Transferencia

FUNDACIÓN BOTÍN

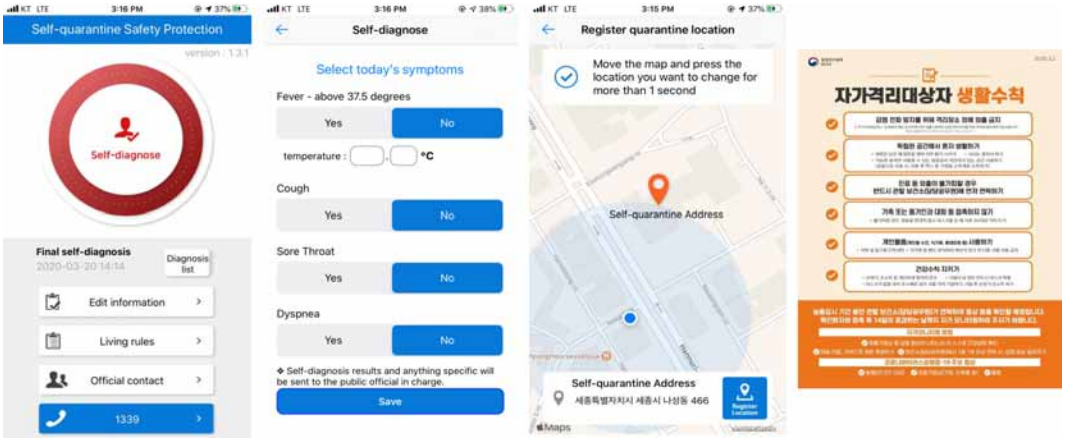
NÚMERO 19 | NOVIEMBRE DE 2020

LA MEDICINA EN LA ERA DEL BIG DATA

La medicina en la era del *big data*

A medida que los datos y la inteligencia artificial se integran en nuestras vidas, cada vez se hace más evidente su potencial para mejorar la salud de las personas, la práctica médica y el desarrollo de nuevos tratamientos. Ya lo han advertido tanto el mundo académico como empresas farmacéuticas y tecnológicas: el *big data* será fundamental para la medicina del futuro.

Estamos en la era de los datos, y la pandemia actual lo ha puesto de manifiesto con más contundencia que nunca. El país que ha mantenido a raya al coronavirus de forma más eficaz no ha sido el más grande ni el más rico, sino el que mejor ha aprovechado los datos y las herramientas que existen para analizarlos: Corea del Sur.



Aplicación que permite monitorizar a las personas en aislamiento por covid-19 a través de *big data* en Corea del Sur. Fuente, Flattening the curve on COVID-19: How Korea responded to a pandemic using ICT

Ante la rápida expansión de la enfermedad en enero y febrero, la administración surcoreana puso en marcha un conjunto de medidas extraordinarias que revolucionaron su sistema sanitario[1]. Lo hizo basándose en dos principios. El primero, realizar test de diagnóstico de forma masiva a la población. El segundo, rastrear los contactos de los casos confirmados mediante herramientas de *big data*, la principal diferencia respecto al resto de países. El 1 de marzo, cuando la pandemia no había hecho más que empezar a nivel mundial, Corea del Sur ya había aplanado la curva, sin que su población se tuviera que confinar de forma generalizada[2]. Este ejemplo

[1] Fuente, Flattening the curve on COVID 19: How Korea responded to a pandemic using ICT, Gobierno de la República de Corea (2020): undp.org/content/seoul_policy_center/en/home/presscenter/articles/2019/flattening-the-curve-on-covid-19.html

[2] Fuente, Business Insider: businessinsider.com/how-south-korea-controlled-its-coronavirus-outbreak-2020-4?IR=T

no se puede extrapolar a países como España o el resto de la Unión Europea, donde las leyes que protegen la privacidad impiden que las autoridades accedan a los datos de sus ciudadanos como lo ha hecho el gobierno coreano, pero sirve para ilustrar el enorme potencial del *big data* cuando se aplica a la salud.

David Pisano, Big Data and Health Professor en el IE Business School. Fuente, David Pisano

La era de la información

El *big data* (en español podría traducirse como macrodatos) no es un fenómeno nuevo ni se restringe al ámbito de la epidemiología. "El *big data* existe desde el momento en que empezamos a digitalizar la información y a almacenar y acumular datos", declara David G. Pisano, Big Data and Health Professor en el IE Business School. "Aunque el concepto se acuñó en los noventa, realmente ha sido en la última década cuando se ha hecho popular".

