Fundación InstitutoRoche 15 aniversario

Informes Anticipando

LOS DATOS

EN LA ERA DE LA MEDICINA PERSONALIZADA DE PRECISIÓN

Observatorio de Tendencias | DE MEDICINA PERSONALIZADA DE PRECISIÓN

Informe Anticipando coordinado por:

Fernando Martín-Sánchez

Profesor de Investigación. Director del Programa de Salud Digital, Cronicidad y Cuidados del Instituto de Salud Carlos III.

Alfonso Valencia Herrera

Profesor de Investigación ICREA en el Centro de Supercomputación de Barcelona - Centro Nacional de Supercomputación (BSC-CNS).

Expertos colaboradores:

Fátima Al-Shahrour

Jefa de la Unidad de Bioinformática del Centro Nacional de Investigaciones Oncológicas (CNIO). Co-directora del Máster en Bioinformática aplicada a la Medicina Personalizada y la Salud (ISCIII).

Nuria Malats Malats

Jefa del Grupo de Genética y Epidemiología Molecular del Centro Nacional de Investigaciones Oncológicas (CNIO).

Víctor Maojo García

Catedrático de Universidad del Departamento de Inteligencia Artificial de la Universidad Politécnica de Madrid.

Arcadi Navarro Cuartriellas

Profesor de Investigación ICREA y Catedrático de Genética de la Universidad Pompeu Fabra, Barcelona.

David Pérez Fernández

Responsable dentro del gabinete del Secretario de Estado de Avance Digital del Plan de tecnologías del lenguaje.

Pablo Serrano Balazote

Director de Planificación del Hospital Universitario 12 de Octubre e Investigador del Instituto de Investigación del Hospital Universitario 12 de Octubre (i+12).

Comité Asesor Observatorio de Tendencias de Medicina Personalizada de Precisión:

Joaquín Arenas

Director del Instituto de Investigación del Hospital Universitario 12 de Octubre (i+12).

Ángel Carracedo

Director de la Fundación Pública Gallega de Medicina Genómica (Servicio Gallego de Salud) y Coordinador del Grupo de Medicina Genómica de la Universidad de Santiago de Compostela (CIBERER).

Pablo Lapunzina

Jefe de grupo de investigación del Instituto de Genética Médica y Molecular (INGEMM) del idiPaz y Director científico del CIBERER.

Nº de depósito legal: M-33102-2019

www.institutoroche.es

Con la colaboración de Ascendo Consulting Sanidad&Farma

Contenidos

PRESENTACIÓN

Los Informes Anticipando, elaborados en el marco del Observatorio de Tendencias de Medicina Personalizada de Precisión (MPP) promovido por la Fundación Instituto Roche, surgen con el objetivo de contribuir a la generación y puesta en común del conocimiento, así como a la difusión de los avances que se producen en la evolución de la MPP para colaborar a traer al presente la medicina del futuro.

El Observatorio cuenta con un Comité Asesor de expertos formado por el Dr. Ángel Carracedo, el Dr. Joaquín Arenas y el Dr. Pablo Lapunzina. Entre sus funciones se incluye la selección de las temáticas que abordan estos informes, la identificación de expertos y la validación de los contenidos.

Este informe que versa sobre los datos en la era de la MPP está coordinado por el Dr. Fernando Martín y por el Dr. Alfonso Valencia y en su elaboración han participado como expertos la Dra. Fátima Al-Shahrour, la Dra. Nuria Malats, el Dr. Víctor Maojo, el Dr. Arcadi Navarro, el Dr. David Pérez y el Dr. Pablo Serrano.

