Utilización de tecnología *Big Data* en investigación clínica*

Use of Big Data technology in clinical research

Guillermo Alcalde Bezhold
Comité Ético de Investigación Clínica. Organización Sanitaria Integrada Araba (Vitoria, España)

Iciar Alfonso Farnós
Comité de Ética de la Investigación con medicamentos de Euskadi, Dirección de Farmacia, Gobierno Vasco (Vitoria, España)

DOI: 10.14679/1133

Resumen / Abstract: La utilización de la tecnología *Big data* en la práctica clínica puede suponer grandes avances para el sistema sanitario, tanto para la asistencia

* Artículo recibido el 5 de junio de 2019 y aceptado para su publicación el 16 de julio de 2019.
Para la realización de este trabajo se ha contado con el apoyo del Proyecto de Investigación BIGDATIUS (Uso de datos clínicos ante nuevos escenarios tecnológicos y científicos –*Big Data*–. Oportunidades e implicaciones jurídicas. UPV/EHU. Ref.: DER2015-68212-R), financiado por el Ministerio de Economía y competitividad y Fondo Europeo de Desarrollo Regional (FEDER).

como para la investigación clínica. Para conseguir este objetivo es imprescindible la integración de las diferentes fuentes de información actualmente disponibles. De forma simultánea a estas ventajas potenciales, la aplicación de estas tecnologías puede suponer una fuerte amenaza para la intimidad, por lo que es imprescindible disponer de medidas adecuadas de control de la información, así como implantar procedimientos adecuados, transparentes y seguros, que garanticen el máximo nivel de confidencialidad asegurando el respeto a los derechos y libertades de las personas. Para poder utilizar de forma segura esta tecnología es imprescindible disponer de un marco ético y jurídico adecuado, que permita compaginar la protección de datos con la investigación clínica relevante y la mejora de la asistencia.

The use of Big data technology in clinical practice can mean great advances for the healthcare system, both for care and for clinical research. To achieve this goal, it is essential to integrate the different sources of information currently available. Simultaneously with these potential advantages, the application of these technologies can pose a strong threat to privacy, so it is essential to have adequate control measures for information, as well as to implement adequate, transparent and secure procedures, which guarantee the highest level of confidentiality ensuring respect for the rights and freedoms of individuals. In order to be able to use this technology safely, it is essential to have an adequate ethical and legal framework that allows data protection to be combined with relevant clinical research and improved care.

Palabras clave / Keywords:

Investigación clínica / Grandes cantidades de datos / Inteligencia artificial / Asistencia sanitaria.

Clinical research / Big Data / Artificial intelligence / Health care.

1. Introducción

El *Big Data* sanitario es una tecnología que se ha incorporado con fuerza al sector sanitario porque supone una gran fuente de oportunidades para la asistencia y la investigación. Sin embargo, como cualquier otro avance tecnológico, su implantación no está exenta de amenazas. El objetivo de esta revisión es analizar las controversias que han surgido de la aplicación de esta tecnología al sector sanitario y cómo se pueden compaginar las ventajas y los posibles inconvenientes.

Existen múltiples definiciones de lo que representa la tecnología *Big Data*. Una definición habitualmente aceptada es "el acceso a grandes cantidades de información disponible con el fin de dar valor añadido a su análisis integrado, con el objetivo de extraer información oculta o correlaciones imprevistas, no deducibles ni inferibles con los métodos de análisis convencionales" (1).

Otras definiciones del *Big Data* se centran más en la tecnología empleada que en el uso que se da al análisis de los datos. Entre estas tecnologías se

incluyen las diferentes variantes de la inteligencia artificial, que se clasifican según la necesidad de intervención humana, progresivamente menor según avanza la tecnología computacional. Todas estas tecnologías parten del análisis de datos a gran escala a partir del cual se diseñan modelos informáticos.

El *Big Data* sanitario también se caracteriza por reunir las siguientes características claves (3 uves): volumen de los datos; velocidad de procesamiento y variedad de las fuentes. Posteriormente, se han añadido otras características relevantes que deben reunir los datos, como que sean veraces o aporten valor (5 uves) (1,2). Por tanto, no existe una única definición de *Big Data*, sino definiciones que abordan diferentes aspectos complementarios.

La aplicación de la tecnología *Big Data* para el sector sanitario puede ser una fuente generadora de recursos económicos, ligados al conocimiento que generan los datos y las tecnologías para su procesamiento y análisis, de tal forma que se ha atribuido a los datos el papel del nuevo petróleo de la economía digital. Como ejemplo de la importancia de esta tecnología, la Unión Europea ha elaborado el documento "Redesigning health in Europe for 2020", en el que se plantean los grandes retos de la Unión Europea en la aplicación del *Big Data* sanitario (3):

- Incrementar el conocimiento de los ciudadanos sobre las oportunidades de la salud digital (eHealth).

- Generar recursos ligados a la salud digital, entendido como la aplicación de tecnologías digitales de la información y comunicación para mejorar la salud.

- Mejorar la salud de la población, reducir las desigualdades sociales o aumentar la eficiencia del sistema sanitario.

- Fomentar la integración de datos en grandes sistemas europeos y mejorar las posibilidades de acceso para los investigadores, así como conseguir una mayor integración entre la investigación y la práctica clínica.

Por otra parte, el avance progresivo de la tecnología de procesamiento de datos ha supuesto que cada vez sea más factible reidentificar los datos anonimizados, de tal manera que el principio de la utilización de datos anónimos como garantía absoluta de protección de la privacidad ha dejado de ser válido (4). Como ejemplo, se pueden citar estudios previos que han demostrado

que combinando datos básicos del censo de EEUU, tales como la fecha de nacimiento, el sexo y el código postal de residencia, el 87% de la población tiene características únicas, y por tanto identificables (5). De este modo, datos que inicialmente parecen anónimos se pueden reidentificar posteriormente de forma directa mediante la combinación de valores de diferentes variables individuales o mediante el cruce adicional con otras bases de datos relacionadas (4).

2. *Big Data* e inteligencia artificial

La **inteligencia artificial** es una rama de la informática dedicada a crear sistemas que realizan tareas que normalmente requieren la intervención de la inteligencia humana.

Los primeros fundamentos de inteligencia artificial consistían únicamente en la programación por un informático mediante unas reglas preestablecidas según el criterio de expertos sin que interviniese el sistema informático (6). Estas técnicas basadas en la programación eran efectivas en entornos con reglas formales y abstractas (como los programas capaces de jugar al ajedrez), pero se enfrentaban a limitaciones importantes en situaciones intuitivas y sencillas para cualquier persona como reconocer imágenes o entender una conversación. Por este motivo, pronto se llegó a la conclusión de que los sistemas de inteligencia artificial necesitaban tener la capacidad de adquirir su propio lenguaje extrayendo patrones a partir de los datos sin procesar, capacidad denominada aprendizaje de máquina *(*machine learning*)*.

Los primeros sistemas de "machine learning" confiaban en expertos humanos para clasificar las características relevantes a partir de las cuales se entrenaba el sistema, bien fuesen datos o imágenes, como ha ocurrido con los programas capaces de diagnosticar la retinopatía diabética a partir de imágenes clasificadas previamente por oftalmólogos.

Posteriormente, se han incorporado los sistemas denominados de aprendizaje por representación *(representation learning)*, que consisten en algoritmos que aprenden a extraer las características relevantes a partir de los datos sin que se hayan definido previamente por un experto humano.

Por último, los modelos más avanzados de inteligencia artificial se han basado en las técnicas de aprendizaje profundo *(deep learning)* mediante redes neuronales, en las que las características que aprende la máquina son compuestas o jerárquicas en varios niveles, de lo simple a lo más complejo,

sin que el modelo tenga ningún tipo de supervisión humana. Estos sistemas tienen el inconveniente de que no se puede explicar cómo llega el sistema a sus conclusiones, funcionando como una "caja negra" cuya lógica interna es difícil de entender e interpretar para los profesionales expertos en la materia, aunque la predicción sea igual de precisa (7,8,9,10).

