

La Inteligencia Artificial como Herramienta para la Gestión y Explotación de Datos, Informaciones y Conocimientos Biomédicos en Entornos “Big Data” en la Nube



Diego Fernández Edreira^a, José Liñares Blanco^a, Brais Castiñeiras Galdo^b, Víctor Maojo García^c y Alejandro Pazos Sierra^d

^[a] Grupo RNASA-IMEDIR de la Universidade da Coruña

^[b] Grupo ATIS-INIBIC (Instituto de Investigación Biomédica de A Coruña)

^[c] Grupo GIB de la Universidad Politécnica de Madrid

^[d] Grupo RNASA-IMEDIR de la Universidade da Coruña. INIBIC (Instituto de Investigación Biomédica de A Coruña). CITIC (Centro de Investigación en Tecnologías de la Información y las Comunicaciones) de la UDC



RESUMEN

Nos encontramos en la era de la Medicina Personalizada y la Medicina de Precisión que tienen como objetivo último mejorar la calidad asistencial e incrementar la sostenibilidad de los sistemas públicos de salud.

El Big Data (BD) facilita el acceso, más o menos abierto, a una enorme cantidad de informaciones y datos de todo tipo, destacando los “ómicos”, residentes en múltiples bases de datos y repositorios alimentados por cohortes que ascienden a los cien-

tos de miles de pacientes. Además, es necesario tener en cuenta la aportación proveniente de la inmensidad de dispositivos de Internet de las cosas (IoT), generando datos de forma continua.

En este contexto que todo indica será expansivo, las técnicas y procedimientos de análisis basadas en la Inteligencia Artificial (IA), por su capacidad de extraer nuevos conocimientos a partir de datos e informaciones, ofrecen una posibilidad, tan necesaria como real, como herramientas de explotación y análisis. Todo parece indicar que la IA, por su potencial predictivo, clasificador, etc. en entornos del mundo real, tendrá un papel principal en todos los pasos del proceso asistencial: prevención, diagnóstico, tratamiento, control y seguimiento, así como en la implantación generalizada de la “cultura de salud”. Estamos ante un cambio de paradigma en la atención a la salud con tres características principales: primero, promoción del posicionamiento del paciente en el centro del proceso asistencial, e incluso de la I+D+i en el ámbito de la salud; segundo, cambio de dimensión del estudio y abordaje de los procesos patológicos, desde un nivel sistémico a una realidad molecular, de una escala “macro” a una “nanométrica”; y, tercero, crecimiento exponencial en los últimos años de los avances científicos y tecnológicos, siendo de especial repercusión aquellos que generan una transferencia lo más directa posible desde lo que se denomina como Ciencia básica hacia el ámbito clínico-asistencial.

En este artículo, se ofrecerá una panorámica, no exhaustiva, de la situación actual en la I+D+i de la Inteligencia Artificial en el campo Biomedicina, los pros y los contras, así como algunos de los “hitos” ya alcanzados en este ámbito.

CONTEXTO Y ANTECEDENTES


Fue en el año 1977 cuando Fredrick Sanger [1] consiguió secuenciar el primer ácido nucleico completo, el del bacteriófago Phi-X174. La posibilidad de secuenciación de genomas, a partir de dicho momento, ha ocasionado incontables mejoras en el campo de la Biomedicina y la asistencia para incrementar el estado de salud de las personas. El sucesor de este acontecimiento fue la ejecución del Proyecto Genoma Humano (en adelante, PGH) [2]. Este proyecto ha sido uno de



“Todo parece indicar que la IA, por su potencial predictivo, clasificador, etc. en entornos del mundo real, tendrá un papel principal en todos los pasos del proceso asistencial: prevención, diagnóstico, tratamiento, control y seguimiento, así como en la implantación generalizada de la “cultura de salud”

los mayores hitos en la historia de la biomedicina. A diferencia del bacteriófago secuenciado por Sanger, este proyecto contó con un consorcio formado por cientos de científicos procedentes de EE.UU., Reino Unido, Japón, Francia, Alemania y China entre otros países. El objetivo, como es bien sabido, fue obtener la primera secuencia completa del genoma de un individuo humano. En este sentido, se invoca como ejemplo la irrupción de la secuenciación completa del genoma humano, por considerarlo como una hoja de ruta con la cual poder abordarse el conocimiento y la explicación de todo tipo de enfermedades. Con el transcurso del tiempo, y la creciente ambición de los objetivos propuestos, la Ciencia en general, y la Biomedicina en particular, se está demostrando que el éxito no está siendo el esperado y depende, en gran medida, de la colaboración e intercambio de datos, informaciones y conocimientos entre los distintos grupos de investigación y la permeabilidad y colaboración entre los científicos, el personal asistencial de los sistemas de salud y los propios pacientes.