El Dr. Fernando Martín es Profesor de Investigación en Informática Biomédica y Director del Programa de Salud Digital, Cronicidad y Cuidados en el Instituto de Salud Carlos III . Desde 2015 hasta 2017 fue Catedrático de Informática de la Salud en Weill Cornell Medicine (Cornell University) e investigador en la Iniciativa de Medicina de Precisión de EE.UU. Antes de esto (2011-2015), fue Chair y Catedrático de Informática de Salud en la Facultad de Medicina y Director fundador (2013) del Centro de Investigación en Informática Biomédica y de la Salud (HaBIC) en la Universidad de Melbourne (Australia). Es Doctor en Informática y en Medicina; Máster en Ingeniería del Conocimiento y Licenciado en Bioquímica y Biología Molecular. Sus intereses de investigación incluyen métodos y sistemas informáticos en investigación traslacional (integración y análisis de datos), medicina de precisión (procesamiento de datos del exposoma) y salud participativa (redes sociales, Quantified-self). Es miembro electo (*fellow*) de ACMI (American College of Medical Informatics) y ACHI (Australasian College of Health Informatics), miembro fundador de la International Academy of Health Sciences Informatics (IAHSI) y catedrático honorario de la Facultad de Medicina de la Universidad de Melbourne.

El Dr. Alfonso Valencia, doctorado por la UAM en Bioquímica y Biología Molecular, realizó su investigación PostDoctoral en Bioinformática en EMBL Heidelberg. Actualmente es profesor en el Instituto Catalán de Investigación y Estudios Avanzados (ICREA), Director del Departamento de las Ciencias de la Vida del Barcelona Supercomputing Center (Centro Nacional de Supercomputación, BSC-CNS), donde se encuentra el Supercomputador MareNostrum. Es Director del Instituto Nacional de Bioinformática (INB-ISCIII) y líder del nodo español de la Infraestructura Europea para la Información de las Ciencias de la Vida (ELIXIR) y miembro fundador, *fellow* y Presidente de la Sociedad Internacional de Biología Computacional (ISCB). Miembro electo de la Organización Europea de la Biología Molecular (EMBO). Editor Ejecutivo de Bioinformatics, OUP y editor de otras revistas como eLIFE, Profesor Honoris Causa por la Universidad Técnica danesa (DTU) y asesor de múltiples Instituciones. El Dr. Alfonso Valencia es pionero en la aplicación de las ciencias computacionales a la resolución de problemas biológicos, y reconocido por su liderazgo en este campo. A lo largo de su carrera se ha centrado en el análisis de grandes colecciones de datos genómicos, especialmente en el estudio de las redes de interacción en Epigenómica, Biología del Cáncer y la Medicina de Precisión, así como la aplicación de la metodología de minería de textos para solucionar problemas biomédicos.

La Dra. Fátima Al-Shahrour, es jefa de la Unidad de Bioinformática del Centro Nacional de Investigaciones

Oncológicas (CNIO). Posee una amplia experiencia en el estudio del cáncer bajo una perspectiva genómica. Su investigación se centra en la aplicación y el desarrollo de métodos computacionales para la medicina de precisión, la interpretación de genomas del cáncer, la identificación de nuevos biomarcadores y la predicción de terapias contra el cáncer. Es codirectora del Máster en Bioinformática aplicada a Medicina Personalizada y salud del INS-ISCIII. Miembro de ISCB, EACR y SEBBM y editora asociada en revistas del campo de la Bioinformática.

La Dra. Nuria Malats, es jefa del Grupo de Epidemiología Genética y Molecular del Centro Nacional de Investigaciones Oncológicas (CNIO) desde 2007. Su investigación se centra, principalmente, en cáncer de páncreas, vejiga y mama. Coordina estudios nacionales e internacionales de gran tamaño que integran diferentes niveles de información, incluidos los datos ómicos, tanto en relación al desarrollo como la progresión de la enfermedad. Tiene 250 publicaciones y es revisora externa de agencias de financiación nacionales e internacionales y revistas científicas de primer rango. La Dra. Malats presidió la acción COST EUPancreas (BM1204), es miembro de la junta directiva del Consorcio Internacional de Casos-Controles de Cáncer de Páncreas (PanC4) y preside el Research Work Stream de la plataforma Pancreatic Cancer Europe (PCE).

