Big data y ciencia de los datos para una nefrología personalizada: ¿estamos preparados para una "nefrología inteligente"?

Miguel Hueso[1,2], José Ibeas[2,3], Ignacio Revuelta[2,4], Francisco-J. Santos-Arteaga[2,5], María José Soler[2,6], Juan Manuel Buades[2,7]

[1]Servicio de Nefrología. Hospital de Bellvitge. L'Hospitalet de Llobregat. Barcelona
[2]Grupo de Trabajo BigSEN de la Sociedad Española de Nefrología.
[3]Servicio de Nefrología. Parc Taulí Hospital Universitari. Institut d'Investigació i Innovació Parc Taulí I3PT. Universitat Autònoma de Barcelona. Sabadell. Barcelona
[4]Servicio de Nefrología y Trasplante Renal. Hospital Clínic. Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS). Barcelona
[5]Faculty of Economics and Management. Free University of Bolzano. Bolzano. Italy
[6]Nephrology Research Group. Vall d'Hebron Research Institute (VHIR). Nephrology Department. Hospital Universitari Vall d'Hebron. Universitat Autònoma de Barcelona. Barcelona
[7]Servicio de Nefrología. Hospital Universitario Son Llàtzer. Palma de Mallorca

NefroPlus 2019;11(2):1-10

RESUMEN

El objetivo de la medicina personalizada es ofrecer a cada paciente un tratamiento adecuado en el momento preciso. En la actualidad, para demostrar que un tratamiento es eficaz es necesario obtener datos de costosos ensayos clínicos controlados, aleatorizados y multicéntricos (medicina basada en la evidencia). Sin embargo, los resultados pueden no demostrar eficacia en la población seleccionada, por lo que en la actualidad se recomienda un análisis previo de datos retrospectivos recogidos en grandes bases de datos *(big data),* que pueden proceder de fuentes heterogéneas (datos de laboratorio, datos moleculares y genéticos, cursos clínicos, imágenes, fármacos, etc.) con el objetivo de conocer qué variables modifican los resultados del tratamiento médico en la "vida real". Para el análisis de este *big data* y para obtener nuevo conocimiento es necesario utilizar nuevas herramientas y potentes tecnologías, generando una "nueva" ciencia basada en los datos. Esta ciencia de los datos se encuentra en continua expansión con el desarrollo de nuevos métodos dedicados a la recogida de datos, su almacenamiento, depuración, procesamiento y análisis. Este conocimiento generado por los datos tendrá un impacto en la práctica clínica con la instauración de tratamientos personalizados, diseño de fármacos inteligentes, identificación de poblaciones de riesgo o rastreo de información en la historia clínica informatizada. El objetivo de esta revisión es mostrar algunas de las oportunidades que ofrece la ciencia de los datos a la nefrología, señalar a qué retos se enfrenta y proponer posibles aplicaciones para una nefrología personalizada.

Palabras clave: *Big data*. *Data science*. Inteligencia artificial. Medicina de precisión. Medicina personalizada. Nefrología.

Correspondencia: Miguel Hueso
Servicio de Nefrología.
Hospital Universitari Bellvitge y Bellvitge Research Institute (IDIBELL).
C/ Feixa llarga, s/n. 08907 L'Hospitalet de Llobregat, Barcelona.
mhueso@idibell.cat

Revisión por expertos bajo la responsabilidad de la Sociedad Española de Nefrología.

DEFINICIÓN DE *BIG DATA* Y DE MEDICINA PERSONALIZADA

El término *big data* (BD) hace referencia a un enorme volumen de datos de formato heterogéneo (estructurados, semiestructurados, como las hojas de cálculo, o desestructurados, como vídeos o documentos de audio) que se generan continuamente y cuyo análisis mediante las tecnologías habituales se ve dificultado[1]. Sus características principales siguen la regla de las 3 v: *v*olumen (gigantesca cantidad de datos generados), *v*elocidad (los

datos se generan rápidamente) y *v*ariedad (datos de diversas fuentes y con múltiples formatos). Con posterioridad se han añadido otras características para mejorar la definición, como son la veracidad, aunque es todavía más un deseo que una realidad, y el valor (valioso para los ciudadanos, los pacientes y su entorno, y los gestores, con el fin de mejorar el conocimiento a través de la información). En la actualidad, el término *big data* se ha expandido para incorporar el análisis e interpretación de los datos, constituyendo una ciencia de los datos *(data science).*

El objetivo de la ciencia de los datos en medicina es contribuir al conocimiento de los mecanismos de las enfermedades para personalizar el tratamiento médico buscando tendencias, asociaciones o patrones en grandes volúmenes de información médica (datos clínicos, moleculares y epidemiológicos), que no se encontrarían en muestras más pequeñas. Este concepto enlaza con la llamada "medicina personalizada", que pretende conseguir un diagnóstico preciso reduciendo la complejidad de las enfermedades a una alteración de algún componente molecular que se pueda medir, para proporcionar un tratamiento personalizado de máxima eficacia, usando la dosis correcta en el momento adecuado y minimizando sus efectos adversos[2]. Para la mayoría de los problemas clínicos esta estrategia es todavía más un deseo que una realidad.

La diferencia entre los ensayos clínicos convencionales aleatorizados, prospectivos y multicéntricos que evalúan la eficacia de un tratamiento y el análisis del BD es que en los primeros el razonamiento es inductivo, de lo particular a lo general, con una hipótesis y unas variables definidas a priori, mientras que el BD presenta un razonamiento deductivo, es decir, de lo general a lo particular, en el que se recopilan todas las variables para la identificación de unos patrones y para generar una hipótesis. En este sentido, se ha propuesto utilizar el análisis de BD para buscar vías moleculares inesperadas en las enfermedades. Evidentemente, cada nuevo mecanismo molecular identificado tendrá que ser validado posteriormente.