Si se compara la situación actual con la del PGH, existen dos factores que han ido evolucionando



“Un aspecto clave en los últimos años ha sido la multidisciplinariedad de los equipos de investigación. Aunar esfuerzos desde diferentes campos de investigación ha producido grandes avances en el cuidado de la salud de las personas. Es muy ilustrativo el ejemplo de la Inteligencia Artificial en el ámbito Biomédico

en los diferentes proyectos biomédicos. En primer lugar, el tipo de secuenciación que se realiza cada vez es más avanzada y más barata. Lo que favorece al segundo factor, el tamaño muestral de los estudios. Estos factores influyen principalmente en los resultados que se extraen de los proyectos, cada vez más generalizables y más robustos. En el PGH, el objetivo fue secuenciar el genoma de un individuo. Más tarde, otros proyectos internacionales han ido secuenciando genomas de cientos de individuos, posteriormente de miles, y actualmente, los objetivos que se están marcando apuntan a la secuenciación del rango de cientos de miles de individuos. Además, es importante destacar que no solo el número de pacientes aumenta, sino también la variabilidad de los datos. Actualmente, no sólo se dispone de datos de la secuencia genómica de los individuos, sino también de las diferentes “ómicas” (proteómica, lipidómica, metabolómica, epigenómica, exposómica, transcriptómica, ...) y de otras fuentes como pueden ser la propia historia clínica, con sus múltiples pruebas complementarias incluidas las imágenes de todo tipo, o incluso de dispositivos IoT que proporcionen datos de la calidad del aire, de determinados tipos de radiaciones, etc.

La generación de todos estos datos, informaciones y conocimientos (DIC) ofrecen la posibilidad del estudio de las diferentes condiciones clínicas desde otro punto de vista. Entendiendo por **dato** a cualquier valor que pueda adoptar una variable (por ejemplo, temperatura corporal 41°C); por **Información** a cualquier dato que adquiere un significado (por ejemplo, temperatura corporal 41°C significa fiebre); y reservándonos la etiqueta semántica de **Conocimiento** para cuando una información adquiere una característica de utilidad (por ejemplo, temperatura corporal 41°C significa fiebre y requiere que se tomen una serie de medidas). Hoy en día, con las posibilidades que ofrece internet, existe un exceso de datos, e incluso de informaciones, disponibles que pueden llevar a mermar el conocimiento de quién los maneja, si este no posee una adecuada formación que ayude a convertir los datos e informaciones en nuevo conocimiento dando utilidad a los mismos.

De esta forma, se ha pasado de estudiar los problemas desde un nivel sistémico a un nivel orgánico, posteriormente a un nivel tisular, pasando a un nivel celular y, últimamente, bajando a un nivel molecular, e incluso submolecular, no siendo descabellado aventurar que pronto se identifica-

rá la causa de muchas patologías a nivel atómico o subatómico. Estos cambios de enfoque han ido de la mano de dos factores principales: el desarrollo y utilización de las TICs, englobando estas a todas las tecnologías bioinformáticas disponibles (incluyendo las utilizadas en las “omics”, la Inteligencia Artificial, el Big Data, la robótica médica, el diagnóstico por imagen, etc.), y la generación de recursos que puedan ser utilizados por la comunidad científica dentro de la filosofía “Open Science” de la Unión Europea.