El Dr. Víctor Maojo, doctor en Medicina por la Universidad de A Coruña y doctor en Informática por la Universidad Politécnica de Madrid, es actualmente Catedrático del Departamento de Inteligencia Artificial de la Universidad Politécnica de Madrid. Su dilatada carrera como investigador incluye entre otros, la participación en más de veinte proyectos nacionales y doce proyectos financiados por la Comisión Europea, en dos de ellos como coordinador. En el campo de los datos en salud, cabe destacar que, en 1994, dirigió uno de los primeros proyectos financiados por el Fondo de Investigación Sanitaria en el área del machine learning y desde 1996 varios proyectos de integración de bases de datos. El proyecto INFOGENMED (2001-2004), que dirigió en la UPM, fue el primer proyecto de la Comisión Europea centrado en la integración de bases de datos clínico-genómicas. En 2011 fue elegido Fellow del American College of Medical Informatics (ACMI), por sus contribuciones en la informática biomédica y en 2017 fue elegido miembro fundador de la nueva International Academy of Health Sciences Informatics, asociada a la International Medical Informatics Association. En 2019 recibió el Premio Nacional de la Sociedad Española de Informática de la Salud (SEIS).

El Dr. Arcadi Navarro es profesor de investigación ICREA desde el año 2006 y Catedrático de Genética en la Universidad Pompeu Fabra donde dirige un grupo de investigación en Genómica Evolutiva en el Departamento de Ciencias Experimentales y de la Salud. Su actividad investigadora se centra en la genómica computacional y la medicina evolutiva a través del estudio de la evolución del genoma y como los patrones de diversidad del genoma se relacionan con el envejecimiento o la susceptibilidad diferencial de diferentes personas a ciertas enfermedades. Destaca su participación como codirector en el European Genome-Phenome Archive (EGPA), proyecto conjunto entre el European Bioinformatics Institute (EBI) y el Centro de Regulación Genómica (CRG) que consiste en un servicio para el archivo permanente y en intercambio de todo tipo de datos genómicos y fenotípicos de identificación personal que resultan de proyectos de investigación biomédica.

El Dr. David Pérez, doctor en Matemáticas, licenciado en Matemáticas y Físico Teórico, pertenece al Gabinete del Secretario de Estado de Avance Digital (MINECO). Dirige el Plan Nacional de Tecnologías del Lenguaje (PlanTL: https://www.plantl.gob.es) de la SEAD, que presenta entre sus ejes principales el sector Salud, cuyo objetivo es desarrollar tecnologías y herramientas en procesamiento de lenguaje natural, sistemas de traducción automática y de diálogo aplicados a la salud. El equipo de minería de texto clínico perteneciente al PlanTL es el responsable de las principales campañas de evaluación y desarrollo de plataformas de procesamiento de lenguaje natural y traducción automática para la Administración. Dicho equipo se enmarca en la colaboración con el Centro de Supercomputación de Barcelona (BSC).

El Dr. Pablo Serrano es Director de Planificación del Hospital Doce de Octubre, ha ocupado puestos de gestión en hospitales y paralelamente investiga en la historia clínica electrónica e interoperabilidad. Ha participado en proyectos nacionales y europeos sobre representación de la información de salud y su utilización en asistencia e investigación. Su grupo trabaja en la actualidad sobre resultados de salud y estudios observacionales a partir de los registros de salud.

RESUMEN EJECUTIVO

Desde hace mucho tiempo, gracias a la capacidad de generación, almacenamiento, procesamiento y análisis de todo tipo de datos, se está produciendo una revolución tecnológica que afecta a distintos ámbitos y, en concreto, está siendo esencial para el progreso de la biomedicina. La tendencia en el campo de la salud consiste en incorporar cada vez más datos que provengan de fuentes muy diversas, de manera que ofrezcan información que puede resultar de gran relevancia para los investigadores, profesionales sanitarios y, sobre todo, para los pacientes.