FUENTES DE *BIG DATA* EN MEDICINA Y CONCEPTO DE eSALUD: HISTORIA CLÍNICA ELECTRÓNICA, DATOS DE LABORATORIO Y DATOS ÓMICOS, REGISTROS OBTENIDOS POR DISPOSITIVOS MÉDICOS PORTÁTILES, IMÁGENES Y BIOPSIAS DIGITALIZADAS

Se calcula que los datos generados por los dispositivos electrónicos (*electronic health records,* EHR) alcanzará los 25.000 petabytes en el 2020[3]. Industrias como Google o Amazon se han aprovechado del acceso libre y gratuito de los datos proporcionados por el propio consumidor durante búsquedas *online* o para realizar compras para obtener información precisa y personalizada en tiempo real. Sin embargo, la situación es diferente en la sanidad, porque los datos médicos son privados, son custodiados cuidadosamente en el mismo hospital y no se dispone de ellos libremente para que se exploten con métodos de BD. Además, los datos médicos son complejos y requieren de un procesamiento previo para poder analizarlos. Tampoco existe la infraestructura técnica que permita el movimiento, manipulación y gestión de los datos médicos. En este contexto surge el término eSalud *(eHealth),* que se refiere al uso de la tecnología de la información y las comunicaciones (TIC) a la salud[4]. Algunos de los componentes fundamentales de la eSalud son la historia clínica electrónica (EHR), la telesalud (incluye la telemedicina), la mSalud (uso de dispositivos móviles), el *eLearning* (aprendizaje a distancia), la educación continuada en TIC (incluye las publicaciones electrónicas, el acceso abierto, la alfabetización digital y el uso de las redes sociales) y la estandarización e interoperabilidad (hace referencia a la comunicación entre diferentes tecnologías y aplicaciones de *software* para el intercambio y uso de datos para hacer viable la gestión integrada de los sistemas de salud).

El uso de los EHR empezó a generalizarse en los últimos 15 años. Desde el principio se ha considerado de vital importancia que la información disponible en los EHR no solo sirva para el abordaje diario de los pacientes, sino que también se pueda utilizar para la ayuda a la toma de decisiones y para la exploración de la información con fines de gestión, epidemiológicos o de investigación. En los EHR que más frecuentemente se utilizan, aunque hay parte de la información estructurada (principalmente diagnósticos, procedimientos en los informes de alta, etc.), la mayor parte de la información útil está en texto libre. Por ello, se ha realizado un esfuerzo en aplicar una de las ramas de la inteligencia artificial, el procesamiento de lenguaje natural (NLP, del inglés *natural language processing*), que permite explotar información de los informes o anotaciones del paciente en texto libre. Actualmente, en varios hospitales del país ya se usan programas como *Savana*[5], que se utilizan en casos de éxito.

Por otro lado, la necesidad de intercambio de información entre distintos sistemas informáticos hace que los EHR tengan tendencia a intentar aumentar el grado de almacenamiento de la información de forma estructurada y normalizada, siguiendo recomendaciones de normas o estándares internacionales (HL7 V3, openEHR, UNE-EN ISO 13606, ISO 21090: 2011, etc.). Sin embargo, en muchas ocasiones se enfrenta al rechazo de los profesionales de la salud, que se han formado en y para el uso de texto libre. Por ello, es muy importante "implementar" mejoras en la usabilidad que favorezcan su implantación. El objetivo que se persigue es alcanzar la interoperabilidad. La interoperabilidad de la información es la capacidad de un producto de información originado en un sistema A para comunicarse, representarse y utilizarse en otro sistema B sin necesidad de intervención humana. A lo largo de la vida de un ciudadano, implica que su información preserve su valor, aunque cambien los sistemas de información que la gestionan. En términos prácticos, lo que ha de ser interoperable son los extractos de información.

Uno de los aspectos más importantes en el desarrollo de sistemas de historias clínicas electrónicas es cómo modelar la información que contienen. Los modelos de información cumplen este papel. Desde la aparición de un modelo "dual" para el modelado de la historia clínica electrónica, los modelos de información se dividen en modelos de representación de la información, también conocidos como *modelos de referencia* (es la forma en la que se almacena la información en los programas

informáticos para que conste de forma fiel el origen de la información tal y como se originó, es decir, deben ser estables a lo largo de la vida), y modelos de contenido del dominio, también conocidos como *arquetipos o plantillas,* que deben evolucionar según las necesidades de información producidas por la actualización de las guías de práctica clínica. Los arquetipos son un mecanismo de representación formal de conceptos clínicos basado en las entidades del modelo de referencia que puede ser automáticamente procesable por un sistema informático. Es decir, es información que puede ser entendida, no solo recibida, por otro ordenador y, por lo tanto, es inmediatamente procesable. Dentro de los elementos que forman los arquetipos, en las variables más importantes se utilizan terminologías normalizadas, tanto para las etiquetas (p. ej., presión arterial en brazo izquierdo), como para los contenidos (p. ej., el problema de salud enfermedad renal crónica). Las terminologías clínicas son, por tanto, un componente que, por lo general, acompaña a las estructuras o modelos de información clínica, constituyéndose como un elemento de enlace que aporta significado preciso y asegura la interoperabilidad semántica de los datos.

Las terminologías son una de las clases de recursos que se usa en el ámbito de la informática sanitaria (tabla 1). La más conocida y por la que apuesta el Ministerio de Sanidad de nuestro país es SNOMED CT. Es una terminología clínica, no una clasificación como CIE-9, CIE-10, ATC, etc. (las cuales están más enfocadas a usos secundarios de explotación de la información). SNOMED CT es una terminología centrada en el paciente que proporciona conceptos para describir situaciones clínicas con precisión, y se acerca más al lenguaje clínico habitual que, por ejemplo, CIE-10. Permite la utilización de sinónimos, por lo que se adapta a usos y expresiones de carácter más local, así como a diferentes idiomas. Permite mapeos con otros recursos terminológicos y clasificaciones. Abarca múltiples dominios de las ciencias de la salud: medicina, farmacia, enfermería, laboratorio, veterinaria, etc. Incorpora cambios de forma dinámica gracias a su mecanismo de extensiones. Mantiene la trazabilidad de sus componentes a lo largo del tiempo y entre versiones, dado que no permite ni la reutilización ni el borrado de los códigos. Incorpora una gramática para la composición de expresiones clínicas poscoordinadas (es decir, generar un nuevo código mediante la combinación de conceptos en el caso de que no se encuentre ninguno disponible). Actualmente, el Ministerio de Sanidad, al ser España miembro de SNOMED International, proporciona la licencia de forma gratuita a todos los centros y organizaciones del SNS, así como a instituciones y organizaciones de carácter privado, empresas y particulares[6].

