: en eso consiste el machine learning.

El problema es que no existen en la actualidad suficientes expertos para los usos crecientes del Big Data en la sociedad (*data scientist y chief data officer*). La Comisión Europea prevé que hasta 2020 (a la vuelta de la esquina) se precisarán 112.000 tecnólogos, ingenieros y matemáticos, al año mientras que el déficit actual de expertos en Big Data se cifra en 800.000 en la Unión Europea, de ellos un 10% en España[20]. Teniendo en cuanta las aplicaciones de Big Data en el campo industrial y mercantil y, por tanto, la demanda de los escasos especialistas por este tipo de actividades, sin duda mejor remuneradas que la investigación, se pone de manifiesto el problema con que la Medicina se enfrenta al pretender realizar análisis de Big Data.

¿Qué importancia tiene este hecho desde el punto de vista de la Medicina? Está claro que no podemos esperar que los médicos se conviertan en expertos en Big Data, al igual que no pueden ser consumados estadísticos. Pero, lo mismo que en los grupos de investigación de cierto tamaño se dispone de bioestadísticos, en un futuro inmediato será preciso contar con otro tipo de expertos versados en el manejo y análisis de Big Data. Y, lo que es más importante los médicos deberán entender los principios de este nuevo tipo de análisis de una forma semejante a como hoy tienen conocimientos básicos de estadística; podrán así orientar el diseño de los estudios de investigación y, lo que es clave, entender la validez metodológica de los resultados de esos estudios, propios o ajenos. Todo ello aconseja que se comiencen a introducir este tipo de enseñanzas en la formación curricular de los futuros médicos, que sin duda se van a enfrentar a una práctica de la Medicina muy diferente de la actual.

>>Utilidad pública frente a privacidad

El empleo de Big Data plantea otra serie de problemas, alguno de ellos de naturaleza ética. Los estudios basados en muestreo requieren un número limitado de muestras de alta calidad y, en el caso de los estudios médicos, el consentimiento informado de los pacientes que participen. Pero los estudios de Big Data, que se basan en todo el universo de datos o, en el caso que nos ocupa, todos los individuos de una determinada población, se enfrentan a serias dificultades. Por una parte, es prácticamente imposible conseguir consentimiento informado general, pero por otra no puede emplearse de forma arbitraria información sensible sobre la población sin contar con su conocimiento. Como quiera que el resultado de dichos estudios puede repercutir de forma positiva sobre la salud de la sociedad, nos encontramos con el dilema de si debe primar el respeto de la privacidad individual o el bien común.

Un ejemplo de los problemas que este tema plantea lo representa el proyecto sueco LifeGen[21], que pretendía enrolar a una cohorte de 500.000 personas de entre 18 y 45 años y seguirlos de forma prospectiva durante al menos 20 años con objeto de desentrañar, entre otros objetivos, las interacciones entre genes y entorno. En 2011 las autoridades detuvieron el estudio debido a que consideraban ilegal, por excesivamente amplio, el consentimiento que los participantes otorgaban. Pero es claro que si se quieren obtener datos relevantes para la salud de la población de forma prospectiva el consentimiento no puede ser excesivamente específico, si lo que se pretende es poder abordar aspectos que en el momento actual no conocemos. Tras una serie de peripecias legales, se revisó la legislación al respecto y se pudo reanudar el estudio tres años más tarde.

El debate, que está abierto, será motivo de discusión en los próximos años a medida que se haga evidente el beneficio potencial de estudios de este tipo. Se han propuesto a este respecto diversas soluciones[22,23,24]. La Unión Europea acaba de instaurar un Reglamento General de Protección de Datos[25], y existen iniciativas dentro de la misma Unión Europea y en el marco del Horizonte 2020 para abordar en profundidad estos temas (Big Data for Better Outcomes, Policy Innovation and Healthcare System Transformation)[26]. En Diciembre de 2018 se publicó en España la Ley Orgánica 3/2018, de Protección de Datos Personales y garantía de los derechos digitales[27].

Pero el tema no termina aquí pues surgen nuevos interlocutores con argumentos de índole distinta. En una serie países están comenzando a surgir asociaciones de pacientes o ciudadanos, conscientes de que los datos —sus datos— son valiosos y que quieren tener algo que decir respecto a su utilización. La iniciativa Salus Coop[28] (Cooperativa Ciudadana de Datos de Salud), por ejemplo, llama a los ciudadanos a "compartir y gobernar conjuntamente sus datos de salud" al tiempo que propone que éstos tengan un papel activo en la investigación biomédica, quizá contribuyendo a sugerir campos de estudio. Es más, ya existen diversas empresas de secuenciación de genoma que ofrecen sus resultados mediante blockchain, de forma que el propietario pueda cederlos para investigación a cambio de un valor[29]. En definitiva, Big Data plantea escenarios distintos a los que la comunidad médica estaba acostumbrada, y el debate sobre muchos de sus aspectos deberá escapar de los restringidos límites del mundo sanitario.