Estos avances han hecho posible definir el siguiente gran objetivo de la biomedicina y la asistencia en salud: la Medicina Personalizada (MPe) y la Medicina de Precisión (MPr). Entendiendo por MPe aquella en la que los procesos de diagnóstico, terapia, control y seguimiento se hacen “ad hoc” para cada persona, en función de sus características “ómicas” (genómicas, epigenómicas, proteómicas, metabolómicas, exposómicas, ...) y de la interacción con su propio entorno. Y nos referimos a la MPr cuando se opta por realizar los actos asistenciales de la manera menos intrusiva y con los menores efectos colaterales adversos posibles, recurriendo incluso a técnicas de escala nanométrica; por ejemplo, administrar nanomoléculas que actúen específicamente sobre células, orgánulos, proteínas específicas; hacer microincisiones quirúrgicas para evitar infecciones o nanointervenciones para sustituir determinadas bases genéticas, etc. Estos dos conceptos, Mpr y Mpe, se interpretan aquí diferentes, aunque en la literatura muchas veces se manejan indistintamente, considerándolos como equivalentes. Ambas abren nuevas oportunidades para tratar mejor las enfermedades complejas y una ventana de oportunidad para aquellos pacientes de enfermedades raras o muy poco frecuentes que en el anterior paradigma tenían dificultades para su atención debido a no tener una amplia prevalencia en la población. MPr y MPe, se potencian a partir del lanzamiento estadounidense en enero de 2015, del reto “Precision Medicine Initiative” (PMI). En las propias palabras del presidente Obama *“para acercarnos a la curación de enfermedades como el cáncer y la diabetes, y para darnos a todos acceso a la información personalizada que necesitamos para mantenernos a nosotros y a nuestras familias*

más saludables”. Este enfoque de tratamiento y prevención de enfermedades busca maximizar la efectividad teniendo en cuenta la variabilidad individual en los genes, el medio ambiente y el estilo de vida. De esta manera, se propone y obliga a un cambio en el paradigma de asistencia e investigación en todo lo referente a la salud. Así, se creó la necesidad de la individualización del proceso asistencial en todas sus etapas, considerando que la generación de diferentes y múltiples datos, informaciones y conocimientos (DIC) de cada persona, y su posterior análisis, es capaz de crear un entorno adecuado para desempeñar una asistencia integrada específica para cada individuo.

Actualmente, el trabajo que está haciendo la comunidad científica internacional para implementar unas MPe y MPr generales y robustas en el proceso asistencial se centran en dos aspectos:

- **Catalogación genómica de las enfermedades:** el objetivo principal de este campo es la creación de un mapa genómico completo que catalogue cada una de las variantes posibles de todas las enfermedades y su interacción con el ambiente que lo rodea. De esta manera, al no presentar toda la información de variabilidad de una enfermedad en términos moleculares, siempre existirán pacientes que no puedan disponer de un diseño de tratamiento personalizado.
- **Integración y análisis de datos multi-ómicos y de entorno de forma eficiente:** la descripción de una enfermedad no puede ser dada únicamente con un tipo de dato genómico, sino que tiene que ser complementada con una gran variedad de datos provenientes de múltiples fuentes heterogéneas (proteómicos, epigenómicos, ambientales, clínicos, histopatológicos, etc.). Uno de los mayores retos a los que se enfrentan los investigadores es la creación de modelos informáticos que puedan asimilar de forma eficiente todos estos datos para integrarlos en los modelos predictivos. Así, en un futuro, al paciente se le extraerán los diferentes datos “ómicos”, así como datos de su entorno, estilo de vida, imágenes médicas y de otras pruebas clínicas, etc. Se gestionarán en plataformas in-

teligentes, y serán estas plataformas los que proporcionen un *output* predictivo que ayude al clínico en la toma de decisiones.

Si bien es cierto que la investigación biomédica ha experimentado últimamente un avance significativo, existen muchos campos en los que todavía queda mucho por avanzar y desarrollar nuevas metodologías. Un aspecto clave en los últimos años ha sido la multidisciplinariedad de los equipos de investigación. Aunar esfuerzos desde diferentes campos de investigación ha producido grandes avances en el cuidado de la salud de las personas. Es muy ilustrativo el ejemplo de la Inteligencia Artificial en el ámbito Biomédico.

La incorporación gradual de las plataformas de secuenciación masiva en la rutina diaria de un laboratorio y la reducción de los costes, favorece que cada día se genere una gran cantidad de datos que permitirán estudiar con mayores posibilidades de éxito problemas biomédicos complejos. Además, los avances tecnológicos producidos en la generación de diversos datos “ómicos” propician la generación de dichos datos con una mucha mayor calidad y velocidad de obtención. Todos estos avances, tanto en el campo de la biotecnología, como en el de la bioinformática, ponen a disposición del investigador y del clínico un amplio y heterogéneo catálogo de DIC que pueden ser explotados para el mejor abordaje de las diferentes enfermedades.

Gracias al auge y el éxito que están obteniendo actualmente las técnicas de Inteligencia Artificial (IA), en concreto el Machine Learning (ML) y “Deep Learning” (DL), sumado a su éxito cuando se enfrenta a conjuntos de datos de muy elevadas dimensiones (BD), se ha abierto una gran oportunidad incluyendo la utilización de datos biomédicos, de entorno y de la propia historia clínica de las personas para el entrenamiento de algoritmos de ML/DL que son capaces de encontrar patrones en datos complejos y posteriormente realizar predicciones, clasificaciones,... que la estadística convencional tendría serias dificultades, sino imposibilidad real, de abordar con éxito en sus resultados.