El tema de la interoperabilidad también es muy importante en nefrología. Los servicios y unidades de diálisis pueden tanto utilizar los EHR del hospital y adaptarlos a las necesidades de nefrología, como utilizar programas departamentales específicamente diseñados para nefrología. Los programas departamentales en general disponen de mayores posibilidades de almacenar información de interés nefrológico de forma estructurada, habitualmente no disponibles en los programas generalistas. Esto incrementa la capacidad de exploración de la información y permite su remisión automática a un registro o base de datos. Sin embargo, obliga a hacer un esfuerzo de integración con los sistemas hospitalarios. Para esta integración, el uso de arquetipos y terminologías comunes facilitaría mucho el intercambio de información. En el futuro, se debería tender a que la información en nefrología sea lo más estructurada posible sin afectar a la usabilidad y que, además, priorice su interoperabilidad, con la incorporación de arquetipos, plantillas y uso de terminología clínica. Hay organismos internacionales que trabajan con arquetipos que se pueden utilizar o adaptar como openEHR, que puede facilitar la homogeneidad en los arquetipos, sobre todo en aspectos tan importantes como alergias, alertas o problemas de salud.

Tabla 1. Recursos de terminología CTtO

C (clasificaciones)	CIE-9 MC, CIE-10, CIE-10-ES, ATC, CIAP
T (terminologías)	SNOMED CT, UMLS (Unified Medical Language System)
t (tesauros)	Léxicos, glosarios, patrones sintácticos, etc.
O (ontologías)	ORPHA, BioPortal

Es probable que el uso de mayor información estructurada no solo permita su explotación de forma más precisa y ayude en la toma de decisiones, sino que facilite el intercambio de información entre distintas bases de datos, lo que aumenta la posibilidad de crear grandes bases de datos nefrológicas a nivel nacional e internacional.

Por último, la tendencia a la estructuración y uso de elementos de interoperabilidad de los EHR no es contradictoria con la sofisticación progresiva de explotación de la información con procesamiento del lenguaje natural, pues la presencia de terminologías clínicas facilita mucho dicha labor[5].

PROBLEMAS Y DESAFÍOS EN LA GESTIÓN DEL *BIG DATA:* RECOGIDA, AGREGACIÓN, PROCESAMIENTO Y FIABILIDAD DE LOS DATOS

De entrada, hay una falta de consenso sobre la definición de BD en salud que dificulta la comparación entre diferentes estudios. Sin embargo, el mayor desafío se encuentra en el acceso a los datos. Para conseguir obtener conocimiento de los datos existentes en los sistemas de salud se necesitaría buscar soluciones, entre otros problemas, a la búsqueda y captura de datos de fuentes heterogéneas, su procesamiento y su almacenamiento en bases de datos, su análisis, su gestión, su visualización y la capacidad de compartirlos con otros investigadores en salud[7]. No hay que olvidar los problemas de seguridad y privacidad y la obligación legal de obtener un consentimiento informado para poder utilizar los datos, aunque se hayan anonimizado.

La minería de datos o *data mining* es un proceso diseñado para obtener información relevante de las grandes bases de datos[8] y

algunos de sus objetivos son aprender, identificar y buscar conocimiento, patrones o regularidades a partir de los datos. Una vez recogidos los datos, se deben procesar para eliminar los errores de tal forma que sean fiables. Las fuentes de datos pueden contener sesgos o errores, o los datos estar incompletos. La dificultad está en cómo depurar la gran cantidad de datos, y cómo decidir qué datos son fiables y qué datos son útiles. Otro reto es sincronizar las fuentes de datos y las plataformas distribuidoras de BD (aplicaciones, suministradores, sensores, redes, etc.) con las infraestructuras internas de cada organización, con el objetivo de maximizar la fuerza de los modelos predictivos utilizados para el análisis. Es necesario codificar los datos que proceden de diferentes fuentes para asegurar su almacenamiento eficiente, el acceso fácil a múltiples consultas, su seguridad y su privacidad. Otro problema tiene relación con la capacidad del sistema informático que depende de la arquitectura del ordenador, que puede enlentecer el acceso a los datos y la ejecución y escalabilidad de las aplicaciones de BD.

Existe un problema especial con los datos que están poco representados. Las técnicas clásicas de aprendizaje no están adaptadas al análisis de muestras minoritarias, porque están basadas en medidas globales y no consideran las diferencias entre el número de muestras que dependen de cada clase. Sin embargo, las clases que están infrarrepresentadas pueden representar casos importantes para identificar. Para resolver este problema se han desarrollado diferentes métodos. Varios de ellos utilizan técnicas de clasificación binarias que puedan aplicarse en problemas de clasificación multiclase, como por ejemplo el análisis discriminante, los árboles de decisión y algoritmos "k vecinos más próximos" (*k-nearest neighbours* o k-NN), las redes bayesianas, las redes neuronales, las máquinas de vectores de apoyo, etc. Otra estrategia se conoce como DEM *(decomposition and ensemble methods),* y consiste en descomponer problemas de clasificación multiclase en un grupo de problemas que pueden ser resueltos como clasificadores binarios (BC). Esto significa que las nuevas observaciones se clasifican estableciendo una estrategia agregativa basada en las predicciones derivadas de los BC[9].