>>Nuevas opciones curriculares

El currículo de los médicos descansa en parte en su capacidad investigadora, reflejada indirectamente en sus publicaciones, que revisten gran importancia en aspectos como la acreditación en el campo académico. El número de publicaciones, la categoría de las revistas, su factor de impacto, la posición relativa de los distintos firmantes, son todos ellos determinantes del valor relativo de los currículos de los médicos. Los autores son los que han contribuido a diseñar el trabajo, recoger los datos, analizarlos y redactar el documento final. Pero en aquéllos estudios que descansen en análisis de Big Data hay otro aspecto que hay que tomar en consideración: quién ha contribuido aportando los datos que maneja. Para que el análisis incluya todos los datos que se quieren analizar, hay que incentivar a aquéllos que los detentan, puesto que sin su participación el trabajo no es factible; y en la actualidad ese papel está escasamente reconocido. Pese a que hace unos años algunas de las principales revistas médicas adoptaron la política de reconocer como autores a todos los investigadores locales —a veces cientos— de los ensayos clínicos multicéntricos, este tipo de autoría sigue siendo escasamente reconocido. Seguramente en el futuro habrá que reconsiderar este aspecto, reconociendo el valor de los médicos que aportan datos que otros incluirán en análisis de Big Data.

>>Big Data y Nutrición

El conocimiento sobre las consecuencias de consumir determinados alimentos sobre nuestra salud es limitado, a pesar de la cantidad de información generada a lo largo de los años. Las razones estriban en la dificultad de realizar estudios adecuadamente diseñados. Los estudios doble ciego son imposibles de realizar, por razones obvias, e incluso los estudios prospectivos de intervención son escasos. Gran parte de la información actual se basa consideraciones apriorísticas o en estudios observacionales, cuando no retrospectivos, extrayendo la información de los datos autoreportados por los participantes. Por otra parte, datos como relación cantidad-efecto, intervalos en la alimentación, forma de preparación, interacción entre alimentos diferentes, interrelación con la actividad física, influencia del sustrato genético etc. hacen que la cantidad de variables a considerar sea enorme. Es en campos como este donde el Big Data ofrece un gran potencial, si se puede reunir una cantidad de observaciones suficientes para someterlas a un análisis desprovisto de hipótesis previas. El proyecto RICHFIELDS[30], sustentado por la Unión Europea a través de su programa 2020, es un primer intento de aproximación a este tipo de análisis. Según sus propias afirmaciones, RICHFIELDS pretende "diseñar una plataforma de datos de los consumidores para recopilar y conectar, comparar y compartir información sobre nuestros comportamientos alimentarios, para revolucionar la investigación sobre las opciones que en este campo se toman a diario en toda Europa." Posiblemente en el futuro este tipo de estudios se completen con la información obtenida de los puntos de venta de alimentos en cada zona, así como otra proveniente de la capacidad de rastrear alimentos determinados mediante blockchain, e incluso de la que puedan aportar dispositivos como frigoríficos "inteligentes" que puedan detectar tipo de alimento, entrada y salida.

>>Conclusiones

La capacidad creciente de acumulación y análisis de datos son la base sobre la que se asienta lo que

se conoce como Big Data. Aplicado a la Medicina constituye una nueva herramienta de obtención de conocimiento que, básicamente, supone substituir la técnica del muestreo por el análisis de todos los datos de un conjunto, o al menos de una abrumadora mayoría, esperando encontrar relaciones de coincidencia en vez de relaciones de causalidad. Esto va a suponer una serie de cambios a distintos niveles. En primer lugar, a nivel conceptual, pues el principio de causalidad está sólidamente arraigado en la mentalidad biomédica; pero además va a requerir que nos familiaricemos con los métodos de análisis de Big Data, para contribuir al diseño de estudios y entender de forma crítica sus resultados. Por otra parte, actualizará el debate entre privacidad de datos y bien común, e incluso puede que aconseje rediseñar los actuales criterios curriculares. Pero, en definitiva, nos debe obligar a interesarnos en un campo con un gran futuro, para que en su desarrollo posterior en el campo de la Medicina estemos presentes como agentes, y no como receptores pasivos, de lo que tecnologías de este tipo nos ofrecen.

Conflictos de intereses

El autor declara no tener conflicto de intereses.