Los modelos "deep learning" han tenido gran éxito en diferentes campos, como en la clasificación de imágenes médicas, en el reconocimiento del lenguaje o en el procesamiento del lenguaje natural. Los modelos más complejos de aprendizaje por redes neuronales tienen su campo de acción principal en las especialidades que utilizan imágenes de forma masiva.

De este modo, estas técnicas se han incorporado a la práctica clínica diaria mediante ayudas al diagnóstico, simplificando tareas como la interpretación automática de los electrocardiogramas, el recuento celular de los hemogramas, la lectura automática de citologías vaginales, la interpretación de retinografías para la detección de retinopatía diabética, la detección de pólipos de colon o microsangrados cerebrales en imágenes radiológicas o la interpretación de fotografías para el diagnóstico de cáncer cutáneo, en ocasiones con mejor exactitud que médicos especialistas (7,11).

Por otra parte, también se han desarrollado modelos predictivos mediante sistemas de aprendizaje profundo a partir de los datos de la historia clínica electrónica, que han demostrado su utilidad para predecir la mortalidad durante el ingreso, el riesgo de reingreso no programado a los 30 días, la prolongación de la estancia y los diagnósticos al alta (12). En un futuro, la integración de estas técnicas con herramientas de apoyo a la decisión clínica, tales como alertas automáticas o ayudas al diagnóstico, podrá ofrecer a los médicos información relevante en tiempo real para mejorar las decisiones clínicas (13).

3. Generación de datos en la práctica clínica

La práctica clínica genera diariamente una enorme cantidad de información relevante para conocer la incidencia, prevalencia y evolución de las enfermedades. El desarrollo de la historia clínica electrónica en cada sistema de salud, como repositorio de toda la información asistencial con un identificador único por paciente que enlaza las diferentes bases de datos, ha facilitado el acceso a esta enorme cantidad de información, que según su posibilidad de procesamiento ulterior se puede considerar como estructurada o no estructurada (14).

La historia clínica contiene información estructurada en forma de datos codificados o numéricos que se pueden procesar directamente, como los resultados de laboratorio o las constantes clínicas de los pacientes (temperatura, frecuencia cardiaca o presión arterial). En otras ocasiones, los datos requieren una labor previa de codificación, como ocurre con la codificación de los diagnósticos y procedimientos de cada episodio asistencial mediante sistemas de codificación estándar, de los cuales el utilizado habitualmente es la Clasificación Internacional de Enfermedades (CIE), promovida por la Organización Mundial de la Salud. La codificación mediante la CIE permite convertir los diagnósticos, procedimientos y otros problemas de salud en códigos alfanuméricos homogéneos internacionalmente, lo que facilita su almacenamiento y posterior recuperación para el análisis de la información (15). A partir de la codificación según la CIE se han desarrollado herramientas de análisis de la información asistencial, tales como los GRD (Grupos relacionados con el diagnóstico), que agrupan los episodios de hospitalización por diagnósticos al alta relacionados con un similar consumo de recursos. Estas herramientas son capaces de aportar a los Sistemas de Salud información comparativa de la eficiencia de la actividad clínica (16). Sin embargo, la codificación clínica es un proceso complejo que requiere personal entrenado.

Otros ejemplos de información estructurada son los datos administrativos de los contactos con el sistema sanitario (atenciones en consultas externas, urgencias, hospital de día o ingresos hospitalarios) o los datos obtenidos de los sistemas de prescripción electrónica. Estos datos, cruzados con otra información estructurada como las variables sociodemográficas (edad, sexo o lugar de residencia) y con la codificación de los diagnósticos y/o procedimientos pueden proporcionar información muy valiosa sobre el pronóstico y el manejo de diferentes enfermedades.

En otras ocasiones se ponen en marcha registros específicos para tener un conocimiento exhaustivo de algunas enfermedades o condiciones con especial relevancia para la salud pública. Un registro es un sistema organizado para la recopilación, almacenamiento, extracción, análisis y diseminación de la información sanitaria. En el entorno sanitario se pueden mencionar los ejemplos de los registros de mortalidad, los registros de cáncer, las enfermedades de declaración obligatoria o más recientemente, los registros de enfermedades raras (17).

Pero la práctica clínica genera una cantidad aún mayor de información no estructurada. Los evolutivos o los informes en texto libre de las historias clínicas electrónicas son una fuente no estructurada de información, cuya

finalidad principal es garantizar que los profesionales sanitarios tengan la información clínica disponible para la asistencia y la puedan interpretar según su criterio. La información en texto libre no es directamente procesable, por lo que para que pueda convertirse en información de utilidad para la toma de decisiones y la investigación clínica son necesarios sistemas automáticos capaces de extraer y procesar esta información. Hoy en día se están desarrollando estas herramientas informáticas con el objetivo de tener conocimiento sobre los resultados y la eficiencia de los procesos asistenciales. Eventualmente, estas herramientas podrían ser capaces de detectar patrones de comportamiento clínico y sugerir cursos de acción mientras se está llevando a cabo la asistencia, comparando las características del paciente que se está atendiendo con patrones previos. Sin embargo, los clínicos deberán tener en cuenta que los resultados que se obtienen no dejan de ser predicciones estadísticas y mantener un alto índice de sospecha sobre su exactitud (11).

Existen, por otro lado, otras fuentes de información no estructurada como las imágenes médicas, tanto de pruebas complementarias (radiología, ecografía, medicina nuclear, resonancias, electrocardiografía, retinografías, etc), como fotografías que se adjuntan a la historia clínica. También se incluyen en este apartado las grabaciones en video de procedimientos (2) o incluso la información contenida en las publicaciones científicas, que ha aumentado de forma exponencial en los últimos años (18).

Por último, cada vez van a ser más importantes las fuentes de información alternativas, provenientes de aplicaciones móviles médicas o relacionadas con la salud, generadas por diferentes dispositivos portátiles, además de los mensajes de las redes sociales generales o específicas de salud o los metadatos (datos relacionados con la generación de otros datos, como es el caso de la geolocalización de los datos que genera un dispositivo).

En resumen, la actividad asistencial genera una enorme cantidad de información, de la que gran parte no está estructurada, por lo que sin un desarrollo tecnológico apropiado difícilmente se puede aprovechar para su tratamiento a gran escala.

4. Oportunidades del *Big Data* sanitario para las organizaciones sanitarias

El entorno sanitario es de gran complejidad. La mejora en las condiciones de vida de la población, los progresos en el tratamiento de las enfermedades

crónicas junto con el envejecimiento progresivo de la población han cambiado los patrones de las enfermedades. En este entorno cambiante, la irrupción de los sistemas de información sanitaria informatizados, incluyendo la aparición de la telemedicina o la historia clínica única entre los diferentes niveles asistenciales y sociosanitarios, ha permitido nuevos desarrollos y formas de atención impensables previamente.