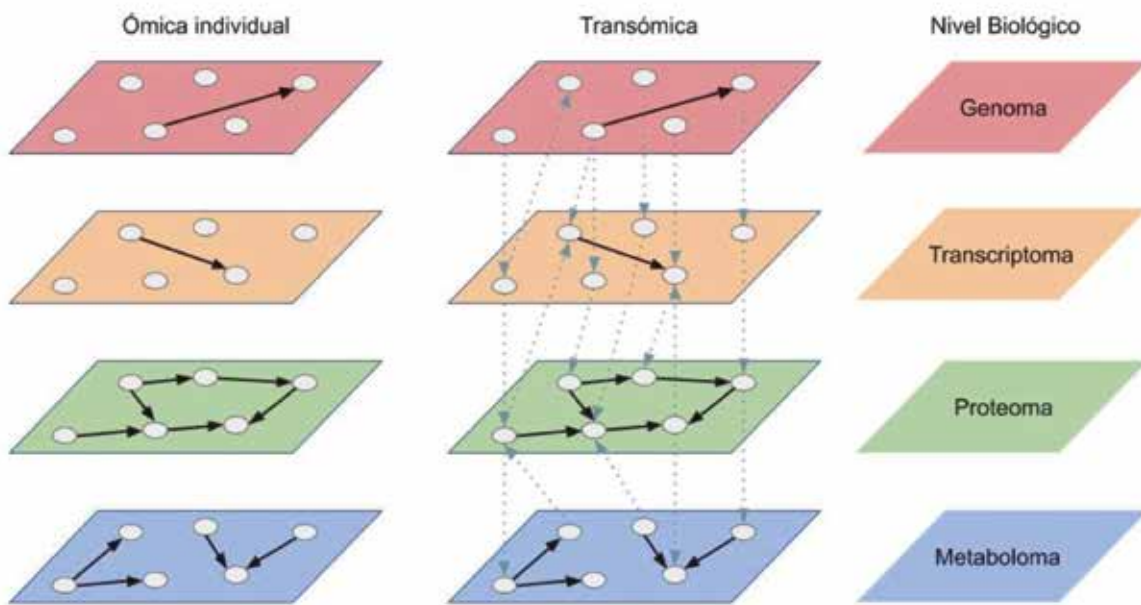
Una vez que se puede disponer de herramientas con la potencia suficiente para gestionar y analizar todas estas grandes cantidades de DIC, el

principal cuello de botella se encuentra en la calidad de esos DIC que han de ser gestionados y analizados. Por ello, se están haciendo grandes esfuerzos en la generación de datos de alta calidad para poder tener éxito en entrenamiento de las herramientas de IA que asegure un buen funcionamiento de las mismas cuando se encuentren en fase de ejecución. Además, en aras de una mayor colaboración de la comunidad científica que promocióne la multidisciplinariedad, se ha establecido una creciente tendencia a alojar y almacenar datos en la nube para que estos estén al alcance de todos los investigadores y clínicos interesados de una forma gratuita, propiciado por filosofías como la “Open Science” de la Unión Europea”. En esta búsqueda de calidad se encuentra también la iniciativa FAIR (Findible, Accesible, Interoperable y Reutilizable) de la propia Unión europea, que establece que los datos y resultados de los proyectos de I+D+i se han de poder localizar, acceder, integrar y reutilizar.

DATOS BIOMÉDICOS DE ACCESO PÚBLICO

Cada célula humana tiene una secuencia de ADN aproximada de 3000 millones de pares de bases distribuidos en 23 pares de cromosomas, donde se encuentran alrededor de 20.000 genes. Estos genes producen entre 10^5 y 10^6 moléculas de ARNm, que posteriormente darán lugar a unas 100.000 proteínas diferentes, expresadas específicamente en cada célula y tejido. Todo esto propicia que los esfuerzos actualmente se centren a nivel molecular y submolecular, aludiendo a diferentes niveles del sistema biológico con el fin de encontrar biomarcadores a los que orientar la atención investigadora y, o, asistencial. Estos distintos datos “ómicos” pueden provenir de diferentes niveles biológicos. Atendiendo al nivel en el que se centre la investigación, se podría hablar de datos: genómicos (secuencias de DNA, variación del número de copias (CNV, etc.), transcritómicos (expresión génica, splicing alternativo, etc.), proteómicos (expresión de proteínas y modificaciones post-traducción, etc.), etc. Además, cabe resaltar que estos datos “ómicos” no son datos estáticos; es decir, a lo largo de los diferentes niveles del sistema biológico se van a tener interacciones (regulación, inhibición, ...), que pueden