MATEMÁTICAS Y COMPUTACIÓN PARA EL *BIG DATA:* TÉCNICAS ANALÍTICAS, PLATAFORMAS E INSTRUMENTOS

La toma de decisiones en la resolución de problemas relacionados con sistemas complejos requiere utilizar modelos matemáticos y computacionales. Cuando un problema es complejo porque implica información incompleta o poco precisa puede ser difícil distinguir los distintos objetos y es más conveniente agrupar la información para abordar el problema (computación granular [GrC]). Los gránulos están formados por objetos que se agrupan porque no se pueden identificar como objetos individuales, o tienen formas o funciones similares[10]. Un ejemplo sería una agrupación de palabras que alcanza un significado que de por sí no tenían o las pinceladas de un cuadro y el resultado final del cuadro después de agregar todas las pinceladas. Los sistemas basados en GrC se pueden utilizar para construir modelos computacionales enfocados a la minería de datos, análisis de documentos, organización y recuperación de grandes bases de datos multimedia, datos médicos, y sensores remotos y biométricos. Las técnicas de GrC pueden servir como herramientas de procesamiento efectivo para sistemas inteligentes del mundo real y ambientes dinámicos como el FDS (Fuzzy Dynamic Decision Systems). La integración de GrC y la inteligencia computacional se han convertido en un área de gran interés investigador para desarrollar modelos eficientes de toma de decisiones dedicados a resolver problemas complejos de BD. La GrC se puede ejecutar asumiendo diversos tipos de imprecisión en las variables analizadas mediante conjuntos difusos *(fuzzy sets),* conjuntos en bruto *(rough sets),* conjuntos aleatorios *(random sets),* etc.[11].

Para procesar BD se suelen utilizar plataformas en código abierto de Apache basadas en Hadoop[12], que utiliza miles de servidores convencionales de bajo coste y permite almacenar datos sin limitación de volumen. Hadoop está compuesto de varios módulos: *a)* Hadoop Common (contiene bibliotecas y herramientas); *b)* Hadoop Distributed File System (HDFS), un sistema de almacenamiento de archivos; *c)* YARN, una plataforma de gestión de recursos, y *d)* MapReduce, un modelo de programación en Java. Su principal ventaja es la capacidad para procesar de forma rápida enormes grupos de datos, gracias a sus clústeres paralelos y el sistema de distribución de documentos. Hadoop no copia en la memoria todos los datos para ejecutar operaciones y ejecuta los datos desde donde están almacenados.

En la actualidad se dispone de varias herramientas tecnológicas en código abierto (tabla 2) para trabajar con Hadoop, con el objetivo de capturar (Sqoop, Flume, Chukwa), procesar (MapReduce, YARN), consultar (Pig, JAQL, Hive), almacenar (HDFS, Hbase, Hive), transmitir (Storm, Spark) y analizar datos (Mahout, R).

Los análisis retrospectivos son útiles para identificar factores de riesgo, realizar estudios de coste-eficacia, selección de pacientes o estudios epidemiológicos, pero es necesario tener en cuenta las limitaciones y el apropiado uso de los datos, así como los riesgos de inexactitudes[13]. El estándar de referencia de la medicina basada en la evidencia son los ensayos clínicos controlados, aleatorizados y multicéntricos. Sin embargo, dada la alta frecuencia con la que estos ensayos dan un resultado negativo, es recomendable el uso previo de estudios retrospectivos mediante EHR. Con este objetivo se ha constituido en Estados Unidos una red (National Patient Centered Clinical Research Network) que proporciona datos de los EHR accesibles a los investigadores[14].

Sin un análisis adecuado, los datos no son útiles. Existen varias técnicas de análisis de datos, incluyendo la minería de datos (que permite encontrar patrones y extraer valores escondidos entre los datos), la visualización, el análisis estadístico y el *machine learning*. A las técnicas clásicas de minería de datos, como *association mining, clustering* y clasificación, les falta eficiencia, escalabilidad y precisión cuando se aplican a BD en un ambiente dinámico. Debido al tamaño de los datos, la rapidez a la que se van generando y su variabilidad, no es posible

Tabla 2. Herramientas tecnológicas en código abierto

Tecnología en código abierto	Fuente	Función	Características
Apache Sqoop™	https://sqoop.apache.org	Captura de datos	Transferencia de datos bidireccional entre Hadoop y bases de datos
Apache Flume™	https://flume.apache.org	Captura de datos	Transmisor de datos en tiempo real hacia Hadoop. Captura, agrega y mueve archivos de registros. Almacena datos en HDFS y HBase
Apache Chukwa™	http://chukwa.apache.org/	Captura/análisis de datos	Basado en HDFS y MapReduce
MapReduce™	https://www.tutorialspoint.com/es/hadoop/hadoop_mapreduce.htm	Procesamiento de datos	Resolución de problemas con *datasets* de gran tamaño
Apache Hadoop YARN™	http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html	Procesamiento/análisis de datos	Mejor escalabilidad que MapReduce. Permite múltiples aplicaciones. Gestor de recursos avanzados
Apache pig™	https://pig.apache.org/	Consulta de datos	Plataforma para crear programas MapReduce
JAQL	https://developer.ibm.com/hadoop/docs/biginsights-ibm-open-platform/biginsights-value-add/jaql/jaql-overview/	Lenguaje de programación	Incorpora funciones *(in-built functions),* operadores principales *(core operators)* y adaptadores E/S
HDFS (Hadoop Distributed File System)	https://hadoop.apache.org/	Almacenamiento de datos	
Apache HBase	https://hbase.apache.org/	Almacenamiento de datos	Base de datos no relacionada de tipo NoSQL
Apache Hive	https://hive.apache.org/	Almacenamiento de datos	Permite representar datos en una base estructurada. Está principalmente basado en tablas
Apache Storm	https://storm.apache.org/	Transmisión de datos	La interface ISpout acepta cualquier tipo de dato en tiempo real. La interface iBolt se adapta a cualquier sistema de salida de datos
Apache Spark	https://spark.apache.org/	Transmisión de datos	Análisis de datos
Apache Mahout	https://mahout.apache.org/	Análisis de datos	Ofrece librerías para *clustering,* clasificación, filtros de colaboración y minería de patrones frecuentes y de textos. Herramientas adicionales
R	https://www.r-project.org/	Análisis de datos	Análisis estadístico y gráficos
Ricardo software	https://software.ricardo.com/	Análisis de datos	Permite hacer simulaciones y análisis

almacenarlos de forma permanente para analizarlos después. Por lo tanto, se necesitan nuevas herramientas como el *data stream learning* para optimizar las técnicas de análisis de datos, procesarlos en un período limitado con recursos limitados (p. ej., de memoria) y producir en tiempo real resultados precisos. La variabilidad del *stream* (flujo) da lugar a cambios impredecibles (p. ej., cambios en la distribución de las variables).