La aplicación de los sistemas de inteligencia artificial a la organización del sistema sanitario puede aportar grandes ventajas potenciales. Se mencionan a continuación algunos de los posibles beneficios más destacados (14,19):

- **Sistemas de salud que aporten calidad en la asistencia**: las técnicas de *Big Data* permiten obtener datos a partir de la práctica clínica diaria para desarrollar modelos predictivos que se puedan incorporar como ayudas a la decisión tanto para profesionales como para gestores;

- Sistemas sanitarios que **fomenten la investigación**: la aplicación de tecnología *Big Data* permite construir modelos expertos capaces de autoaprendizaje que puedan generar conocimiento para la mejora de la salud, así como la posibilidad de realizar ensayos clínicos pragmáticos en condiciones de práctica clínica real;

- Sistemas sanitarios **seguros**: que incorporen herramientas de análisis de seguridad clínica, incluyendo prevención de errores y detección de incidentes, así como integrar los sistemas de farmacovigilancia (notificación automática de reacciones adversas de fármacos);

- Sistemas sanitarios **eficientes**: que sean capaces de comparar sus datos y resultados con otros sistemas de salud, que puedan medir la eficiencia de las innovaciones y detectar las intervenciones de escaso valor, así como poder seleccionar el perfil de pacientes que más se van a beneficiar de determinadas actuaciones asistenciales;

- Salud **centrada en los ciudadanos**: estas tecnologías pueden favorecer la progresiva incorporación de los ciudadanos a la toma de decisiones en salud, así como responder a la exigencia de transparencia en las administraciones en sus relaciones con los ciudadanos (14); permite incorporar los datos de sistemas de monitorización continua de múltiples variables de los usuarios para mejorar los resultados de salud, desarrollar apps móviles para la mejora de la

asistencia o la investigación, o incluso detectar desigualdades en el manejo de pacientes, o centrar las intervenciones en los colectivos más desfavorecidos;

- Sistemas con **orientación a la salud pública**: mediante el cruce de datos poblacionales de múltiples fuentes (datos socioeconómicos; evaluación de intervenciones poblacionales, detección precoz de epidemias y riesgos para la salud) con los datos de la historia clínica electrónica para aumentar el conocimiento en el ámbito de la salud pública;

- Sistemas que favorecen el **desarrollo de los profesionales**: facilitando la innovación y la incorporación del conocimiento a la práctica clínica mediante la difusión por la organización sanitaria y la formación de los profesionales; evitando tareas repetitivas de escaso valor que permitan a los profesionales dedicarse a los aspectos más personales de la asistencia (20).

5. Oportunidades del *Big Data* Sanitario en medicina personalizada

Además de la aplicación a la mejora del sistema sanitario en su conjunto, estas técnicas se aplican también a los modelos personalizados de medicina. La medicina personalizada o de precisión, consiste en la adaptación del tratamiento médico a las características individuales de cada paciente y de su enfermedad, como consecuencia de la integración de diferentes fuentes de datos a los modelos de enfermedad (21). De este modo, las posibles aplicaciones más destacadas de las tecnologías *Big Data* en este ámbito serían (19):

- Modelos **más completos de enfermedad**, desde el nacimiento hasta la muerte y desde la molécula a la sociedad, que incluyan los datos de diferentes fuentes y no únicamente de la historia clínica;

- Descubrimiento de **procesos biológicos fundamentales**, cruzando secuencias exómicas (regiones de los genes que codifican las proteínas) con los datos longitudinales de las historias clínicas;

- Desarrollo de **nuevas asociaciones** patológicas o terapéuticas analizando datos a gran escala de individuos que desarrollan patologías o cruzando información de la efectividad y eficiencia de los diferentes tratamientos;

- Descubrimiento de **nuevos mecanismos y diferentes subtipos de enfermedad** relevantes para el tratamiento, por ejemplo, variantes de tumores que responden de forma distinta a los diferentes tratamientos;

- Desarrollo de nuevas **dianas farmacológicas** o descubrimiento de nuevas indicaciones para fármacos ya existentes, acortando los procesos de desarrollo de nuevos fármacos;

- **Implantación de la medicina de precisión**, integrando el genoma en la historia clínica electrónica para elaborar modelos predictivos de enfermedad, utilizar los fármacos más eficaces y seguros para cada paciente, o incluso poner en marcha actividades de prevención del riesgo cardiovascular o de otro tipo más adecuadas al perfil genético (22);

- **Estimación personalizada** de riesgos y beneficios de las diferentes intervenciones sanitarias.

6. Riesgos y limitaciones de la tecnología *Big data*

Sin embargo, la aplicación de esta nueva tecnología no está exenta de amenazas. Como se ha mencionado, la utilización de tecnología *Big Data* supone unos riesgos para la privacidad que se pueden mitigar, pero no eliminar del todo, en vista de las posibilidades de reidentificación que ofrece el avance tecnológico.

Por este motivo, como se desarrolla a continuación, la legislación en materia de protección de datos ha evolucionado hacia un enfoque más proactivo desde la entrada en vigor del Reglamento Europeo (UE) 2016/679 general de protección de datos (en adelante, RGPD). Este enfoque está basado en los principios de la protección de datos: licitud, lealtad y transparencia del tratamiento (incluyendo el consentimiento y el derecho de información); la limitación de la finalidad; la minimización de los datos; la exactitud de los datos; la limitación del plazo de conservación y la integridad y confidencialidad de los datos. Por otra parte, el RGPD establece que el responsable del tratamiento será el responsable del cumplimiento de estos principios y capaz de demostrarlo (responsabilidad proactiva), lo que incluye la protección de datos desde el diseño y por defecto, la evaluación previa del impacto en los tratamientos que puedan suponer un alto riesgo para las libertades de

los sujetos o incluso, si es necesario, la consulta previa a la autoridad de protección de datos.

Son especialmente preocupantes las situaciones en las que una base de datos obtenida de forma lícita se combina posteriormente con otras bases de datos, de tal manera que se puedan reidentificar a los sujetos o se puedan utilizar estos datos para elaborar perfiles automatizados. A continuación, se mencionan algunas de las situaciones de riesgo que se pueden dar con el uso de estas tecnologías:

- Datos inicialmente anónimos por lo que no están sujetos a revisión ética o control legal que posteriormente se pueden reidentificar cruzándolos con otras bases de datos.

- Bases de datos obtenidas con un propósito lícito en las que se reutilizan los datos para usos secundarios no supervisados, tal y como podría ocurrir cuando la información obtenida para investigaciones previas o para la asistencia se transfiera a terceros o se reutilice con otros fines sin mediar un nuevo consentimiento.

- Transmisión de información a países cuya legislación no garantiza un nivel adecuado de protección de datos o a empresas con intereses comerciales sin el consentimiento de los sujetos.

- Generación de perfiles de riesgo de padecer enfermedades, determinadas condiciones o la puesta en marcha de sistemas de monitorización de conducta humana, lo que se puede realizar incluso con datos anónimos, exponiendo a los sujetos a situaciones de discriminación.

- Posibilidad de estigmatizar poblaciones vulnerables o mantener o incrementar desigualdades sociales, incluso en estudios con datos anonimizados.

- Considerar los datos disponibles en redes sociales como públicos, sin tener en cuenta la finalidad para la cual se hicieron públicos, utilizándolos para otros usos no lícitos o sin autorización.

- Conflictos en la legislación aplicable cuando la información se deposita en los entornos en línea, que pueden estar ubicados en terceros países con legislaciones que garantizan un menor nivel de protección de lo establecido en el RGPD.

- Propiedad de los datos depositados en redes sociales, en los que se autorizan unas condiciones de uso de los datos sin prestar atención a los términos por falta de una cultura de privacidad de los usuarios, teniendo en cuenta que dicha autorización se plantea siempre como un requisito obligatorio para acceder al servicio.

- Movimientos a favor de la transparencia de la información, que proponen que toda la información pertinente debe estar a disposición del público, incluyendo la exigencia de los editores de las publicaciones científicas de hacer disponibles las bases de datos originales de los estudios como medida para prevenir el fraude científico (23).

- Falta de formación, tanto de investigadores como de los miembros de los Comités de Ética de la Investigación, sobre las tecnologías de *Big data* y sus implicaciones éticas y legales.

Otra posible fuente de conflictos surge por el aprovechamiento comercial de los datos, dado que los grandes sistemas de bases de datos de salud pertenecen en su mayor parte a los sistemas públicos, mientras que la tecnología *Big Data* es propiedad generalmente de empresas privadas. Este hecho puede originar problemas éticos, legales e incluso políticos, cuando la cesión de datos de la Administración puede resultar en beneficios económicos a terceros.