FIGURA 1. GRÁFICO DE LOS DIFERENTES NIVELES BIOLÓGICOS Y DE LOS POSIBLES ENFOQUES QUE SE PUEDEN TOMAR CON SUS DATOS “ÓMICOS”.



ser intranivel, internivel e incluso con el ambiente con el cual está en contacto el propio sistema, por lo que, si se observan los datos “ómicos” como un proceso, las diferentes alteraciones que pueden producirse a lo largo del mismo son casi incalculables (ver Figura 1).

Realizar modelos y analizar toda esta complejidad molecular y submolecular, a partir de DIC y en un entorno multidisciplinar, pone en el foco la necesidad inmediata de una estandarización de todos los procesos y de la creación de un lenguaje común, una ontología “ómica” que asegure la calidad de los DIC que han de ser gestionados y analizados. Se están realizando grandes esfuerzos a nivel internacional para disponer de datos de alta calidad.

A lo largo de los últimos años, han ido surgiendo iniciativas y proyectos que secundan estas bases y presentan bases de datos de libre acceso que pueden ser utilizadas por cualquier grupo de investigación y, o, investigador individual. Un ejemplo es el proyecto *The Cancer Genome Atlas* (TCGA). El TCGA es un proyecto de genómica

sobre el cáncer, donde se caracterizó molecularmente a más de 11.000 pacientes de cáncer y personas sanas, permitiendo el estudio de más de 33 tipos y subtipos de cáncer, incluyendo 10 cánceres raros. Actualmente, provee un repositorio de diferentes tipos de datos que van desde secuencias de ADN, diferentes tipos de ARN, mutaciones en el número de copias, datos de expresión, datos de metilación o datos de expresión proteica. La característica más importante, además de ser una amplísima muestra, es que estos datos han sido puestos de forma abierta para el acceso de cualquier investigador. De esta manera, diferentes grupos de trabajo pueden aportar nuevas metodologías y, o, enfoques que ayuden en el abordaje terapéutico del cáncer.

Un proyecto todavía mayor es el *The International Cancer Genome Consortium* (ICGC), el cual engloba diferentes cohortes de pacientes (incluyendo el TCGA), relacionados con cáncer, poniéndolas a disposición también de los investigadores. En este caso, el ICGC consta de 86 proyectos de diferentes países. Esta base de da-

tos no sólo presenta datos genómicos, sino que también alberga datos de diferentes plataformas biotecnológicas, capaces de inferir niveles de expresión genética, proteínas, perfiles de metilación, anotación de mutaciones o incluso niveles de diferentes transcripciones.

Además de las bases de datos relacionadas con la genómica, existen otras más relacionadas con la química, como es el caso de ChEMBL. Esta base de datos recoge información de moléculas bioactivas con propiedades similares a las de los medicamentos y está disponible de forma gratuita. Incluye información sobre cómo las pequeñas moléculas interactúan con sus dianas proteicas, cómo estos compuestos afectan a las células y a organismos enteros, e información sobre absorción, distribución, metabolismo, excreción y toxicidad. La base de datos es única debido a su enfoque en todos los aspectos del descubrimiento de fármacos y a su tamaño, ya que contiene información sobre más de 1,8 millones de compuestos y más de 15 millones de registros de sus efectos en los sistemas biológicos. De esta manera, se puede acceder a información, ya no solo para la caracterización de los pacientes, si no también es posible estudiar los diferentes tratamientos.

Estos que hemos citado son sólo algunos de los ejemplos que existen en la actualidad de bases de datos, plataformas y proyectos que favorecen una investigación transversal y global en biomedicina. De esta manera, grupos de investigación con una especialización más técnica, y sin posibilidad de generar datos masivos de pacientes, puedan ofrecer soluciones en el diagnóstico y tratamiento de las diferentes patologías. Además, como se ha comentado anteriormente, la gran cantidad de DIC generados y puestos a disposición en esta filosofía "Open Science" crea un contexto idóneo para la aplicación de nuevas herramientas de gestión y análisis. En los últimos años, la gestión y el análisis de DIC ha sido llevado a cabo principalmente por técnicas y procedimientos basados en la Inteligencia Artificial. Estas técnicas y procedimientos de IA están experimentando un crecimiento realmente exponencial y están siendo trasladadas, cada vez más, al entorno biomédico investigador y asistencial, como apoyo a la toma de decisiones en el ámbito asistencial.