Al principio, el tratamiento de los datos se hacía de forma manual. "Los primeros genomas se distribuían en libros", recuerda Alfonso Valencia, profesor ICREA y director del Departamento de Ciencias de la Vida del Barcelona Supercomputing Center (BCS-CNS). Pero, a medida que esos datos se fueron haciendo más y más complejos, nacieron por necesidad nuevas herramientas para analizarlos. "La humanidad acumula información muy rápido. El *big data* es una consecuencia inevitable de internet y de todos los instrumentos que hay conectados. Se ha convertido en un paradigma ubicuo", subraya Valencia.

Alfonso Valencia, director del Departamento de Ciencias de la Vida del Barcelona Supercomputing Center (BCS-CNS). Fuente, BSC

"El *big data* es una consecuencia inevitable de internet y de todos los instrumentos que hay conectados. Se ha convertido en un paradigma ubicuo", afirma Alfonso Valencia

El *big data*, pues, engloba las tecnologías que permiten almacenar, gestionar y procesar grandes volúmenes de información para revelar tendencias que a priori no son evidentes. Ha ido calando gradualmente en todas las áreas de nuestras vidas que implican la generación de datos, y la salud no es una excepción. Desde las historias clínicas de los pacientes a las llamadas tecnologías ómicas[3], la práctica médica es un generador de información extremadamente valiosa, que puede servir para investigar nuevos fármacos, aplicar medicina de precisión, ayudar a la toma de decisiones de médicos y enfermeros y optimizar el propio sistema sanitario.

"El volumen masivo de esta información, su complejidad y la necesidad de acceder rápido a ella hacen necesario el uso de *big data* en la biomedicina", afirma Pisano. "Los sistemas computacionales por sí solos no van a curar las enfermedades, pero sin estos datos y sistemas de análisis no se van a poder curar de forma racional", subraya Valencia. Un ejemplo es la búsqueda de nuevos tra-

[3] Las tecnologías ómicas estudian macroconjuntos de datos biológicos a nivel molecular. Son, por ejemplo, la genómica (que estudia datos de genomas), la proteómica (datos de proteínas) o la metabolómica (datos sobre el metabolismo).

"El *big data* engloba las tecnologías que permiten almacenar, gestionar y procesar grandes volúmenes de información para revelar tendencias que a priori no son evidentes"

tamientos para la covid-19. Ahora, ante la emergencia global, grupos de investigación de todo el mundo están valiéndose de datos sobre la estructura molecular del virus para hallar pistas sobre qué fármacos pueden ser más efectivos. "Manualmente llevaría mucho tiempo. En cambio, creando modelos computacionales podemos ver qué compuestos funcionarían mejor sobre las proteínas del virus, lo que sirve para guiar la búsqueda en librerías de fármacos", explica Alfonso Valencia. El objetivo es acelerar el proceso tanto como sea posible.

También las empresas del sector de la salud apuestan por las aplicaciones del *big data*. "La industria farmacéutica está invirtiendo mucho en estas tecnologías para desarrollar fármacos cada vez más rápido, y que sean más seguros, eficaces y adecuados a cada paciente. Creo que esto va a ser un salto de gigante en la investigación biomédica", valora David Pisano. Esta apuesta puede comenzarse en el ámbito nacional: en Madrid, la farmacéutica Roche tiene uno de sus mayores centros de tecnologías de la información, que recientemente se ha reconvertido para desarrollar proyectos de *big data*, explica Pisano.

"A nivel de compañías, empieza a haber un movimiento interesante y consistente en *big data*", coincide Valencia. Aunque en España no hay grandes farmacéuticas que podrían hacer avanzar al campo a mayor velocidad, "sí tenemos muchas compañías pequeñas, que son una gran riqueza. Tenemos uno de los hubs más grandes de Europa en el ámbito de la biomedicina y la informática", destaca. "Siempre hemos sido un país puntero en Europa pese a la escasa financiación pública en investigación, pero corremos el peligro de quedarnos atrás a medida que otros países hacen grandes inversiones y van cobrando ventaja. Conforme los ensayos clínicos se complican y van incluyendo más información genómica, los países que tengan acceso a las tecnologías necesarias para gestionar los datos serán los que tengan mayor capacidad de atraer ensayos clínicos. Y eso es fundamental para la ciencia, el desarrollo de la industria y el beneficio de los pacientes. A pesar de los recortes, tradicionalmente esto en España ha funcionado muy bien, pero si en el futuro nos quedamos atrás y perdemos la capacidad de atraer proyectos y compañías, perderemos la oportunidad de traslación de la investigación a la práctica clínica", remarca el investigador del BSC.