Una limitación de los sistemas predictivos es que pueden ser útiles para el nivel de organización sanitaria, pero su utilización en pacientes individuales todavía no está contrastada respecto a la referencia, que es el juicio clínico de un profesional sanitario. Teniendo en cuenta que estas tecnologías se basan en predicciones a partir del procesamiento de datos retrospectivos, es lógico concluir que presentan limitaciones intrínsecas cuando se enfrentan a situaciones cambiantes o novedosas (11).

Otro riesgo que han señalado algunos autores es la probabilidad de que las aplicaciones de inteligencia artificial puedan poner en riesgo las interacciones personales que se dan en la asistencia entre facultativos y pacientes, influyendo desfavorablemente en las experiencias de ambos grupos. Estas amenazas, traducidas en desempleo médico y pérdida de habilidades clínicas, parecen exageradas. Del mismo modo que la toma automatizada de la tensión arterial o el recuento celular automatizado en el laboratorio han liberado a los clínicos de tareas repetitivas, la incorporación de la inteligencia artificial podría volver a dar sentido a la práctica de la medicina al tiempo

que se consiguen nuevos niveles de eficiencia y precisión. Los médicos deben guiar, supervisar y monitorizar, de una forma proactiva, la incorporación de la inteligencia artificial como un complemento en el cuidado del paciente (20,24).

7. Retos futuros del *Big Data* en la práctica clínica

La aplicación de técnicas de *Big Data* en el sector sanitario se enfrenta a algunos retos tecnológicos como los que se mencionan a continuación (2):

- Integrar la información de las diferentes fuentes disponibles (sanitario, socioeconómico, genómica y proteómica, imagen, redes sociales, sensores, aplicaciones móviles,…);
- Documentar la información de manera digital sin que requiera un esfuerzo añadido a los profesionales de la salud (mediante sistemas de codificación automática y de extracción de información a partir de texto libre);
- Analizar y procesar los datos no estructurados de salud (imagen, texto);
- Disponer de sistemas de almacenamiento físico de datos que puedan albergar una cantidad ingente de información;
- Disponer de medios técnicos y legales para compartir e intercambiar datos;
- Disponer de sistemas para asegurar la calidad de los datos (especialmente cuando se toman decisiones relevantes a partir de los resultados);
- Disponer de sistemas que aseguren la confidencialidad de los datos.

8. Experiencias de *Big Data* con grandes bases de datos sanitarios

Algunas de las experiencias infructuosas en el pasado con *Big Data* sanitario pueden servir como motivo de reflexión sobre las implicaciones éticas de estos proyectos.

El proyecto care.data fue promovido por el Servicio Nacional de Salud de Inglaterra (National Health Service England) en enero 2014 para utilizar los datos de las historias clínicas electrónicas de Atención Primaria con fines de investigación biomédica. Antes de su puesta en marcha se enviaron un total de 22 millones de folletos para informar a la población general del alcance del proyecto (25), a partir de los cuales se recibieron más de un millón de solicitudes de rechazo a la utilización de datos por la investigación. Posteriormente trascendió a los medios de comunicación que no se había respetado la opinión de los usuarios que se habían negado al tratamiento de sus datos (26). El proyecto se canceló en julio 2016 por la recomendación de la Autoridad de Protección de Datos de Salud Británica de establecer un modelo claro de oposición al tratamiento que abarcase diferentes opciones de retirada del consentimiento (27).

Una experiencia similar fue el proyecto VISC+ (Más valor a la información de salud en Cataluña) promovido por la Agencia de Calidad y Evaluación Sanitaria de Cataluña. Este proyecto consistía en contratar a una empresa privada para que desarrollase una plataforma para ceder los datos anonimizados del sistema sanitario público catalán, tanto para realizar proyectos de investigación como para otros fines con ánimo de lucro relacionados con las empresas farmacéuticas o biotecnológicas. Los diferentes informes que se elaboraron sobre los problemas éticos que planteaba este proyecto coincidieron en la preocupación suscitada por las posibles vulneraciones de los derechos de los ciudadanos y la falta de transparencia y debate público informado del proyecto (28). Finalmente, debido a la contestación pública y política, se desestimó el proyecto antes de su adjudicación (29).

El proyecto VISC+ ha sido sustituido por el programa PADRIS (Programa de Analítica de Datos para la Investigación e Innovación en Salud), gestionado exclusivamente por entidades públicas. Los datos anonimizados del PADRIS se suministrarán únicamente a los centros de investigación acreditados, públicos o sin ánimo de lucro, mediante una solicitud que incluya el protocolo de investigación y la aprobación previa por un Comité de Ética de la Investigación (CEI) (30). Aunque se prioriza la investigación independiente con financiación pública, no se impide el acceso a la investigación con financiación privada o mediante fórmulas de mecenazgo (31).

Otros proyectos de *Big Data* sanitaria son el proyecto CALIBER, plataforma promovida por el Servicio Nacional de Salud de Inglaterra, que ofrece variables "listas para la investigación" extraídas de diferentes bases de datos clínicas interrelacionadas. Esta plataforma incluye datos clínicos de atención primaria y

especializada sobre diagnósticos, constantes clínicas, resultados de laboratorio, prescripción farmacológica, vacunaciones, además de causas de fallecimiento y datos de deprivación social. Esta base de datos incluye información de 10 millones de personas con un seguimiento total de 400 millones de años-paciente. Las solicitudes de proyectos de investigación, que deben ser aprobadas por un Comité asesor científico independiente, tienen que acompañarse de la memoria del proyecto, así como del curriculum del investigador (32).

Las historias clínicas electrónicas de los Servicios Autonómicos de Salud son grandes repositorios de información que incluyen los datos de atención primaria y especializada, prescripción farmacéutica, resultados de laboratorio e información sociodemográfiica. Sin embargo, aún están pendientes de desarrollo las aplicaciones informáticas que permitan aprovechar esta información. Un ejemplo es BIFAP (Base de datos para la investigación fármaco-epidemiológica en Atención Primaria), gestionada por la Agencia Española del Medicamento y Productos Sanitarios con la colaboración de 9 Comunidades Autónomas, a la que aportan sus datos 6.800 médicos de familia y pediatras con más de 9 millones de historias clínicas anonimizadas (33).

Otros ejemplos de aplicación del *Big Data* sanitario son los sistemas comerciales que analizan los datos sanitarios a partir de la información ya disponible. IBM ha desarrollado la herramienta Watson Health, un sistema de inteligencia artificial aplicable al diagnóstico y tratamiento de enfermedades, a la gestión de servicios sanitarios, al desarrollo de nuevos fármacos o a la selección de participantes en ensayos clínicos (34). Una iniciativa similar es la propuesta por Savana, empresa española que ofrece unos módulos de Consulta (guías clínicas con datos en tiempo real); Research (extracción de variables automáticas para la investigación); Manager (descripción de patrones de consumo de recursos); Predict (herramienta de predicción de resultados) o EHRead (extractor de información clínica a partir de la historia clínica en castellano) (35). Otras herramientas están dedicadas exclusivamente a la investigación, como el Researchkit de Apple, que permite interactuar con los participantes en un proyecto de investigación, tanto para el reclutamiento como para la recogida de datos de forma automática o autorreferida por el participante (36).

9. Aplicación de la tecnología *Big Data* a la investigación clínica

Los sistemas de *Big Data* tienen la posibilidad de cambiar tanto la asistencia como la forma en la que se genera el conocimiento en que se basa la práctica clínica.

El modelo tradicional de investigación parte de una hipótesis que se intenta confirmar diseñando un experimento, incorporando posteriormente a la asistencia aquellas intervenciones que han demostrado su efectividad. El nuevo paradigma de la investigación biomédica, a partir de la implantación de tecnología *Big Data*, es la generación de hipótesis a partir de los datos obtenidos en la práctica clínica, que posteriormente se deben confirmar mediante una investigación (37).