A continuación, se tomará como referencia este contexto y se hará un repaso de cómo las técnicas y procedimientos basadas en IA han ofrecido soluciones en el campo de la biomedicina a partir de bases de datos públicas.

INVESTIGACIÓN BASADA EN ML/DL A PARTIR DE DATOS PÚBLICOS

La ML/DL son técnicas de IA por las cuales los computadores adquieren la capacidad de aprender de una forma automática, a partir de ejemplos en lugar de instrucciones, ofreciendo, una vez son entrenadas adecuadamente, la capacidad de ejecutar acciones o comportamientos que, si lo llevase a cabo un humano, se diría que es inteligente. Esto se logra mediante diferentes programas informáticos que permiten al computador aprender, después de haber sido entrenado con un conjunto de ejemplos, un comportamiento (predecir, clasificar, identificar, etc.). Estos programas trabajan con ejemplos que han sido etiquetados previamente, intentando encontrar un patrón que, dadas las variables de entrada, les asigne una etiqueta de salida adecuada. Estos programas son entrenados con un subconjunto del total, denominado conjunto de entrenamiento, que le permite configurar los parámetros internos del mismo para ofrecer una minimización del error de salida. El subconjunto que no pertenece al entrenamiento se denomina "test", y tiene como finalidad hacer posible la validación del modelo, así como la comprobación de su rendimiento. Una vez entrenado se busca que el programa de ML/DL tenga la mayor capacidad de generalización posible para que el modelo funcione adecuadamente, no sólo con los casos con los que ha aprendido, sino también con casos con los que no haya tenido contacto en el proceso de entrenamiento.

Dentro del contexto biomédico en el que nos encontramos, donde existe una gran cantidad de datos abiertos al público, las técnicas de ML/DL pueden ofrecer una solución en la gestión y el análisis de este tipo de DIC. A diferencia de las técnicas estadísticas convencionales, diseñadas para el análisis de datos con pocas variables y distribuciones muy concretas, las técnicas de ML/DL son capaces de extraer patrones no li-



“Dentro del contexto biomédico en el que nos encontramos, donde existe una gran cantidad de datos abiertos al público, las técnicas de ML/DL pueden ofrecer una solución en la gestión y el análisis de este tipo de DIC

neales a partir de datos complejos, donde existen muchas más variables que observaciones en nuestros datos. Con esta ventaja que presentan las técnicas de ML/DL, se apunta que un punto crucial es el diseño metodológico de los experimentos. Este diseño debe estar estandarizado y utilizando siempre la misma etiqueta lingüística para el mismo concepto, de cara a obtener resultados robustos y reproducibles por los demás investigadores.

Para dar una idea de las oportunidades que ofrecen estas técnicas, y su utilización con datos públicos, a continuación, se recogen algunas metodologías que han logrado grandes resultados en el campo de la oncogenómica.

Por ejemplo, muchos investigadores han utilizado los datos del TCGA para el entrenamiento de ciertos algoritmos basados en ML. El objetivo ha sido muy variado si uno se fija en los diferentes trabajos publicados. Entre ellos, la predicción en la respuesta de tratamientos inmunológicos. Actualmente, una de las terapias más exitosas y prometedoras contra el cáncer son los medicamentos que actúan contra las dianas inmunológicas. La activación de ciertos puntos de control inmunológico de la célula por parte de las células tumorales puede hacer que las células de nuestro sistema inmunológico no sean capaces de eliminar a las células tumorales. El tratamiento de la mayoría de los tipos de tumores se ve favorecido por este tipo de terapia, aunque hay algunos que no responden de la misma manera. En el trabajo de Dai, Y. et al

[3], desarrollaron y validaron una firma de ocho características basada en la imagen radiológica de células CD8 para la respuesta a fármacos contra PD-1 y PD-L1. Este predictor de imágenes ofrece una forma prometedora de predecir el fenotipo inmunológico de los tumores e inferir resultados clínicos para los pacientes de cáncer que han sido tratados con anti-PD-1 y PD-L1.