El objetivo del *machine learning* es obtener conocimiento y tomar decisiones inteligentes. Generalmente se divide en 3 subdominios: aprendizaje supervisado *(supervised learning),* aprendizaje no supervisado *(unsupervised learning)* y *reinforcement learning*[15]. El *deep learning* (que se traduciría como "aprendizaje en profundidad") es un campo de investigación muy activo dentro del *machine learning* y el reconocimiento de patrones. Juega un papel importante en aplicaciones para el análisis predictivo como la visión artificial, el reconocimiento de voz y el procesamiento del lenguaje natural. Se basa en el aprendizaje jerárquico y la extracción de diferentes niveles de abstracción de datos complejos. Es útil para simplificar el análisis de grandes volúmenes de datos, definir índices semánticos, marcar datos *(data tagging),* recuperar información y realizar tareas de discriminación como clasificación y predicción.

Sin embargo, a pesar de estas ventajas, el BD todavía presenta retos: *a)* la gran cantidad de datos. Las computaciones iterativas de los algoritmos de aprendizaje son muy difíciles de paralelizar. Se necesita crear algoritmos paralelos eficientes y escalables para mejorar las etapas de entrenamiento del *deep learning; b)* heterogeneidad. Las soluciones están restringidas por el tiempo de ejecución y la complejidad de los modelos; *c)* ruido de fondo y distribuciones no estacionarias como las derivadas de datos incompletos, la desaparición de categorías *(missing labels),* o la imprecisión inherente a la categorización de los datos *(noisy labeled datasets)*[16]; *d)* alta velocidad. Los datos se generan a una gran velocidad y se deben procesar en tiempo real. Además, los datos no suelen ser estacionarios y presentan cambios de distribución en el tiempo.

Para caracterizar datos no estacionarios (aquellos que pueden cambiar en el tiempo) que puedan utilizarse en métodos de aprendizaje basados en grandes flujos de datos y en el llamado "cambio de concepto" (cambio de distribución de variables en el tiempo), se han propuesto los algoritmos incrementales. Ejemplos de algoritmos incrementales incluyen los árboles de decisión (IDE4, ID5R), las reglas de decisión, las redes neuronales, las neuronas gaussianas, las redes RBF (learn++, ARTMAP) o el incremental SVM *(support-vector machine).* Cabe destacar los algoritmos *ensemble* o de conjunto, que son más flexibles y se adaptan mejor[17].

SEGURIDAD Y PRIVACIDAD DE LOS DATOS

En la actualidad, la privacidad es el gran problema que va en aumento con la implementación del *big data*. La generación diaria de miles de datos nos hace cada vez más vulnerables a la exposición de nuestra privacidad. La legislación protege la privacidad de personas mediante un método de notificación y consentimiento. Es necesario conocer que cuando se firma un consentimiento para que un sistema informático grabe nuestros datos estamos ofreciendo información que puede utilizarse como rastro posterior o huella digital. A modo de ejemplo, nuestras búsquedas en Google sin protección expresa de privacidad nos llevan a recibir posteriormente anuncios relacionados con estas[18].

La seguridad en el tratamiento de los datos personales y su libre circulación en el ámbito de la Unión Europea está regulada por el Reglamento General de Protección de Datos (RGPD) 2016/679 del Parlamento Europeo[19], aprobado el 27 de abril de 2016 (aplicable desde el 25 de mayo de 2018) y, en el ámbito nacional, por la Ley Orgánica de Protección de Datos y Garantía de los Derechos Digitales (LOPDGDD), aprobada el 18 de octubre de 2018.

APLICACIONES DEL *BIG DATA* A LA NEFROLOGÍA. ¿DÓNDE ESTAMOS Y QUÉ PODEMOS HACER? ¿ESTAMOS PREPARADOS?

En la actualidad se encuentra una vorágine de artículos y nuevas publicaciones en el ámbito de la medicina y concretamente en la nefrología relacionados con el uso de la inteligencia artificial como una herramienta de ayuda en el diagnóstico, evaluación del riesgo y abordaje del paciente con enfermedad renal, y que se podría definir como "nefrología inteligente". Los estudios diseñados hasta la actualidad cubren un amplio abanico de la nefrología, que incluye la nefrología clínica con predictores de progresión de ADPKD y nefropatía IgA, estudios de estimación y variabilidad del área bajo la curva de los inhibidores de la calcineurina en el trasplante renal, estudios en anemia y el peso seco óptimo en pacientes en programa de hemodiálisis y la identificación de patógenos responsables de la infección bacteriana en diálisis peritoneal[20].

Últimamente, la bibliografía ha comenzado a desarrollar modelos híbridos compuestos por diferentes tipos de técnicas de evaluación[21]. Consideremos, por ejemplo, la categorización de pacientes dependiendo de su evolución durante el proceso de trasplante. De manera intuitiva, las redes neuronales artificiales suelen considerar un objetivo definido por una variable dividida en diferentes categorías. El algoritmo correspondiente intenta aprender, y posteriormente replicar, la relación existente entre las variables independientes y la subsecuente clasificación de los pacientes entre las diversas categorías. La composición de dichas categorías no es necesariamente homogénea, pudiendo dar lugar a categorías con un número de observaciones que resulta insuficiente para inferir resultados suficientemente robustos estadísticamente. En este sentido, la definición de las diferentes categorías mediante procesos previos de optimización destinados a orientar el algoritmo durante su fase de aprendizaje constituye una parte fundamental del proceso de evaluación de los pacientes. En consecuencia, la bibliografía ha comenzado a diseñar estructuras híbridas que incluyen modelos de optimización (basados en variables con diversos tipos de imprecisión inherentes a su categorización), que tanto se utilizan para seleccionar diferentes muestras de pacientes dependiendo

de su eficiencia[22] como para clasificar la muestra e implementar una red neuronal en su posterior categorización[23].