Los datos que se manejan actualmente en salud son altamente complejos, heterogéneos y deben mantener siempre su carácter confidencial. Actualmente, pueden obtenerse grandes volúmenes de datos a gran velocidad relacionados con diferentes aspectos que condicionan la salud de los individuos. Estos datos, gracias a los avances tecnológicos, pueden integrarse con otros datos que antes o bien no se tenían en cuenta o bien no era posible analizarlos. Por tanto, el diseño de futuras estrategias en salud estará ligado a la integración de los datos provenientes de distintas fuentes, desde historias clínicas a sensores de monitorización, pasando por los datos ómicos. Como consecuencia, se ha impulsado el diseño y mejora de herramientas e infraestructuras computacionales que tengan capacidad de hacer uso de estos datos con la mínima intervención humana.

Estas herramientas para el análisis de los datos, como la minería de datos o la inteligencia artificial (IA), destacando por su aplicabilidad en el campo de la salud lo que se conoce como aprendizaje automático o *machine learning*, tienen un gran potencial como vehículo para impulsar un cambio real en la práctica clínica.

El análisis de los datos en salud, desde historias clínicas a datos ómicos, con herramientas de IA y el conocimiento derivado de los mismos previsiblemente jugarán un papel fundamental en la configuración de la medicina del futuro, ya que dan lugar a nuevas oportunidades y aplicaciones en todas las áreas, desde la investigación biomédica y traslacional hasta la práctica clínica y salud pública. El reposicionamiento de fármacos basado en la identificación de nuevos marcadores, el descubrimiento de multimorbilidades ocultas, el diseño de nuevos modelos de ensayos clínicos, la identificación de patrones y predicción de riesgos asociados a enfermedades a través de redes sociales y los Sistemas de Apoyo a la Decisión Clínica, son algunos ejemplos de las aplicaciones de los datos en salud que están impulsado el desarrollo de lo que se conoce como Medicina Personalizada de Precisión.

INTRODUCCIÓN

LOS DATOS Y LA VISIÓN INTEGRADORA DE LA MEDICINA PERSONALIZADA DE PRECISIÓN

El sector Salud emplea una gran cantidad de datos que provienen de fuentes muy heterogéneas y que permiten obtener información que puede ser de gran relevancia para los investigadores, profesionales sanitarios y para los pacientes. El hecho de que los datos y la información derivada de los mismos resulten fundamentales en el campo de la salud es una realidad reconocida desde hace mucho tiempo. Por tanto, la generación y recolección de grandes cantidades de datos forma parte del campo de la biomedicina desde hace décadas,[1] siendo una más de las "ciencias de datos", entendidas en este caso como aquellas áreas científicas en las que la generación, almacenamiento, procesamiento y análisis de datos son esenciales en su desarrollo.

El conocimiento acumulado tras décadas de investigación ha demostrado que las enfermedades son el resultado de la interacción entre la carga genética de cada individuo y los factores ambientales que lo rodean y a los que está expuesto.[2,3] El desarrollo de la genómica, y posteriormente de otras ómicas (que, de manera proporcional a la aparición de nuevas tecnologías experimentales, cuentan con un número creciente de fuentes de datos ómicos), han supuesto un cambio de paradigma en el estudio y abordaje de las enfermedades que han permitido estudiar aspectos relacionados con el individuo que condicionan su estado tanto de salud como de enfermedad. Las ómicas, junto con la revolución tecnológica en el campo del procesamiento y análisis de grandes cantidades de datos son dos grandes hitos que han contribuido al desarrollo de lo que se conoce como Medicina Personalizada de Precisión.