Se conocen bien algunos de los factores genéticos específicos de cada tumor, así como ciertos “pathways” que influyen en el proceso de desarrollo de los tumores. Aunque la identificación del estado de una ruta metabólica es una cuestión compleja, es de gran utilidad en el diagnóstico, la estratificación y el tratamiento de los pacientes. El trabajo desarrollado por Way, G. et al [4] ofrece una solución a este problema. Utilizaron un enfoque de ML basado en el concepto de la regresión logística, integrando datos de mutación, número de copias y expresión de todos los datos disponibles en el TCGA, con la finalidad de detectar variantes de activación del “pathway” RAS (KRAS, HRAS, o NRAS) en tumores. Obtuvieron un 0.86 de AUC en el rendimiento de su modelo a la hora de clasificar los datos entre las diferentes variantes.

CONCLUSIONES

Como se puede observar por lo previamente citado, las técnicas y procedimientos de Inteligencia Artificial tienen una gran capacidad de

EMPLEO DE IA POR ÁREAS TERAPÉUTICAS

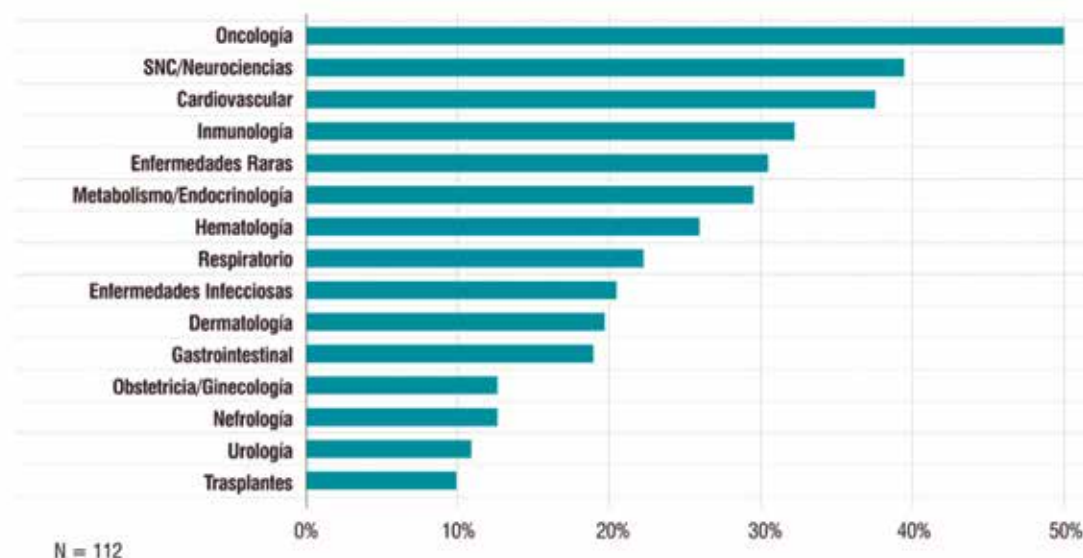


FIGURA 2. UTILIZACIÓN DE LA IA POR ÁREAS TERAPÉUTICAS.
(FUENTE: TUFTS CENTER FOR THE STUDY OF DRUG DEVELOPMENT)

explotación de las grandes cantidades de DIC generados en el mundo de la biomedicina, siendo capaces de extraer el conocimiento subyacente de un conjunto de datos e informaciones, por lo que es relevante comprender la idoneidad de los mismos. En otras palabras, estas técnicas y procedimientos deben usarse con ciertas precauciones y los investigadores y los clínicos deben ser conscientes de que las conclusiones que obtienen pueden estar alteradas debido a una mala selección de los DIC sobre los que se aplican estas técnicas y procedimientos de IA o a una inadecuada metodología de ejecución y análisis de los procesos.

Es crucial que los DIC, metodologías y resultados sean compartidos y abiertos para que los distintos interesados (investigadores, personal asistencial, gestores, ...) puedan beneficiarse de ellos y contribuir en el avance del conocimiento de los procesos que afectan a la salud. En este contexto, dos conceptos crecen en importancia: *Open Science* y *FAIR*. El primero de ellos, "*Open Science*", como ya se ha dicho, hace referencia a que deben de ponerse a disposición de forma abierta a todo el mundo con el fin de que la colaboración interdisciplinar pueda ayudar en la resolución de

los problemas complejos relacionados con la salud. Es complicado imaginar, a día de hoy, algún avance significativo en Ciencia sin la aportación de muchos investigadores, provenientes de muy diferentes ámbitos disciplinares, trabajando de manera colaborativa sobre el mismo problema. El segundo concepto, "*FAIR*", alude a la calidad de los DIC y los repositorios. Se están haciendo grandes esfuerzos para que todos los datos puestos de forma abierta cumplan los principios FAIR [5] [6] ya citados: ser localizables, accesibles, interoperables y reproducibles.