La aplicación de la tecnología *Big Data* en la investigación biomédica permite realizar de forma relativamente sencilla estudios que hasta hace poco tiempo hubieran sido implanteables. De este modo, la disponibilidad de datos masivos procedentes de la asistencia sanitaria simplifica la realización de estudios observacionales, tanto retrospectivos como prospectivos. Son estudios en los que la información no se recoge con motivos de investigación, sino que se obtiene de la práctica clínica diaria (datos de la vida real o real world data) (16,38).

Como ejemplos de posibles diseños de estudios utilizando esta tecnología se pueden citar los siguientes:

- Estudios de cohortes prospectivas con seguimiento naturalístico a través de los datos del sistema de información sanitario, recogidos originalmente para la asistencia.

- Estudios de cohortes retrospectivas, seleccionando pacientes a partir una característica sucedida en el pasado y realizando el seguimiento desde ese momento hasta la actualidad.

- Estudios de casos y controles anidados en cohortes.

- **Ensayos clínicos pragmáticos** (aleatorizados), con seguimiento naturalístico (39). Este tipo de ensayos clínicos prometen ser uno de los avances con mayor capacidad de transformar la investigación clínica, cuyas ventajas más importantes serían las siguientes (40):

- Permite reclutar pacientes excluidos habitualmente de los ensayos clínicos (por ejemplo, grandes comorbilidades, enfermedad renal crónica avanzada o mujeres en edad fértil).

- Permite analizar variables de resultados duras (fallecimiento, entrada en diálisis), que con otros diseños de ensayos clínicos necesitan un

número elevado de pacientes seguidos durante mucho tiempo, lo que encarece extraordinariamente la investigación.

- Aumenta la eficiencia de los estudios buscando variables de resultado diferentes o seleccionando pacientes según características que hagan que los resultados sean más probables.

- Permite sistemas de ayuda a la selección de participantes utilizando algoritmos de análisis de lenguaje natural a partir de los evolutivos en texto libre de las historias clínicas electrónicas.

- Ofrece datos de efectividad en la práctica real, con información de los riesgos y beneficios a largo plazo de las intervenciones.

- Permite aportar datos en situaciones en las que no es posible realizar un ensayo clínico convencional, por no ser ético o impracticable.

- **Estudios exploratorios de relaciones entre datos** (data mining, data driven). Los nuevos modelos de investigación basada en datos parten de una hipótesis, pero en vez de diseñar un experimento para confirmarla, los investigadores analizan los datos disponibles y entonces diseñan experimentos o incluso utilizan datos adicionales para validar la hipótesis planteada. Los modelos guiados por datos (data driven) dan un paso adicional analizando los datos existentes sin una hipótesis fijada, este análisis genera una hipótesis que luego se confirma con el análisis de nuevos datos (38).

- **Estudios que utilizan técnicas de inteligencia artificial** (machine learning, redes neuronales), que constituyen los avances más significativos incorporados a la investigación. En estos estudios se procesan grandes cantidades de datos para entrenar un modelo de inteligencia artificial, que se utiliza para evaluar nuevos datos y ayudar a la toma de decisiones. Algunas de las herramientas más avanzadas de inteligencia artificial, tales como las redes neuronales profundas, no requieren una comprensión del mecanismo subyacente del modelo para lograr resultados predictivos increíblemente precisos. Este enfoque podría ser una herramienta valiosa para explorar datos cruzados de diferentes fuentes y apoyar decisiones clínicas, incluso cuando muchas interconexiones y relaciones entre estos datos son todavía desconocidas o ausentes (8,10,38).

Estas técnicas de inteligencia artificial son de algún modo contradictorias con el principio de minimización de datos que marca la legislación de protección de datos, dado que cuando se utilizan estas técnicas no se conocen a priori las variables que pueden aportar valor a los modelos predictivos.

En resumen, los estudios que utilizan la tecnología *Big Data* pueden ser de gran utilidad, tanto para mejorar los diseños de la investigación como para facilitar la realización de estudios en aspectos tales como la selección de participantes o el seguimiento electrónico.

10. Limitaciones de la investigación con *Big Data*

Sin embargo, la investigación con "datos de la vida real" tiene algunas limitaciones intrínsecas.

En primer lugar, todos los estudios observacionales incluyen sesgos difíciles de controlar ante la falta de aleatorización. Estos sesgos se pueden controlar con los ensayos clínicos pragmáticos aleatorizados, en los que la intervención se asigna al azar pero el seguimiento se hace en condiciones de práctica clínica real.

Una de las limitaciones principales de la investigación observacional es la dificultad de establecer relaciones de causalidad. En este tipo de investigaciones no se puede tener la certeza de que una asociación sea debida a una relación causa-efecto o por el contrario, a algún factor de confusión relacionado simultáneamente con las dos variables estudiadas, aunque se intente hacer un ajuste por las variables que el investigador considere más significativas (41). Otra causa importante de confusión es el sesgo de observación, por el cual a los pacientes sometidos a un seguimiento más intenso por su comorbilidad o por estar en tratamiento con un fármaco se les detecta más efectos secundarios que a una población control (39). En los estudios observacionales con seguimiento también puede aparecer el sesgo de selección, en el que los pacientes prevalentes de una condición, o los que ya están recibiendo un tratamiento, se comportan de forma diferente a aquellos en los que se acaba de hacer el diagnóstico o en los que se comienza el tratamiento, por lo que estos estudios deberían ajustar el resultado para los casos prevalentes e incidentes (42).

Otra limitación de los estudios observacionales es que los acontecimientos que definen el resultado estudiado se recogen a partir de la codificación de los

diagnósticos y procedimientos de los episodios asistenciales, por lo que dependen de la exhaustividad de esta codificación, que suele ser adecuada para los diagnósticos principales, pero puede ser menor para algunas condiciones crónicas coexistentes (hipertensión arterial, diabetes o tabaquismo) o para los diagnósticos que no están directamente relacionados con el motivo del ingreso. Por otra parte, en los estudios observacionales multicéntricos, puede haber diferencias de criterio en la codificación en los diferentes centros, lo que introduce sesgos sobre las variables de resultado (43).

Algunos autores han llamado la atención sobre los sesgos potenciales inherentes a los estudios con técnicas de inteligencia artificial, incluyendo la posibilidad de aumentar las desigualdades sociales (24). Algunas propuestas para reducir los errores en los algoritmos de aprendizaje de máquina serían la prevención de la confianza excesiva en la automatización, prevenir los sesgos en los datos que alimentan los algoritmos y tener en cuenta que los algoritmos pueden descubrir relaciones estadísticamente significativas, pero sin significación clínica. Se debería prestar atención a los datos que se utilizan para ajustar los algoritmos de inteligencia artificial, para saber qué datos y qué pacientes están ausentes y evitar así que las disparidades existentes en la asistencia aumenten por una confianza irreflexiva o excesiva en la computación (13).

Con el fin de mejorar la publicación de los estudios observacionales realizados a partir de los datos obtenidos de la historia clínica, se han publicado las guías RECORD (REporting of studies Conducted using Observational Routinely-collected health Data statement), que incluyen un listado de comprobación con los apartados que deberían incluirse en este tipo de publicaciones, para que se pueda valorar la validez interna y externa de los resultados (44). Estas guías son una adaptación de las recomendaciones previas STROBE (Strengthening the Reporting of OBservational studies in Epidemiology), dirigidas a la publicación de estudios observaciones en epidemiología (45).

Como conclusión, los estudios con datos de la vida real no sustituyen los datos de seguridad y eficacia obtenidos en los ensayos clínicos aleatorizados, pero sí que pueden contribuir a aumentar su validez externa.