Debido a la heterogeneidad, complejidad y al carácter confidencial de los datos en salud, es evidente la necesidad de contar con sistemas de almacenamiento, procesamiento y análisis de datos que respondan a estas características específicas con el objetivo de poder extraer información que sea de utilidad en el campo de la Medicina en el futuro. Por tanto, las nuevas herramientas que permiten extraer información significativa y proponer hipótesis verificables a partir de los datos en salud, tienen un gran potencial como vehículo para impulsar un cambio sustancial en la práctica clínica, desde la terapia personalizada y el diseño inteligente de medicamentos hasta la detección de poblaciones de riesgo a padecer una determinada enfermedad o la extracción de información de historias clínicas digitales,[4] para asistir en la toma de decisiones en la práctica clínica y la gestión de los sistemas de salud. Si bien se estima que son necesarios dos o tres órdenes de magnitud superiores en tamaño a los que se generan actualmente[5] para resolver, entre otros, los problemas relacionados con la variabilidad interindividual, se puede afirmar que haciendo uso del conocimiento que proporciona el análisis de estos datos se estará contribuyendo en gran medida a la implementación de la Medicina Personalizada de Precisión, que busca identificar y aplicar un abordaje preventivo, diagnóstico y terapéutico más efectivo atendiendo a la variabilidad genética, la interacción con el ambiente y el estilo de vida de cada paciente.[6]

TIPOS, FUENTES Y CARACTERÍSTICAS DE LOS DATOS EN EL CAMPO DE LA SALUD

TIPOS DE DATOS

Los datos que se pueden emplear en el campo de la salud son de distinta naturaleza y de origen variado y, en general, todos los niveles y fuentes de datos sobre la salud de las personas se consideran importantes desde el punto de vista de la Medicina Personalizada de Precisión, ya que permiten obtener información global de cada individuo.

Todos estos datos pueden clasificarse de diferentes maneras, en función de su origen, procedencia, el responsable de la generación de los datos, naturaleza, frecuencia, nivel de organización biológica, grado de procesamiento, grado de estructuración o enfoque (Tabla 1).

Tabla 1. Tipos de datos en función de diferentes aspectos.

CATEGORÍAS	TIPOS
Origen	Análisis de muestras biológicas / Procedimientos clínicos / Cuestionarios / Sensores / Apps / Sistemas de imagen
Procedencia	Pacientes Individuales / Cohortes / Individuos sanos
Responsable	Generados por pacientes / Generados por profesionales sanitarios / Generados por métodos analíticos
Naturaleza	Cuantitativos / Cualitativos
Frecuencia	Puntuales / De monitorización periódica / Continuos
Nivel de organización biológica*	Molecular/ Celular/ Tisular/ Órgano/ Organismo/ Población/ Ecosistema
Grado de procesamiento**	Primarios / Secundarios / Metadatos
Grado de estructuración***	Estructurados (bases de datos) / No estructurados (textos e imágenes)
Enfoque****	Cantidades pequeñas de datos (small data)/ Grandes cantidades de datos (big data)

A continuación, se realiza una breve explicación de las categorías señaladas en la Tabla 1.

* El nivel de organización de los datos indica la complejidad biológica de las entidades sobre las que se recogen los datos. Esta puede ir desde una molécula (por ejemplo, ADN), hasta una célula, tejido, órgano, individuo, población o ecosistema.

** En cuanto al grado de procesamiento, los datos primarios hacen referencia a los datos originales en bruto, los datos secundarios son datos procesados computacionalmente o manualmente curados a partir de datos primarios y, por último, se consideran metadatos al conjunto de datos que a su vez describen otros datos. Por ejemplo, cuando se secuencia un genoma, los datos primarios serían los que proporciona directamente el secuenciador. La secuencia de nucleótidos alineada completa sería el dato secundario. A veces, se generan nuevas versiones de las secuencias disponibles, por ejemplo, empleando nuevas técnicas. Además, normalmente los archivos de secuencias incluyen una puntuación sobre lo fiable que es esa secuencia. Tanto el número de versión como la puntuación de fiabilidad son ejemplos de lo que constituye un metadato.[7]