Es imprescindible poner el foco en la calidad de los DIC. Existen numerosas herramientas que son capaces de explotar los DIC a un nivel muy elevado desde el punto de vista de capacidad computacional y tiempo de ejecución, pero si no se abastecen de DIC de suficiente calidad, no se va a ser capaz de extraer todo su potencial, o incluso se obtendrán soluciones inadecuadas. Son muchos los países y organizaciones que están siguiendo este camino. La misma Unión Europea, planifica para 2025 finalizar la construcción de la "*European Open Science Cloud*" (EOSC) [7], que será un entorno virtual federado de confianza que obvia fronteras y disciplinas científicas para

almacenar, compartir, procesar y reutilizar objetos digitales de investigación (como publicaciones, datos y programas informáticos) que sigan los principios FAIR.

Respecto a los desafíos que se presentan en la gestión de los datos, sobre todo cuando estamos en entornos de “*Big data*”, la IA aprovecha los enfoques de aprendizaje profundo (DL) para superar los obstáculos inherentes a ellos. En los entornos clínicos, la IA funciona, y debe de funcionar, como un asistente que ayuda al personal asistencial a trabajar de forma más eficiente y a realizar diagnósticos más precisos, lo que ayuda a mejorar la calidad asistencial y la productividad de los servicios de salud. A un nivel más amplio, la IA ayuda a las empresas relacionadas con el ámbito de la salud a acelerar y hacer más precisos sus desarrollos; un ejemplo es el caso de la farmaindustria desarrollando técnicas y procedimientos de descubrimiento temprano y reposicionamiento de fármacos para reducir costos y lograr que estén accesibles en el menor tiempo posible.

Entre los principales usos de la IA en el ámbito de la atención a la salud, se encuentran:

- Diagnóstico por imagen
- Detección precoz de procesos cancerígenos
- Descubrimiento temprano y reposicionamiento de fármacos
- Predicción del riesgo de enfermedades
- Control y seguimiento de enfermedades crónicas

Como se puede ver en la Figura 2, se constata la implementación de la IA en las múltiples áreas médicas que comienzan a emplear métodos y procedimientos de trabajo BIOTIC de forma generalizada, postulándose como una aproximación de clara tendencia al alza en el futuro.

BIBLIOGRAFÍA

- Dai Y, Sun C, Feng Y, Jia Q, Zhu B. Potent immunogenicity in BRCA1-mutated patients with high-grade serous ovarian carcinoma. *J Cell Mol Med*. 2018. doi:10.1111/jcmm.13678
- <https://www.genome.gov/human-genome-project/Completion-FAQ>.
- Way GP, Sanchez-Vega F, La K, Armenia J, Chatila WK, Luna A, et al. Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. *Cell Rep*. 2018;23:172–180.e3. doi:10.1016/j.celrep.2018.03.046
- Way GP, Sanchez-Vega F, La K, et al. Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. *Cell Rep*. 2018;23(1):172-180.e3. doi:10.1016/j.celrep.2018.03.046
- Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. As Open as Possible, as Closed as Necessary’--Managing legal and owner. 2016;3.
- <https://www.fair4health.eu/en/news/fair-principles-for-cancer-research>
- https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/european-open-science-cloud-eosc_en

AGRADECIMIENTOS:

This work is supported by: “Proyecto colaborativo de integración de datos genómicos (CICLOGEN)” PI17/01561 funded by the Carlos III Health Institute from the Spanish National plan for Scientific and Technical Research and Innovation 2017-2020; the Spanish Ministry of Economy and Competitiveness through the project BIA2017-86738-R and through the funding of the unique installation BIOCAI (UNLC08-1E-002, UNLC13-13-3503) and the European Regional Development Funds (FEDER) by the European Union. Additional support was offered by the Consolidation and Structuring of Competitive Research Units—Competitive Reference Groups (ED431C 2018/49) and Accreditation, Structuring, and Improvement of Consolidated Research Units and Singular Centers (ED431G/01), funded by the Ministry of Education, University and Vocational Training of the Xunta de Galicia endowed with EU FEDER funds