11. Principios éticos en la investigación con *Big Data*

La irrupción de estas nuevas tecnologías ha supuesto que algunos autores se planteen si la ética de la investigación está obsoleta en la época del *Big*

Data (46). Sin embargo, aunque su aplicación deba adaptarse a esta nueva tecnología, se puede afirmar que los principios de la bioética, establecidos en el informe Belmont y desarrollados posteriormente en las diferentes pautas éticas internacionales sobre investigación biomédica, siguen siendo igualmente aplicables a estos estudios, incluyendo el respecto a la dignidad, autonomía, privacidad y confidencialidad de los participantes en una investigación (47,48,49,50,51). De este modo, las dimensiones éticas relevantes implicadas en este tipo de estudios son las siguientes (48):

Respeto por las personas: cualquier iniciativa que incluya el uso de datos tiene que considerar tanto los intereses privados como los públicos. Se debe permitir a los interesados expresar sus preferencias e informarles sobre la utilización de sus datos.

Respeto por los derechos humanos: los términos de cualquier uso de datos deberían respetar los derechos básicos, tales como la protección de la vida privada o familiar. Esto incluye establecer limitaciones a los gobiernos a interferir con la privacidad de los ciudadanos por motivos de interés público.

Participación e inclusión de los interesados: la utilización de datos debe tener en cuenta los deseos de los ciudadanos y, en lo posible, hacer participar a las personas y las comunidades a las que pertenecen en las decisiones acerca de sus datos.

Responsabilidad y rendición de cuentas: la utilización de datos debería rendir cuentas de forma pública mediante procedimientos legales, penales y políticos, además de rendir cuentas socialmente mediante el compromiso periódico de recoger las expectativas de la población sobre la utilización de sus datos. Las iniciativas que utilizan datos deben informar a los afectados de los fines del tratamiento, incluyendo información transparente sobre las posibles vulneraciones de la confidencialidad y cualquier desviación con la gobernanza establecida (48,50).

En este sentido, las iniciativas de utilización de *Big Data* sanitario deberían contar con el compromiso de la comunidad que aporta sus datos para contribuir a la mejora del sistema sanitario, a cambio del compromiso de los investigadores de que los datos estarán protegidos de revelaciones indebidas. Por este motivo, es imprescindible poner en marcha sistemas de gobernanza del *Big Data*, tanto para los responsables de los datos como para los encargados del tratamiento. Se debe elevar la exigencia a los profesionales

que manejan datos mediante políticas de acceso y privacidad que incluyan consecuencias rápidas y graves para el mal uso o la revelación de datos (52).

Teniendo en cuenta los beneficios esperables de la utilización de datos masivos, los sistemas de gobernanza de las bases de datos deben basarse preferentemente en los beneficios que puede aportar la investigación a la sociedad, sin olvidar los posibles perjuicios que pueda causar.

El cumplimiento de estos principios debería reflejarse en la gobernanza de datos que debería recoger al menos los siguientes aspectos clave (48):

- **Las disposiciones para el almacenamiento de datos** (dónde se custodian, con qué medidas de seguridad, por cuánto tiempo).

- **El acceso o transferencia de los datos** (si se hacen públicos, si se permite un acceso o transferencia controlados o únicamente se suministran datos indirectos).

- **El papel del consentimiento o de otras formas de autorización** (si se solicita consentimiento explícito, implícito dando opción a la no participación o con garantías adicionales como la autorización por comités u otros organismos).

- **Los usuarios autorizados** (incluyendo investigadores académicos, usuarios con intereses comerciales y cómo éstos demuestran (o no) que sus fines tienen un interés público).

Estos principios son similares a los recogidos en el RGPD, aprobado por la Unión Europea como respuesta a los nuevos retos para la protección de datos personales que plantean los avances tecnológicos.

Ya se han descrito precedentes de problemas éticos con el empleo de tecnología *Big Data* en investigación, como muestra de que las precauciones que suscita el uso de esta tecnología no son meramente teóricas. Un ejemplo ilustrativo es la controversia generada por la publicación del resultado de un experimento diseñado con el objetivo de demostrar que es posible producir un contagio emocional masivo sin comunicación directa con los usuarios (53). En este experimento se seleccionaron de forma aleatoria 689.000 usuarios de Facebook, en los que se filtraron las noticias positivas y negativas que recibían en comparación con un grupo control, demostrando que al reducir las noticias positivas, las personas eran menos propensas a

escribir mensajes positivos y viceversa. Los mensajes se calificaban como positivos o negativos mediante un algoritmo automático de análisis del lenguaje, de tal modo que los investigadores no tuvieron acceso al texto de los mensajes. Los investigadores no solicitaron el consentimiento de los usuarios ni se les advirtió de que estaban participando en una investigación, alegando que "el estudio era conforme con la política de utilización de datos de Facebook, aprobada por todos los usuarios antes de abrir una cuenta de Facebook, constituyendo esta aprobación el consentimiento informado para esta investigación" (53).

En una editorial posterior de la misma publicación se expresaron las dudas y la preocupación surgida sobre el respeto a los principios de consentimiento informado y a la opción de rechazar la participación a los implicados, como garantía de las mejores prácticas posibles en investigación. La editorial remarcaba también que el estudio no fue aprobado por un CEI debido a que la investigación fue realizada por Facebook con fines privados (54).

Similares preocupaciones se han suscitado con otros estudios, en los que se han puesto a disposición de la comunidad investigadora información tan sensible como la base de datos de los usuarios de un sitio web de contactos, en la que los usuarios podían identificarse fácilmente, con el argumento de que la información que se proporciona en estas páginas es pública (46).

12. Control ético y legal de los proyectos *Big Data*

Teniendo en cuenta los posibles riesgos mencionados previamente, el control ético y legal de los proyectos de investigación que utilizan tecnologías *Big Data* ha sido un motivo de preocupación para los Comités de Ética de la investigación y los propios investigadores.

La entrada en vigor del RGPD y de la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales (LOPD-GDD) ha cambiado el panorama legislativo de la utilización de datos relacionados con la salud en investigación, tanto para la utilización de datos identificados, seudonimizados e incluso anonimizados.

El consentimiento para utilizar los datos para una investigación se obtiene generalmente en el momento de recabarlos, teniendo en cuenta que la LOPD-GDD permite que este consentimiento para investigación sea con fines amplios. Cuando se plantea reutilizar los datos obtenidos previamente surgen

dos problemas fundamentales: determinar si el consentimiento otorgado aún refleja los deseos del sujeto y si los nuevos usos propuestos se consideran incluidos en el consentimiento original, por este motivo la LODP-GDD exige que la reutilización de datos de investigaciones previas deba estar supervisada por un CEI cuando la recogida de datos sea posterior a la entrada en vigor de la LOPD-GDD.

Hasta la entrada en vigor de esta Ley, la utilización de datos anonimizados carecía de regulación legal, al no considerarse datos de carácter personal. Sin embargo, la LOPD-GDD establece, de acuerdo con el RGPD, que los estudios con datos relacionados con la salud tengan una evaluación previa del impacto que incluya, "*de modo específico, los riesgos de reidentificación vinculados a la anonimización y seudonimización de los datos*". La Ley también prevé que el uso de datos personales seudonimizados con fines de investigación biomédica deberá ser sometido al informe previo de un CEI. Este dictamen también sería aplicable a las investigaciones que utilicen datos anonimizados, tal como se desprende del enunciado de la LOPD-GDD, que exige la incorporación de un delegado de protección de datos o un experto en protección de datos a la composición de los comités de ética de la investigación "*cuando se ocupen de actividades de investigación que comporten el tratamiento de datos personales o de datos seudonimizados o anonimizados*".

La LOPD-GDD establece también el régimen sancionador en las infracciones a la protección de datos, considerando como infracciones muy graves tanto la reversión deliberada de un procedimiento de anonimización, a fin de permitir la reidentificación de los afectados, como la utilización de datos disponibles para otros fines, no autorizados o lícitos, o la transferencia internacional de datos cuando no concurran las garantías, requisitos o excepciones establecidos en el RGPD.