*** Por otro lado, en relación con el grado de estructuración, los datos estructurados, son datos con un esquema o modelo de datos definido. Por lo general, los datos que se almacenan en una base de datos suelen estar estructurados como es el caso de los datos derivados de mediciones o señales. Sin embargo, los datos no estructurados hacen referencia a datos que contienen información que no es fácilmente accesible para los sistemas de gestión de datos computacionales, es decir, la información que contienen no se presenta en una forma con un esquema de datos claro que permita una interpretación y análisis computacional directo. Este tipo de datos generalmente requieren métodos analíticos especializados para extraer y transformar la información que contienen. Ejemplo de ello son los textos libres y las imágenes.

**** Además, los datos pueden clasificarse por un lado en lo que podríamos llamar *small data*, enfoque cuyo objetivo es lograr una mejor descripción a nivel individual, predicción y, en última instancia, control de una unidad específica que puede ser un único individuo, un hospital, una comunidad, una ciudad, etc. Estos datos se denominan "pequeños" solo en el sentido de que se recopilan y utilizan para una sola unidad. Es importante destacar que muchos problemas en biomedicina pertenecen intrínsecamente a esta categoría, por ejemplo, el número de DNA circulante de un paciente que puede obtenerse en un experimento concreto. Por otro lado, el enfoque de *big data* se refiere al uso de datos que se recopilan en un grupo muy numeroso de individuos con el objetivo de mejorar la descripción y la predicción de un fenómeno en otros individuos que no necesariamente tienen que ser aquellos de quienes se recopilaron los datos. Por tanto, aunque se tiende hacia proyectos de *big data*[8] (por ejemplo *UK Biobank* o *All of US*, ver Tabla 3), la mayoría de los métodos de investigación en ciencias de la salud como la epidemiología o los ensayos clínicos y ciertamente las ómicas, se podrían encajar en un enfoque de *small data* (incluso se habla de ensayos clínicos n-of-1 sobre un solo individuo, tema que se aborda más adelante en este informe).

FUENTES DE DATOS

Los datos que se manejan en el campo de la salud son altamente complejos y heterogéneos,[9] estando estas características asociadas no solo a su origen, puesto que reflejan procesos distintos, sino también derivadas de las técnicas y protocolos empleados para obtener dichos datos. A todo ello, se suma la heterogeneidad propia de la biología, existiendo variaciones tanto inter como intrapersonales, estas últimas relacionadas con diferentes factores como pueden ser el momento del día en el que se realiza la medición, el ayuno o el ejercicio físico previo.

Los tipos de datos identificados en el apartado anterior permiten caracterizar el genoma, fenoma y exposoma, dimensiones que permiten definir la salud de un individuo (Figura 1), y pueden provenir de diferentes fuentes como son:

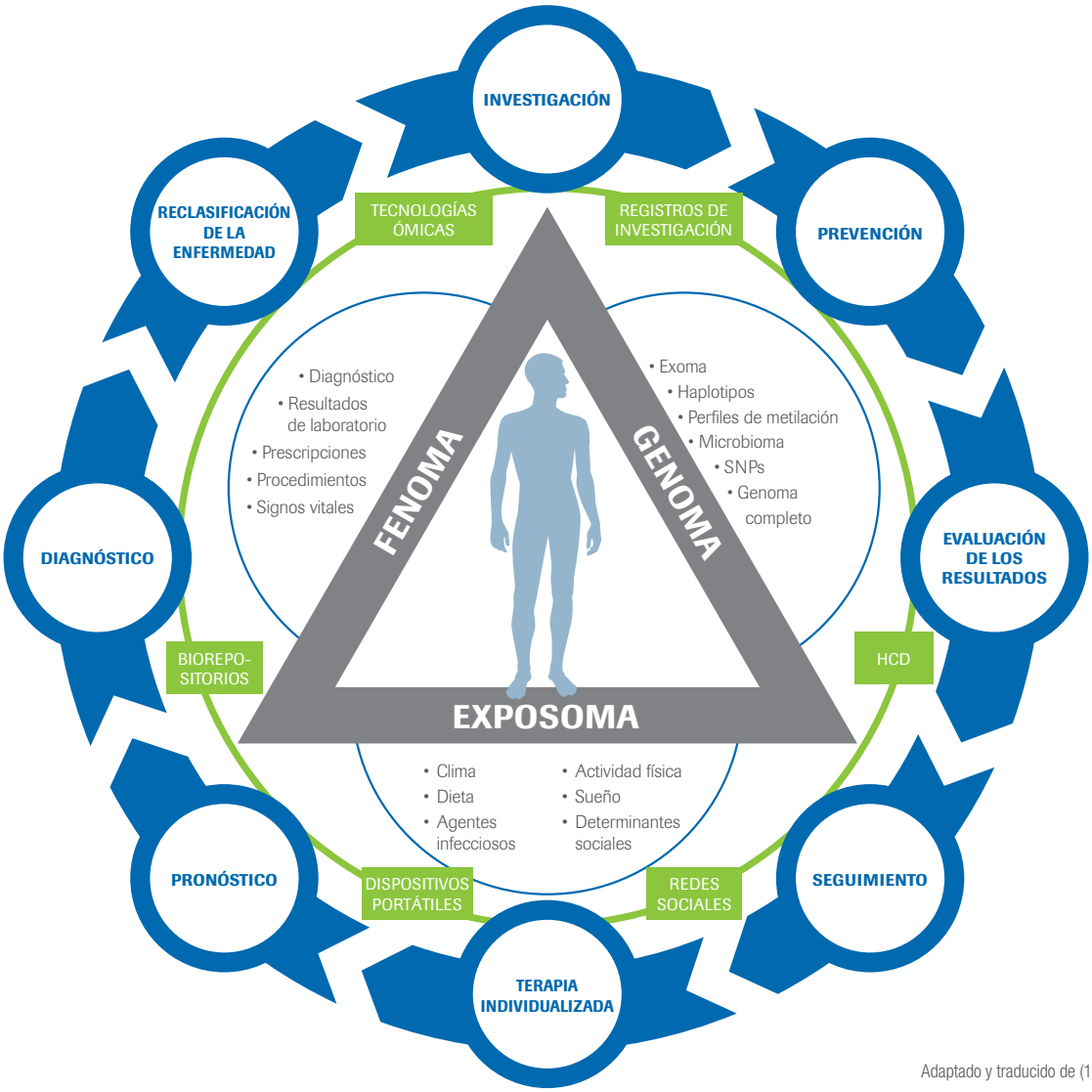
- Tecnologías ómicas. Permiten generar datos genómicos y datos del fenoma molecular (proteómica, metabolómica, etc.), y, por tanto, proporcionan datos sobre los procesos bioquímicos y reguladores que se dan en los organismos vivos, ofreciendo una visión de los diferentes perfiles moleculares, cambios e interacciones biológicas.
- Sistemas de imagen médica. Las ecografías, las radiografías, la resonancia magnética o la tomografía computarizada, que ayudan a la caracterización de fenotipos, ya sean de salud o enfermedad,[10] son una de las principales fuentes de información en el ámbito clínico.
- Instrumentación médica. A través de los instrumentos médicos se recoge una gran cantidad de datos relacionados con la situación clínica del paciente a lo largo del tiempo. Por ejemplo, en las unidades de cuidados intensivos se generan grandes cantidades de datos de origen variado a través de instrumentos médicos, como pueden ser las constantes vitales, los datos de monitorización avanzada, los parámetros del respirador o los parámetros de funcionamiento de equipos complejos como son las bombas de perfusión, los datos de monitorización avanzada o los monitores de diálisis, entre otros.[11]
- Bases de datos científicas. Existen diversos tipos de bases de datos según el tipo de datos que contengan. Por ejemplo, las bases de datos primarias son aquellas que almacenan los datos tal y como han sido depositados por quienes los han generado, de manera que se pueden analizar repetidamente a medida que van surgiendo nuevas herramientas (como por ejemplo EGA[a], la base