Bibliografía

1. Hilbert M, Lopez P. The world's technological capacity to store, communicate, and compute information. *Science* 2011; 332: 60-5.
2. (https://public.dhe.ibm.com/common/ssi/ecm/wr/en/wrl12345usen/watson-customer-engagement-watson-marketing-wr-other-papers-and-reports-wrl 12345usen-20170719.pdf).
3. «Christopher Wylie, el vegano canadiense que ideó la herramienta de guerra psicológica para la campaña de Trump». eldiario.es. Consultado el 31 de Julio de 2019.
4. https://www.lavanguardia.com/vida/20180929/452069931810/influencers-instagram-religion-relacion-estudio-impacto.html Consultado el 31 de Julio de 2019.
5. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data
6. Khalid U, Ahlehoff O, Gislason GH, Skov L, Torp-Pedersen C, Hansen PR. Increased risk of aortic valve stenosis in patients with psoriasis: a nationwide cohort study. *Eur Heart J* 2015; 36 (32): 2177-83. doi: 10.1093/eurheartj/ehv185.
7. Guarnieri M. Matthew Fontaine Maury: The 19th-Century Forerunner of Big Data [Historical] Article in IEEE Industrial Electronics Magazine • June 2018 DOI: 10.1109/MIE.2018.2827861
8. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature* (2009) 457: 1012-1014.
9. Cook S, Conrad C, Fowlkes AL, Mohebbi MH. Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. *PLOS* (2011) 6: e23610.
10. Olson DR, Konty KJ, Paladini M, Viboud C, Simonsen L: Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales. *PLoS Comput Biol* 2013; 9 (10): e1003256. https:// doi.org/10.1371/journal.pone.0023610)
11. https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/ Consultado el 2 de Agosto de 2019.
12. Mayer-Schönberger V, Ingelsson E. Big Data and medicine: a big deal? *J Intern Med* (2017); doi: 10.1111/ joim.12721
13. Duggan M, Levitt SD. Winning isn't everything: corruption inSumo Wrestling. *Am Econ Rev* 2002; 92: 1594-605.
14. https://www.foodnavigator.com/Article/2014/12/11/Big-data-Exhaustive-review-pulls-together-evidence-on-food-groups-and-diet-related-disease
15. Fardet A, Boirie Y. Associations between food and beverage groups and major diet-related chronic diseases: an exhaustive review of pooled/ meta-analyses and systematic reviews. *Nutrition Reviews* Vol. 72 (12): 741-762. doi:10.1111/ nure.12153
16. Medscape National Physician Burnout and Depression Report 2018. www.medscape. com/slideshow/ 2018-lifestyle-burnout- depression-6009235, consultado 7 de agosto 2019.
17. Semmelweis Ignaz (1983). The Etiology, Concept, and Prophylaxis of Chilbed Fever. The University of Wisconsin Press. ISBN 0299093646.
18. Noakes TD, Borresen J, Hew-Butler T, Lambert MI, Jordaan E. (2008). «Semmelweis and the aetiology of puerperal sepsis 160 years on: an historical review». Epidemiol. Infect (2008), 136, 1-9. doi:10.1017/S095 0268807008746

19. https://www.genome.gov/es/genetics-glossary/ Estudio-de-asociacion-de-genoma-completo

20. https://elpais.com/tecnologia/2018/09/25/actualidad/1537892876_048111.html

21. Almqvist C, Adami HO, Franks PW et al. LifeGene–a large prospective population-based study of global relevance. *Eur J Epidemiol* 2011; 26: 67-77.

22. Jacobs B, Popma J. Medical research, Big Data and the need for privacy by design. *Big Data and Society* (2019) https://doi.org/10.1177/2053951718824352

23. Ienca M, Ferretti A, Hurst S, Puhan M, Lovis C, Vayena E (2018) Considerations for ethics review of big data health research: A scoping review. *PLoS ONE* 13 (10): e0204937. https://doi. org/10.1371/journal. pone.0204937

24. Abouelmehdi et al. Big healthcare data: preserving security and privacy. *J Big Data* (2018) 5: 1 https://doi.org/10.1186/s40537-017-0110-7

25. https://www.actasanitaria.com/wp-content/uploads/2019/02/L00001-00088.pdf

26. https://cordis.europa.eu/project/rcn/209468/factsheet/en

27. https://www.boe.es/diario_boe/txt.php?id=BOE-A-2018-16673

28. https://www.saluscoop.org/

29. Ozercan HI, Ileri AM, Ayday E, Alkan C. Realizing the potential of blockchain technologies in genomics. *Genome Res* 2018; 28 (9): 1255-1263. doi: 10.1101/gr. 207464.116

30. https://www.richfields.eu/