13. La utilización de datos anonimizados

Tal como se ha comentado, el paradigma de la anonimización de los datos como garantía absoluta de privacidad tiene actualmente serias limitaciones, dado que las posibilidades de reidentificación que ofrece la informática son cada vez mayores, dependiendo de la información adicional disponible (tanto en el momento actual como en un futuro) y del esfuerzo que se dedique.

La única forma de garantizar la imposibilidad de reidentificar a un individuo es recodificar y agrupar las variables que podrían identificar eventualmente a un individuo, tales como la fecha de nacimiento, el lugar de residencia o la fecha de un ingreso hospitalario (4,55). Cuando se llevan a cabo estas técnicas de anonimización, siempre se distorsiona y pierde parte de la información (el llamado diferencial de privacidad), de tal manera que deberá llegarse a un compromiso entre los fines del tratamiento y las amenazas para la privacidad (55). Una garantía adicional para evitar la reidentificación es que las personas implicadas en el tratamiento de la información personal anonimizada no tengan acceso a los datos personales no anonimizados o que no conozcan los mecanismos y claves de anonimización utilizados (55).

El RGPD exige que se garantice el respeto al principio de minimización de datos personales en el tratamiento de datos con fines de investigación biomédica. El RGPD especifica que se pongan en marcha medidas como la seudonimización, siempre que la investigación se pueda realizar de esta forma, pero que si la investigación se puede realizar con datos anonimizados, deberá llevarse a cabo de ese modo.

La LOPD-GDD ha previsto la utilización de datos seudonimizados o anonimizados para investigación sin el consentimiento del sujeto, siempre que se pongan en marcha las medidas técnicas y organizativas suficientes para prevenir la reidentificación y que los investigadores no tengan acceso a los datos identificativos. Estas medidas incluyen una evaluación del impacto en la protección de datos cuando sea necesario, así como desarrollar un enfoque proactivo en la protección de datos, incluyendo las medidas de protección por defecto y desde el diseño de la investigación. Como garantía adicional, la LOPD-GDD exige que estos proyectos cuenten con el informe previro del comité de ética de la investigación.

En resumen, aunque la entrada en vigor de la LOPD-GDD ha supuesto un gran avance en la regulación de la investigación con datos relacionados con la salud, incluyendo el *Big data* sanitario, todavía existen dudas en la interpretación del texto legal, especialmente en cuanto al tratamiento de datos anonimizados. En ausencia de un desarrollo mediante Reglamento como la anterior LOPD, deberá esperarse la interpretación de los informes que emitan las Autoridades de protección de datos.

14. Conclusiones

- La utilización de la tecnología *Big data* en la práctica clínica puede aportar grandes beneficios para la atención sanitaria y la investigación clínica.

- Para conseguir estos objetivos es fundamental la integración de las diferentes fuentes de información.

- Es necesario garantizar el uso y la aplicación de la tecnología *Big data* en un marco ético y jurídico adecuado, tanto en asistencia como en investigación.

- La utilización de estas tecnologías puede suponer una fuerte amenaza para la privacidad, por lo que es esencial poner en marcha procedimientos de control de la información que sean adecuados, transparentes y seguros y que garanticen el máximo nivel de confidencialidad y el respeto a los derechos y libertades de las personas.

- El personal sanitario, los investigadores, los Comités de ética de la investigación y las autoridades sanitarias y de protección de datos tienen como deber compartido el compaginar de forma equilibrada la protección de datos con la investigación relevante y la práctica clínica.

- La reciente aprobación del RGPD y la LOPD-GDD ha incorporado la regulación del uso de datos a gran escala con fines asistenciales y de investigación, incluyendo los datos identificados, seudonimizados y anonimizados.

15. Bibliografía

1. Aced, E. / Heras, M.R. / Sáiz, C.A., *Código de buenas prácticas en protección de datos para proyectos Big Data,* Agencia Española de Protección de Datos [En línea, consultado el 28/02/2019]. Disponible en: www.agpd.es.
2. Menasalvas, E. / Gonzalo-Martín, C. / Rodríguez-González, A., "Big Data en salud: retos y oportunidades", *Economía Industrial*, Vol. 405, 2017, pp. 87-97.
3. European Commission, *eHealth Task Force Report - Redesigning health in Europe for 2020,* Luxembourg, 2012.

4. Article 29 Data Protection Working Party, *Opinion 05/2014 on anonymisation techniques*, 2014 [En línea, consultado el 28/02/2019]. Disponible en: https://ec.europa.eu/justice/article-29/documentation/
5. Sweeney, L., *Simple demographics often identify people uniquely*, Carnegie Mellon University, Data Privacy Working Paper 3, 2000.
6. Goodfellow, I. / Bengio, Y. / Courville, A., *Deep Learning*. In. Cambridge, Mass: MIT Press; 2016, pp. 1-26 [En línea, consultado el 28/02/2019] Disponible en: http://www.deeplearningbook.org/contents/intro.html
7. Naylor, CD., "On the Prospects for a (Deep) Learning Health Care System", *JAMA*, Núm. 11, Vol. 320, 2018, pp. 1099-1100.
8. Chartrand G. / Cheng, P.M. / Vorontsov, E. / Drozdal, MTS. / Pal, C.J. / Kadoury, S. *et ál*. "Deep Learning: A Primer for Radiologists", *RadioGraphics*, Vol. 37, 2017, pp. 2113-31.
9. Hinton, G., "Deep Learning–A Technology with the potential to transform health care", *JAMA,* Núm. 11, Vol. 320, 2018, pp. 1101-2.
10. Beam, A.L. / Kohane, I.S., "Big Data and Machine Learning in Health Care", *JAMA,* Núm. 13, Vol. 319, 2018, pp. 1317-8.
11. Stead, WW., "Clinical implications and challenges of artificial intelligence and deep learning", *JAMA*, Vol. 320, 2018, pp. 1107-8.
12. Rajkomar A., "Scalable and accurate deep learning with electronic health records", *npj Digital Med*. Núm. 18, Vol. 1, 2018.
13. Gianfrancesco, M.A. / Tamang, S. / Yazdany J. / Schmajuk, G., "Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data", *JAMA Intern Med*., Núm. 11, Vol. 178, 2018, pp. 1544-47.
14. Murdoch, T.B. / Detsky A.S., "The Inevitable Application of Big Data to Health Care", *JAMA*, Núm. 13, Vol. 309, 2013, pp. 1351-2.
15. Clasificación Internacional de Enfermedades. Organización Panamericana de la Salud [En línea, consultado el 28/02/2019]. Disponible en: https://www.paho.org/hq/
16. García López JL. / Del LLano Señarís, J.E. / Del Diego Salas J. / Recalde Manrique, J.M., *Aportación de los "Real World Data" a la mejora de la práctica clínica y del consumo de recursos de los pacientes*, Fundación Gaspar Casal, 2014.
17. Comité de Ética del Instituto de Investigación de Enfermedades Raras, "Guías Éticas de Investigación en Biomedicina", 2009 [En línea, consultado el 29/02/2019]. Disponible en: http://publicaciones.isciii.es/
18. Universidad Politécnica de Madrid, *Big Data en Salud*, Informe de vigilancia tecnológica, 2015 [En línea, consultado el 28/02/2019]. Disponible en: https://fipse.es/sites/default/files/documentos/
19. Hemingway, H. / Asselbergs, F.W. / Danesh, J. / Dobson, R. / Maniadakis, N. / Maggioni, A. *et ál.,* "Big data from electronic health records for early and late translational cardiovascular research: challenges and potential", *European Heart Journal*, Núm. 16, Vol. 39, 2017, pp. 1481-95.
20. Verghese, A. / Shah, N.H. / Harrington, R.A., "What This Computer Needs Is a Physician Humanism and Artificial Intelligence", *JAMA*, Núm. 1, Vol. 319, 2018, pp. 19-20.