Cómo extraer valor de la avalancha de datos

Hasta ahora, las tecnologías ómicas, y especialmente la genómica, han sido el principal caballo de tiro del *big data* en biomedicina. "Probablemente la genómica es la actividad humana en la que los datos se duplican más rápidamente", señala Alfonso Valencia. Sin embargo, este crecimiento tan frenético entraña un inconveniente.

La información crece a un ritmo tan elevado que las bases de datos convencionales no alcanzan a recogerla. "No tienen tiempo de incorporar todo lo que se publica", explica Ferran Sanz, director del Programa de Investigación en Informática Biomédica (GRIB), del Instituto Hospital del Mar de Investigaciones Médicas (IMIM). Ante este problema, Sanz y su equipo crearon hace diez años una base

El grupo Integrative Biomedical Informatics del GRIB, liderado por Laura I. Furlong y Ferran Sanz, trabaja en los proyectos de *big data* DisGeNET y TRANSAFE. Fuente, GRIB

de datos innovadora, llamada DisGeNET[4], que recoge relaciones entre genes y enfermedades, y que abarca desde trastornos con base genética hasta los de etiología compleja, por ejemplo, la depresión o la hipertensión.

"La industria farmacéutica está invirtiendo mucho en estas tecnologías para desarrollar fármacos cada vez más rápido, y que sean más seguros, eficaces y adecuados a cada paciente", explica David Pisano

"La idea vino cuando nos dimos cuenta de que más del 60% de la información sobre asociaciones entre genes y enfermedades no figuraba en ninguna base de datos. Estas asociaciones son muy importantes para la medicina de precisión. Es imprescindible saber qué genes y qué mutaciones concretas de esos genes están relacionados con cada enfermedad", argumenta Sanz. DisGeNET se nutre de información que se extrae automáticamente de los artículos científicos mediante inteligencia artificial y minería de textos[5]. Actualmente contiene más de un millón de asociaciones entre más de 20.000 genes y 30.000 enfermedades y cada mes recibe más de 5.000 consultas de investigadores de todo el mundo.

Este año, el grupo de Sanz ha creado una empresa spin-off llamada MedBioinformatics Solutions para que las compañías farmacéuticas y biotecnológicas puedan beneficiarse también de los datos de DisGeNET con la finalidad de desarrollar nuevos tratamientos, ya que las condiciones iniciales de la base de datos no permitían que se utilizara de forma lucrativa. En la compañía participan el IMIM, la Universidad Pompeu Fabra y las empresas Prous Institute for Biomedical Research y Icrowd+D[6].

[4] DisGeNET es de acceso abierto para fines no lucrativos y se puede consultar libremente en el siguiente enlace: disgenet.org

[5] La minería de textos es el análisis de textos con el fin de capturar conceptos clave y temas, así como descubrir relaciones y tendencias ocultas, sin necesidad de conocer las palabras precisas que los autores utilizan para expresar dichos conceptos. Fuente, IBM: ibm.co/2ZcPf3e

[6] Fuente, Instituto Hospital del Mar de Investigaciones Médicas: imim.cat/noticias/733/la-spin-off-medbioinformatics-solutions-ofrecera-software-y-consultoria-sobre-relaciones-entre-genes-y-enfermedades