Así, y como se ha comentado con anterioridad, en los últimos años existe un gran interés en el estudio y la difusión de la inteligencia artificial en nefrología, con la finalidad de mejorar estrategias diagnósticas, terapéuticas y pronósticas de los pacientes afectados de enfermedad renal, en los siguientes ámbitos: estudios de detección precoz del fracaso renal agudo y prevención[24-28], estudios en histología renal[29], estudios en fisiología renal[30,31], identificación de nuevas dianas terapéuticas en ERC[32-34], identificación de nuevas dianas terapéuticas en HTA[35,36], BD y diálisis[37-43], supervivencia[44] y trasplante[45].

FUTURO Y RETOS DEL *BIG DATA* PARA LA NEFROLOGÍA. PROPUESTAS DE PROYECTOS DE INVESTIGACIÓN. ¿QUÉ NECESITAMOS PARA QUE LA NEFROLOGÍA ESTÉ A LA CABEZA EN INVESTIGACIÓN EN *BIG DATA?*

La nefrología presenta una evolución histórica y es pionera a la hora de implantar los avances de la medicina, así como la tecnología que se va sucediendo. Pero en el siglo XXI, la información se plantea como prioritaria y su tratamiento complejo, así como las diferentes fuentes y su carácter heterogéneo, hacen necesarias herramientas para poder compartir dicha información y poder tener accesibilidad a una base de datos nacional con datos clínicos (públicos y privados) y datos genómicos, que sea lo más homogénea posible y se fundamente de la misma manera para que todos entendamos lo que pasa en nuestros pacientes[46]. Dichos sistemas de información no quedarían solo reducidos a los datos que manejamos en nuestra actividad clínica diaria si se pudiera conseguir que las grandes compañías compartieran sus datos. Todo ello, y de manera más automatizada y homogénea, nos llevaría a poder obtener dicha información, aunque se hacen necesarios estudios para evaluar la calidad de los datos[47].

En el campo de la nefrología, se van realizando estudios en áreas como la enfermedad renal crónica, el fracaso renal agudo y las diferentes opciones de diálisis, así como en el trasplante renal, que conllevan mucho campo para trabajar e investigar. Pero para ello, y acorde a la consecución de una información más veraz, homogénea y más accesible e instantánea, no solo contamos con la información de las historias clínicas y bases de datos hospitalarias, sino que esta se puede extraer de la digitalización de las muestras y diagnóstico histológico[48-50], de las imágenes radiológicas[51,52] y de la obtención de bases de datos poblacionales con datos y mortalidad. No hay que olvidar que es necesario diseñar algoritmos de decisión que ayuden a tomar decisiones médicas, como podría ser, en la asignación de los órganos donados[53,54], estimar la supervivencia de los receptores de injertos renales[55-58] o evitar la transmisión de enfermedades del donante[59].

La inteligencia artificial puede dar un soporte para poder tomar decisiones clínicas, que en ningún momento hasta ahora sustituiría la labor médica, pero ayudaría a tomar decisiones más precisas[60] e incluso a cambiar paradigmas como disponer de información en la consulta del impacto que tendría cada variable sobre la evolución del paciente[61], entendiéndolo como un global y no como objetivos univariables; así, por ejemplo, en el trasplante renal, en la consulta saber cómo está el paciente en el control de la presión arterial respecto a un paciente ideal, entendido como el que no se muere, no rechaza y no desarrolla un cáncer[23]. Pero para ello se necesita conseguir un grupo multidisciplinar basado en matemáticos, físicos, bioingenieros, informáticos, a nivel nacional y dedicado al análisis de los datos (algo así como un *supercomputing group* dedicado a la nefrología), que trabajen en estrecha relación con los investigadores o clínicos en nefrología. Y lo que es importante, conseguir fuentes de financiación nacionales e internacionales que permitan desarrollar todos estos estudios y tecnología clave para poder avanzar en la comprensión de las enfermedades que nos acontecen y poder ofrecer una mejor atención a nuestros pacientes cada día en la consulta.

Con la finalidad de abordar correctamente la inteligencia artificial y BD en nefrología es necesario la creación de equipos multidisciplinares, quizás llamados en un futuro *supercomputing nephrologist group,* formados por matemáticos, nefrólogos, físicos, bioingenieros, informáticos y estadísticos. En la actualidad, mucha financiación de proyectos, tanto en el marco español como en el europeo, está dedicada a la creación de BD e inteligencia artificial, pero creemos que es necesaria la financiación por parte de instituciones sanitarias de dichos proyectos dirigidos a mejorar el abordaje del paciente renal.

Conflicto de intereses

No hay conflictos de interés.

Conceptos clave

1. El conocimiento generado por la ciencia de los datos y la inteligencia artificial tendrá un impacto en la práctica clínica con ayudas al diagnóstico, la instauración de tratamientos personalizados, el diseño de fármacos, la identificación de poblaciones de riesgo o el rastreo de información en la historia clínica electrónica, emergiendo lo que podríamos llamar "nefrología inteligente".
2. La historia clínica electrónica es útil para el abordaje diario de los pacientes, para la ayuda a la toma de decisiones y para la exploración de la información con fines de gestión, epidemiológicos o de investigación.
3. La privacidad de los datos es el gran problema que va en aumento con la implementación del *big data* y la legislación protege la privacidad de personas mediante un método de notificación y consentimiento.
4. En la actualidad se encuentran numerosos artículos en nefrología relacionados con el uso de la inteligencia artificial como una herramienta de ayuda en el diagnóstico, evaluación del riesgo y abordaje del paciente con enfermedad renal.