21. Sociedad Española de Oncología Médica, [En línea, consultado el 28/02/2019]. Disponible en: https://seom.org/informacion-sobre-el-cancer/ique-es-la-medicina-de-precision
22. Cabaleiro, T. / Prieto-Pérez, R. / Ochoa, D. / Abad-Santos, F., "Aplicación de la farmacogenómica y otras nuevas tecnologías al desarrollo de medicamentos", *Medicina Clínica*, Núm. 12, Vol. 140, 2013, pp. 558-63.
23. The Center for Scientific Integrity. Retraction Watch. [En línea, consultado el 28/02/2019]. Disponible en: https://retractionwatch.com
24. Israni, S.T. / Verghese, A., "Humanizing Artificial Intelligence", *JAMA*, Núm. 1, Vol. 321, 2019, pp. 29-30.
25. Nhs Choices, *National Health Service England. Better information means better care*, 2014 [En línea, consultado el 28/02/2019]. Disponible en: https://www.england.nhs.uk/wp-content/uploads/2014/01/cd-leaflet-01-14.pdf
26. Ramesh, R. "NHS disregards patient requests to opt out of sharing medical records", *The Guardian*. 22/01/2015.
27. Nhs England, *care.data NHS England*, 2016 [Consultado el 1/03/2019]. Disponible en: https://www.england.nhs.uk/2013/10/care-data/
28. Llácer, M.R. / Casado, M. / Buisan, L., *Documento sobre bioética y Big Data de salud: explotación y comercialización de los datos de los usuarios de la sanidad pública,* Dret OdBi, editor. Barcelona: Publicacions i Edicions de la Universitat de Barcelona, 2015.
29. Agència de Qualitat i Avaluació Sanitàries de Catalunya (AQuAS), Generalitat de Catalunya, 2016 [En línea, consultado el 28/02/2019]. Disponible en: http://aquas.gencat.cat/es/detall/detall-noticia/VISC_cancelacio_dialegCompetitiu.
30. Agència de Qualitat i Avaluació Sanitàries de Catalunya (AQuAS), *Programa de analítica de datos para la investigación y la innovación en salud (PADRIS)* [En línea, consultado 01/03/2019]. Disponible en: http://aquas.gencat.cat/es/ambits/analitica-dades/padris/index.html
31. Agència de Qualitat i Avaluació Sanitàries de Catalunya, Departament de Salut. Generalitat de Catalunya. rograma públic d'analítica de dades per a la recerca i la innovació en salut a Catalunya –PADRIS–. 2017 [En línea, consultado el 28/02/2019]. Disponible en: http://aquas.gencat.cat/web/.content/minisite/aquas/publicacions/2017 /Programa_analitica_dades_PADRIS_aquas2017.pdf .
32. Caliber. Ucl Institute Of Health Informatics. [En línea, consultado el 28/02/2019]. Disponible en: https://www.ucl.ac.uk/health-informatics/caliber.
33. Agencia Española del Medicamento y Productos Sanitarios. BIFAP. [En línea, consultado el 28/02/2019]. Disponible en: http://www.bifap.org/.
34. Ibm. Ibm Watson Health. [En línea, consultado el 28/02/2019]. Disponible en: https://www.ibm.com/watson/health/index-1.html.
35. Savana. [En línea, consultado el 28/02/2019]. Disponible en: https://savanamed.com/es/.
36. Apple. Researchkit. [En línea, consultado el 28/02/2019]. Disponible en: https://www.apple.com/es/researchkit/.

37. Embi, PJ. / Payne, PR., "Evidence Generating Medicine: Redefining the Research-Practice Relationship to Complete the Evidence Cycle", *Medical Care*, Núm. 8 (Supl 3)), Vol. 51, 2013, s87-91.
38. Zhu, L. / Zheng, W.J., "Informatics, Data Science, and Artificial Intelligence", *JAMA*, Núm. 11, Vol. 320, 2018, pp. 1103-4.
39. Mc Cord, K.A. / Salman, R.A. / Treweek, S. / Gardner, HSD. / Whiteley, W. / Ioannidis, JPA. *et ál.*, "Routinely collected data for randomized trials: promises, barriers, and implications", *Trials,* Núm. 29, Vol. 19, 2018.
40. Lauer, M.S. / D'Agostino, R.B., "The Randomized Registry Trial – The Next Disruptive Technology in Clinical Research", *NEJM,* Vol. 369, 2013, pp. 1579-81.
41. Lazarus, B. / Chen, Y. / Wilson, F.P. / Sang, Y. / Chang, A.R. / Coresh, J. *et ál*, "Proton Pump Inhibitor Use and Risk of Chronic Kidney Disease", *JAMA Intern Med*., Núm. 2, Vol. 176, 2016, pp. 238-46.
42. Danaei, G. / Tavakkoli, M. / Hernán, M.A., "Bias in Observational Studies of Prevalent Users: Lessons for Comparative Effectiveness Research From a Meta-Analysis of Statins", *Am J Epidemiol*, Núm. 4, Vol. 175, 2012, pp. 250–62.
43. Li, L. / Rothwell, P.M., "Biases in detection of apparent `weekend effect´ on outcome with administrative coding data: population based study of stroke", *BMJ*, Vol. 353, 2016, i2648.
44. Benchimol, E. I. / Smeeth, L. / Guttmann, A. / Harron, K. / Moher, D. / Peterson, I. / Sorensen, H.T. / Von Elm, E. / Langan, S.M., "Record Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement", *PLoS Med*, Núm. 10, Vol. 12, 2015, e1001885.
45. Von Elm E. / Altman DG. / Egger, M. / Pocock, S.J. / Gotzsche, P.C. / Vanderbroucke, J.P. *et ál.* "The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies", *PLOS Medicine*, Núm. 10, Vol. 4, 2007, e296.
46. Leetaru, K., "Are Research Ethics Obsolete In The Era Of Big Data?", *Forbes*, 17 Junio, 2016.
47. The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. *Office for Human Research Protections. U.S. Department of Health and Human Services. The Belmont Report. Ethical Principles and Guidelines for the Protection of Human Subjects of Research*, 1979 [En línea, consultado el 28/02/2019]. Disponible en: https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html
48. Nuffield Council on Bioethics, *The collection, linking and use of data in biomedical research and health care: ethical issues,* Londres, 2015 [En línea, consultado el 28/02/2019]. Disponible en: http://nuffieldbioethics.org/project/biological-health-data
49. Consejo de Organizaciones Internacionales de las Ciencias Médicas (CIOMS), *Pautas éticas internacionales para la investigación relacionada con la salud con seres humanos,* Ginebra, 2017.

50. Asociación Médica Mundial, *Declaración de la AMM sobre las Consideraciones Eticas de las Bases de Datos de Salud y los Biobancos*, 2017 [En línea, consultado el 28/02/2019]. Disponible en: https://www.wma.net/es
51. Asociación Médica Mundial, *Declaración de Helsinki. Principios éticos para las investigaciones médicas en seres humanos,* 2013 [En línea, consultado el 28/02/2019]. Disponible en: https://www.wma.net/es
52. European Economic and Social Committee, *The ethics of Big Data: Balancing economic benefits and ethical questions of Big Data in the EU policy context*, 2017.
53. Kramer, ADI. / Guillory, J.E. / Hancock, J.T., "Experimental evidence of massive-scale emotional contagion through social networks", PNAS, Núm. 24, Vol. 111, 2014, pp. 8788-9
54. Verma, I.M. Editor in Chief. Editorial Expression of Concern: Experimental evidence of massivescale emotional contagion through social networks. PNAS. 2014; 111(29): 10779
55. Agencia Española de Protección de Datos, *Orientaciones y garantías en los procedimientos de anonimización de datos personales,* 2016 [En línea, consultado el 28/02/2019]. Disponible en: https://www.aepd.es/media/guias/guia-orientaciones-procedimientos-anonimizacion.pdf