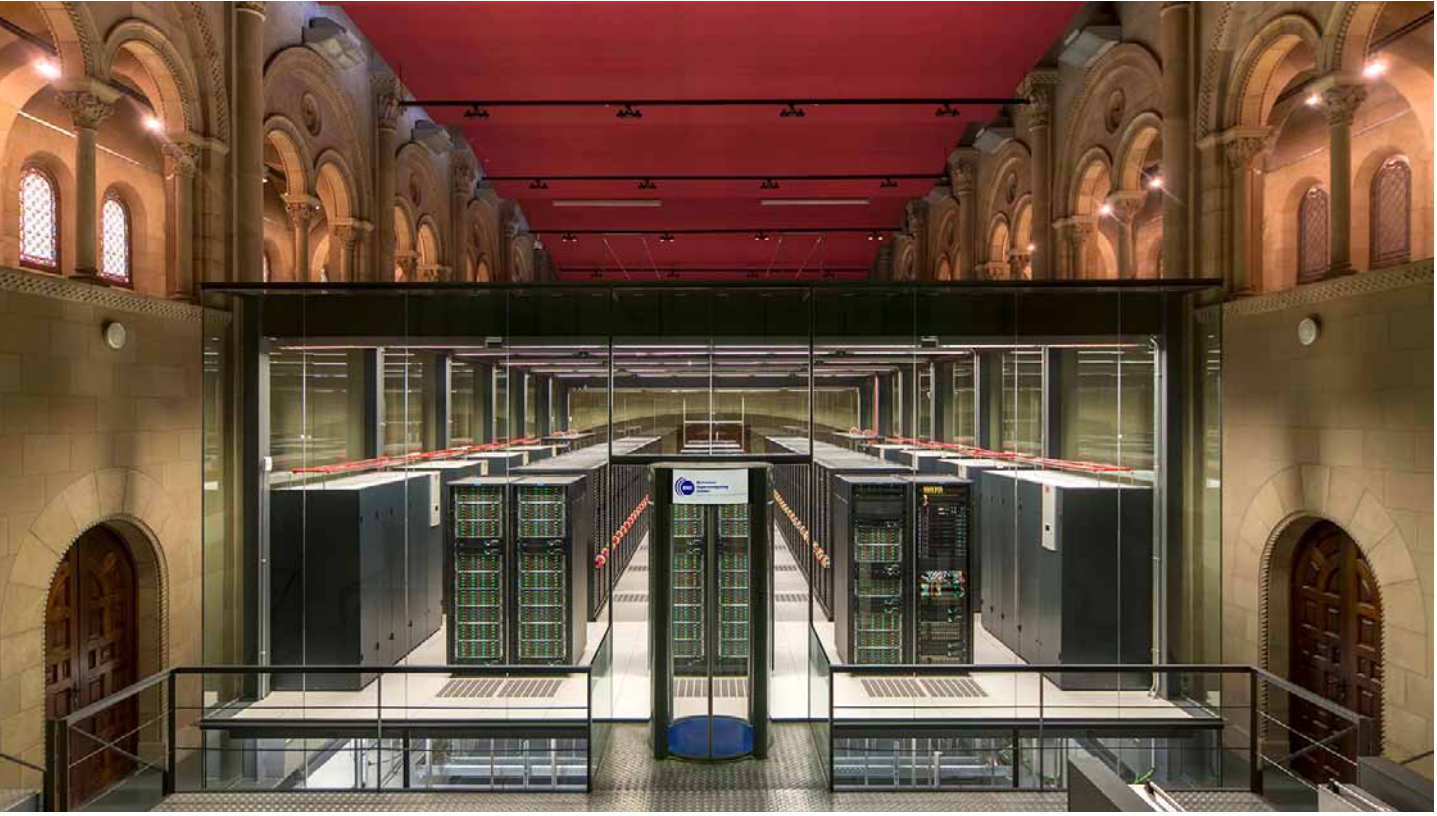
Instalaciones del Parc de Recerca Biomèdica de Barcelona (PRBB), donde se encuentra el GRIB. Fuente, UPF

Por otra parte, el IMIM coordina otro proyecto, llamado eTRANSAFE, para crear un sistema que acelere el desarrollo de nuevos medicamentos. Cada vez que se investiga un nuevo fármaco, los protocolos exigen que se sigan una serie de pasos para garantizar que sea seguro y se pueda ensayar en personas, un proceso que dura años. "Hay que hacer muchos experimentos, como si no empezáramos de cero en cada medicamento que se quiere desarrollar. Pero, si pudiéramos utilizar información histórica de medicamentos similares ya estudiados, si bien no podremos eliminar este proceso de investigación, por lo menos podríamos ahorrarnos una parte, al complementar los experimentos del nuevo medicamento con predicciones a partir de la experiencia anterior", explica Ferran Sanz, que es también el coordinador académico de eTRANSAFE. Se trata de un proyecto europeo de 40 millones de euros que comenzó en 2017 y que cuenta con la participación de ocho instituciones, seis PYMEs y doce grandes empresas farmacéuticas. Así, eTRANSAFE pretende utilizar herramientas de *big data* para destilar la información sobre seguridad y toxicidad de fármacos investigados en el pasado, con el fin de generar modelos predictivos que aceleren la investigación futura. Los resultados se esperan para 2022[7].