REFERENCIAS BIBLIOGRÁFICAS

1. DZone [consultado 19-11-2019]. Disponible en: https://dzone.com/
2. Ashley EA. Towards precision medicine. Nat Rev Genet. 2016;17:507-22.
3. Feldman B, Martin EM, Skotnes T. Big Data in Healthcare Hype and Hope. Big Data; 2012.
4. Eysenbach G. What is e-health? J Med Internet Res. 2001;3:1-5.
5. Hernández Medrano I, Tello Guijarro J, Belda C, Ureña A, Salcedo I, Espinosa-Anke L, et al. Savana: Re-using Electronic Health Records with Artificial Intelligence. International Journal of Interactive Multimedia and Artificial Intelligence. 2017;4:8-12.
6. Muñoz Carrero A, Romero Gutiérrez A, Marco Cuenca G, Abad Acebedo I, Cáceres Tello J, Sánchez de Madariaga R, et al. Manual Práctico de Interoperabilidad Semántica para Entornos Sanitarios Basada en Arquetipos. Madrid: Unidad de Investigación en Telemedicina y e-Salud, Instituto de Salud Carlos III – Ministerio de Economía y Competitividad; 2013.
7. Hulsen T, Jamuar SS, Moody AR, Karnes JH, Varga O, Hedensted S, et al. From Big Data to Precision Medicine. Front Med (Lausanne). 2019;6:34.
8. Yao JT, Yao YY. A granular computing approach to machine learning. Proceedings of the 1st International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'02). 2002.
9. Zhou L, Fujita H. Posterior probability based ensemble strategy using optimizing decision directed acyclic graph for multi-class classification. Information Sciences. 2017;400-401:142-16.
10. Skowron A, Jankowski A, Dutta S. Interactive granular computing. Granular Computing. 2016;1:95-113.
11. Wang G, Yang J, Xu J. Granular computing: from granularity optimization to multi-granularity joint problem solving. Granular Computing. 2017;2:105-20.
12. Bappalige SP. An introduction to Apache Hadoop for Big Data [consultado 19-11-2019]. Disponible en: https://opensource.com/life/14/8/intro-apache-hadoop-big-data
13. Balas EA, Vernon M, Magrabi F, Gordon LT, Sexton J. Big Data Clinical Research: Validity, Ethics, and Regulation. Stud Health Technol Inform. 2015;216:448-452.
14. Terry K, Fridsma D, Haley D, Hepp K; Medical Economics Staff. The future of interoperability. Med Econ. 2014;91:30, 32-4, 36.
15. Qiu J, Wu Q, Ding G, Xu Y, Feng S. A survey of machine learning for big data processing. EURASIP Journal on Advances in Signal Processing. 2016;2016:67.
16. Li J, Wong Y, Zhao Q, Kankanhalli MS. Learning to Learn from Noisy Labeled Data. 2019 [consultado 19-11-2019]. Disponible en: https://github.com/LiJunnan1992/MLNT
17. Sayed-Mouchaweh M, editor. Learning from Data Streams in Evolving Environments: Methods and Applications. Springer; 2019.
18. Monleón Getino A. El impacto del Big-data en la Sociedad de la Información. Significado y utilidad. Historia y Comunicación Social. 2015;20:427-45.
19. Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la directiva 95/46/CE (reglamento general de protección de datos). Diario Oficial de la Unión Europea. 2016:L119/1-88.
20. Niel O, Bastard P. Artificial Intelligence in Nephrology: Core Concepts, Clinical Applications, and Perspectives. Am J Kidney Dis. 2019;74:803-10.
21. Toloo M, Zandi A, Emrouznejad A. Evaluation efficiency of large scale data set with negative data: an artificial neural networks approach. The Journal of Supercomputing. 2015;71:2397-411.
22. Ahmadvand S, Pishvaee MS. An efficient method for kidney allocation problem: a credibility-based fuzzy common weights data envelopment analysis approach. Health Care Manag Sci. 2018;21:587-603.

23. Arteaga FJS, Caprio DD, Cucchiari D, Campistol JM, Oppenheimer F, Diekmann F, et al. Modeling Patients as Dynamical Systems: Evaluating the Efficiency of Kidney Transplantation through Multi-Stage Data Envelopment Analysis [abstract]. Am J Transplant. 2019;19 Suppl 3.
24. Colpaert K, Hoste EA, Steurbaut K, Benoit D, Van Hoecke S, De Turck F, et al. Impact of real-time electronic alerting of acute kidney injury on therapeutic intervention and progression of RIFLE class. Crit Care Med. 2012;40(4):1164-70.
25. He J, Hu Y, Zhang X, Wu L, Waitman LR, Liu M. Multi-perspective predictive modeling for acute kidney injury in general hospital populations using electronic medical records. JAMIA Open. 2019;2: 115-22.
26. Sawhney S, Fraser SD. Epidemiology of AKI: Utilizing Large Databases to Determine the Burden of AKI. Adv Chronic Kidney Dis. 2017;24:194-204.
27. Bagshaw SM, Goldstein SL, Ronco C, Kellum JA; ADQI 15 Consensus Group. Acute kidney injury in the era of big data: The 15 th Consensus Conference of the Acute Dialysis Quality Initiative (ADQI). Can J Kidney Health Dis. 2016;3:5.
28. Nadkarni GN, Coca SG. Temporal trends in AKI: Insights from big data. Clin J Am Soc Nephrol. 2016;11:1-3.
29. Gallego J, Pedraza A, Lopez S, Steiner G, Gonzalez L, Laurinavicius A, et al. Glomerulus Classification and Detection Based on Convolutional Neural Networks. J Imaging. 2018;4:20.
30. Zhao Y, Yang CR, Raghuram V, Parulekar J, Knepper MA. Big: A large-scale data integration tool for renal physiology. Am J Physiol Renal Physiol. 2016;311:F787-92.
31. Huling JC, Pisitkun T, Song JH, Yu MJ, Hoffert JD, Knepper MA. Gene expression databases for kidney epithelial cells. Am J Physiol Renal Physiol. 2012;302:F401-7.
32. Cisek K, Krochmal M, Klein J, Mischak H. The application of multi-omics and systems biology to identify therapeutic targets in chronic kidney disease. Nephrol Dial Transplant. 2016;31:2003-11.
33. Morris AP, Le TH, Wu H, Akbarov A, Van der Most PJ, Hemani G, et al. Trans-ethnic kidney function association study reveals putative causal genes and effects on kidney-specific disease aetiologies. Nat Comm. 2019;10:29.
34. Singh K, Betensky RA, Wright A, Curhan GC, Bates DW, Waikar SS. A concept-wide association study of clinical notes to discover new predictors of kidney failure. Clin J Am Soc Nephrol. 2016;11:2150-8.
35. Giri A, Hellwege JN, Keaton JM, Park J, Qiu C, Warren HR, et al. Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. Nat Genet. 2019;51:51-62.
36. Mahajan A, Rodan ARR, Le THH, Gaulton KJ, Haessler J, Stilp AM, et al. Trans-ethnic Fine Mapping Highlights Kidney-Function Genes Linked to Salt Sensitivity. Am J Hum Genet. 2016;99:636-46.
37. Erickson KF, Qureshi S, Winkelmayer WC. The Role of Big Data in the Development and Evaluation of US Dialysis Care. Am J Kidney Dis. 2018;72:560-8.
38. Vito D, Casagrande G, Bianchi C, Costantino ML. How to extract clinically useful information from large amount of dialysis related stored data. Conf Proc IEEE Eng Med Biol Soc. 2015;2015:6812-5.
39. Barbieri C, Molina M, Ponce P, Tothova M, Cattinelli I, Ion Titapiccolo J, et al. An international observational study suggests that artificial intelligence for clinical decision support optimizes anemia management in hemodialysis patients. Kidney Int. 2016;90:422-9.
40. Barbieri C, Bolzoni E, Mari F, Cattinelli I, Bellocchio F, Martin JD, et al. Performance of a predictive model for long-Term hemoglobin response to darbepoetin and iron administration in a large cohort of hemodialysis patients. PLoS One. 2016;11:e0148938.
41. Canaud B, Barbieri C, Marcelli D, Bellocchio F, Bowry S, Mari F, et al. Optimal convection volume for improving patient outcomes in an international incident dialysis cohort treated with online hemodiafiltration. Kidney Int. 2015;88:1108-16.
42. Barbieri C, Mari F, Stopper A, Gatti E, Escandell-Montero P, Martínez-Martínez JM, et al. A new machine learning approach for predicting the response to anemia treatment in a large cohort of End Stage Renal Disease patients undergoing dialysis. Comput Biol Med. 2015;61:56-61.
43. Ion Titapiccolo J, Ferrario M, Barbieri C, Marcelli D, Mari F, Gatti E, et al. Predictive modeling of cardiovascular complications in incident hemodialysis patients. Conf Proc IEEE Eng Med Biol Soc. 2012;2012:3943-6.
44. Chen B, Fan VY, Chou YJ, Kuo CC. Costs of care at the end of life among elderly patients with chronic kidney disease: patterns and predictors in a nationwide cohort study. BMC Nephrology. 2017; 18:1-14.
45. Massie AB, Kuricka LM, Segev DL. Big data in organ transplantation: Registries and administrative claims. Am J Transplant. 2014;14: 1723-30.
46. Saran R, Steffick D, Bragg-Gresham J. The China Kidney Disease Network (CK-NET): "Big Data-Big Dreams". Am J Kidney Dis. 2017; 69:713-6.
47. Sirgo G, Esteban F, Gómez J, Moreno G, Rodríguez A, Blanch L, et al. Validation of the ICU-DaMa tool for automatically extracting variables for minimum dataset and quality indicators: The importance of data quality assessment. Int J Med Inform. 2018;112: 166-72.
48. Barisoni L, Hodgin JB. Digital pathology in nephrology clinical trials, research, and pathology practice. Curr Opin Nephrol Hypertens. 2017;26:450-9.
49. Hermsen M, De Bel T, Den Boer M, Steenbergen EJ, Kers J, Florquin S, et al. Deep Learning-Based Histopathologic Assessment of Kidney Tissue. J Am Soc Nephrol. 2019;30:1968-79.
50. Denic A, Morales MC, Park WD, Smith BH, Kremers WK, Alexander MP, et al. Using computer-assisted morphometrics of 5-year biopsies to identify biomarkers of late renal allograft loss. Am J Transplant. 2019;19:2846-54.
51. Regge D, Mazzetti S, Giannini V, Bracco C, Stasi M. Big data in oncologic imaging. Radiol Med. 2017;122:458-63.
52. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med. 2019;25:1301-9.
53. Briceño J, Cruz-Ramírez M, Prieto M, Navasa M, Ortiz de Urbina J, Orti R, et al. Use of artificial intelligence as an innovative donor-recipient matching model for liver transplantation: results from a multicenter Spanish study. J Hepatol. 2014;61:1020-8.
54. Dorado-Moreno M, Pérez-Ortiz M, Gutiérrez PA, Ciria R, Briceño J, Hervás-Martínez C. Dynamically weighted evolutionary ordinal neural network for solving an imbalanced liver transplantation problem. Artif Intell Med. 2017;77:1-11.
55. Mark E, Goldsman D, Gurbaxani B, Keskinocak P, Sokol J. Using machine learning and an ensemble of methods to predict kidney transplant survival. PLoS One. 2019;14:0209068.

56. Sapir-Pichhadze R, Kaplan B. Seeing the Forest for the Trees. Transplantation. 2019: doi:10.1097/TP.0000000000002923. [Epub ahead of print].
57. Loupy A, Toquet C, Rouvier P, Beuscart T, Bories MC, Varnous S, et al. Late Failing Heart Allografts: Pathology of Cardiac Allograft Vasculopathy and Association With Antibody-Mediated Rejection. Am J Transplant. 2016;16:111-20.
58. Aubert O, Higgins S, Bouatou Y, Yoo D, Raynaud M, Viglietti D, et al. Archetype analysis identifies distinct profiles in renal transplant recipients with transplant glomerulopathy associated with allograft survival. J Am Soc Nephrol. 2019;30:625-39.
59. Mark E, Goldsman D, Keskinocak P, Sokol J. Using machine learning to estimate survival curves for patients receiving an increased risk for disease transmission heart, liver, or lung versus waiting for a standard organ. Transpl Infect Dis. 2019;21:e13181.
60. Garcia-Vidal C, Sanjuan G, Puerta-Alcalde P, Moreno-García E, Soriano A. Artificial intelligence to support clinical decision-making processes. EBioMedicine. 2019;46:27-9.
61. Fournier M-C, Foucher Y, Blanche P, Legendre C, Girerd S, Ladrière M, et al; DIVAT Consortium. Dynamic predictions of long-term kidney graft failure: an information tool promoting patient-centred care. Nephrol Dials Transplant. 2019;34:1961-9.