Otro tipo de datos que crecen a ritmo vertiginoso son los generados a través de técnicas de imagen médica, como la resonancia magnética o la tomografía computarizada. En este caso, las herramientas de procesamiento de imagen de inteligencia artificial se están convirtiendo ya en un instrumento para ayudar a los médicos a tomar decisiones de forma más rápida, eficiente y fiable. Un ejemplo es la compañía de origen español QMENTA, que aplica este tipo de herramientas para mejorar el tratamiento y la investigación de enfermedades neurodegenerativas, como el alzhéimer o la esclerosis múltiple[8].

[7] Fuente, eTRANSAFE: etransafe.eu

[8] Fuente, QMENTA: qmenta.com. Ver más información en la sección La transferencia es posible del presente boletín.



Instalaciones del superordenador MareNostrum, del Barcelona Supercomputing Center. Fuente, BSC

"La cantidad de información de salud almacenada en lenguaje médico natural es grandísima. Con herramientas de *big data* se puede desenterrar esta información para descubrir nuevas variables que expliquen qué tratamientos funcionan y cuáles no y a qué dosis, así como nuevos síntomas, dianas terapéuticas…", ilustra David Pisano.

Asimismo, las propias historias clínicas de los pacientes constituyen una enorme fuente de datos. "La cantidad de información de salud almacenada en lenguaje médico natural es grandísima. Con herramientas de *big data* se puede desenterrar esta información para descubrir nuevas variables que expliquen qué tratamientos funcionan y cuáles no y a qué dosis, así como nuevos síntomas, dianas terapéuticas…", ilustra David Pisano.

En este sentido, el BSC cuenta con un grupo de investigación en minería de textos que está desarrollando sistemas para reconocer menciones de fármacos y asociaciones con enfermedades en todo tipo de textos médicos. "Inicialmente los desarrollamos para textos en inglés relacionados con ensayos preclínicos de compuestos químicos de empresas farmacéuticas", explica Martin Krallinger, investigador principal del grupo de Minería de Textos del BSC. Después de aplicarlos también a la literatura científica y a la información sobre patentes, "ahora hemos dado el salto para que funcionen en textos en español de casos e informes clínicos, o sea en las historias clínicas de pacientes". Según relata Krallinger, los textos clínicos utilizan un lenguaje muy complejo, con vocabularios muy especializados y un alto grado de ambigüedad. "Los sistemas que procesan lenguaje general no funcionan bien en este dominio, tienen que aprender las características de cómo escriben los médicos y científicos".

El proyecto comenzó hace cinco años, como parte del Plan de Impulso de las Tecnologías del Lenguaje de la Secretaría de Estado de Digitalización e Inteligencia Digital, y se espera

que en otros dos se pueda aplicar a otros idiomas y textos, como los de las redes sociales, "algo clave para poder reconocer mejor efectos adversos o interacciones y medicamentos", añade este investigador del BSC. "Ya tenemos un sistema que no sólo reconoce fármacos y medicamentos, sino también otros conceptos de alto impacto clínico, tales como enfermedades, síntomas, procedimientos médicos o incluso patógenos. En colaboración con hospitales como el 12 de Octubre de Madrid o el Clínic de Barcelona, estamos intentando ver si este sistema es útil para caracterizar mejor y de forma sistemática comorbilidades, síntomas o coinfecciones de pacientes con enfermedades como la covid-19, procesando miles de informes clínicos. Esto implica que se podrán responder preguntas tales como qué síntomas presentaban los pacientes que han evolucionado peor o qué medicamentos estaban tomando los que han tenido una mejor evolución. En definitiva, se podrán encontrar respuestas que previamente eran imposibles de hallar debido a que la información estaba oculta en los textos clínicos, en parte por su complejidad y gran volumen", concluye Martin Krallinger.

El mismo equipo del BSC cuenta con otro proyecto que aplica un principio similar, pero al estudio exhaustivo de una enfermedad, el ictus. Se trata de un trastorno cardiovascular, que se produce cuando un coágulo obstruye la circulación sanguínea del cerebro. Es la segunda causa de muerte más frecuente a nivel mundial, responsable del 11% de los fallecimientos en el planeta, y una importante causa de discapacidad[9]. En 2018, un consorcio europeo formado por diez centros, hospitales e instituciones de España, Francia y Portugal, en el que participa el BSC, puso en marcha el proyecto ICTUSnet para mejorar el tratamiento del ictus a través del *big data*[10]. "El objetivo es establecer una red de colaboración entre diferentes regiones del sur de Europa para la creación de infraestructuras de investigación que incorporen tecnologías innovadoras de análisis de datos para mejorar los sistemas de atención integrada del ictus y reducir el impacto de la enfermedad en la población", declara Marta Villegas, colíder del grupo de Minería de Textos del BSC.

Así, ICTUSnet pretende crear un registro centralizado sobre indicadores que permitan evaluar cómo se trata el ictus en los hospitales y compartir buenas prácticas clínicas. Se nutre principalmente del informe de alta de los pacientes, que contiene datos sobre su evolución, el tratamiento y las técnicas de diagnóstico. Ya que se trata de información compleja, para agilizar el proceso la codificación en el registro se deberá realizar mediante técnicas de minería de textos. En estos momentos, los investigadores están anotando manualmente datos que servirán para entrenar a los modelos de inteligencia artificial necesarios, explica Villegas.

El *big data*, pilar de la medicina del futuro

Una de las aplicaciones del *big data* que más interés despiertan de cara a los próximos años es la optimización del sistema sanitario, como está ocurriendo ya en sectores como la logística o el entretenimiento. "¿Se puede hacer uso de esta tecnología para hacer mejores recomendaciones de salud? ¿Aplicar mejores tratamientos? ¿Mejores diagnósticos? ¿Sería posible evitar errores médicos? ¿Se podría optimizar el sistema y reducir la presión asistencial y económica?", refle-

[9] Fuente, Global Burden of Disease: Data visualizations, The Lancet (2017): thelancet.com/lancet/visualisations/gbd-compare

[10] Fuente, ICTUSnet: ictusnet-sudoe.eu/en/project/objectives/

Placas del MareNostrum, del Barcelona Supercomputing Center. Fuente, BSC

xiona David Pisano. Esta optimización deberá pasar, no obstante, por la digitalización de un sector aún anclado en procesos muy manuales. “Los profesionales del mundo de la salud necesitan incorporar a su formación competencias digitales, y el mayor reto será incorporar a los equipos médicos perfiles tecnológicos que puedan entender la terminología médica. Esto sólo funcionará desde la multidisciplinariedad y la suma de capacidades”, subraya Pisano.

“Debemos cuidar la privacidad, pero no se puede renunciar al potencial de servicio a la sociedad que tiene el *big data*”, sostiene Ferran Sanz

Otro reto por resolver es el equilibrio entre privacidad y acceso a los datos, algo que precisamente ha marcado una enorme diferencia entre países en la gestión de la pandemia. “Hay que hacer compatible la potencia del *big data* para descubrir cosas que puedan ser útiles para la humanidad con el debido respeto a la privacidad de las personas. Las historias clínicas siempre deben analizarse respetando cuidadosamente la anonimidad de los pacientes, y deben custodiarse siempre dentro del hospital. Debemos cuidar la privacidad, pero no se puede renunciar al potencial de servicio a la sociedad que tiene el *big data*”, sostiene Ferran Sanz. “Hay una regulación muy protectora con los datos, pero muchas veces no permite avanzar lo suficiente”, valora David Pisano, que añade que ante la crisis del coronavirus el sector tecnológico español ha intentado poner en marcha iniciativas para ayudar, pero muchas de ellas se han visto coartadas por las restricciones existentes sobre la privacidad.

Más a largo plazo, y a medida que la complejidad de los datos continúe aumentando, integrarlos va a ser cada vez más difícil, lo que hará necesario desarrollar nuevos métodos computacionales, señala Alfonso Valencia. Asimismo, “será necesario hacer explicables los resultados, porque la práctica biológica y médica va

"El desafío de los próximos diez años será disponer de avatares o gemelos virtuales, modelos que simulen en tiempo real lo que pasa en un paciente que interacciona con su entorno", declara Alfonso Valencia

a exigirlo. No basta con encontrar que un fármaco funciona, hay explicar por qué y hacerlo interpretable", sostiene.

Si se superan estos escollos, la próxima década podría suponer un cambio radical en cómo se abordan la medicina y la investigación biomédica. "El desafío de los próximos diez años será disponer de avatares o gemelos virtuales, es decir, modelos que simulen en tiempo real lo que pasa en un sistema, en este caso, en un paciente que interacciona con su entorno. Así, el médico podrá consultar el modelo de su paciente en su ordenador, que simulará las características de su contraparte real, y cuando éste llegue a la consulta lo podrá actualizar con nueva información. El objetivo es que sea suficientemente fidedigno para hacer simulaciones que se utilicen como parte fundamental del sistema de ayuda a la toma de decisiones en medicina. Será el culmen de la tecnología del *big data* y la inteligencia artificial", concluye Valencia.