de datos de secuencias de genomas de interés médico, mantenida por el EBI-EMBL[b] en colaboración con el CRG[c] y el soporte del BSC, INB-ISCIII y la Fundación La Caixa. Las bases de datos de datos derivados o secundarias[12] son aquellas que almacenan el "valor añadido" derivado del análisis de la información depositada en las bases de datos primarias permitiendo descubrir nuevas propiedades o establecer nuevas relaciones entre los datos. Algunas de estas bases de datos son consideras esenciales para la biomedicina y se conocen como bases de datos centralizadas o core databases[12], como por ejemplo la base de datos con información genómica EGA antes mencionada.

- Ensayos clínicos. Los ensayos clínicos constituyen una fuente de información,[13] que permite realizar análisis retrospectivos y observar cómo es la evolución de los participantes.
- Tecnologías participativas de salud digital. Estas tecnologías hacen referencia a las *apps* y *wearables*, entre otros, que permiten registrar y monitorizar hábitos de vida (por ejemplo, el sueño, la actividad física o la dieta) e incluso parámetros fisiológicos (por ejemplo, la frecuencia cardiaca) a nivel individual. Este tipo de datos comienzan a incorporarse como información adicional a la práctica médica y estas tecnologías a aprobarse como dispositivos médicos. En este sentido, el paciente/individuo juega un papel crucial, dado que es quién recopila todos estos datos.
- Redes sociales e Internet. Existe una gran cantidad de datos derivados del uso de redes sociales así como de las búsquedas en Internet[14] con un gran potencial, ya que si son cruzados con otros datos, pueden servir para la identificación de patrones o la predicción de riesgos.
- Sensores ambientales y sistemas de información geográfica. Estos dispositivos dispuestos por organismos públicos y centros de investigación son herramientas que permiten obtener datos de exposición, relacionados por ejemplo con los niveles de contaminación del aire o la calidad del agua, que condicionan la salud de la población.
- Determinantes sociales y económicos de la salud. Hay una serie de determinantes tanto sociales como económicos que también influyen de manera relevante en el estado de salud y que se recogen a través de las encuestas del Instituto Nacional de Estadística (INE) o bases de datos de tipo gubernamental. Estas últimas permiten conocer datos sobre información socioeconómica, demográfica o sobre el nivel cultural y educativo entre otros, reflejados en el código postal, el centro de estudios, etc.[15]
- Datos aportados por los pacientes. Los individuos tienen un papel muy relevante en la generación y recogida de información, no sólo a través de las tecnologías participativas de salud, sino a través de cuestionarios de salud como es el caso de los *Patient Reported Outcomes* (resultados de salud informados por el paciente), que recogen información de la percepción de, en este caso, el paciente sobre su interacción con el sistema de salud; y la *Patient Reported Experience*, que ofrece información para mejorar la calidad de la atención de salud. Cabe destacar el papel cada vez más relevante que tienen las asociaciones de pacientes proporcionando una visión global de los pacientes más allá de los casos individuales.
- Historia Clínica Digital (HCD). La HCD, cuyo uso cada vez está más extendido, es una digitalización del registro clásico en el que se recoge información sobre el paciente[16] por lo que tiene un papel relevante como sistema colector de información. Esta información puede incluir información registrada desde distintos niveles asistenciales, pruebas de imagen y de laboratorio e incluye los tratamientos y procedimientos a los que se ha sometido cada paciente, entre otros, por lo tanto, se trata, no sólo de datos con distinto grado de estructuración, sino que además se almacenan en formatos muy heterogéneos. El personal sanitario tiene un papel relevante para asegurar la calidad de la información, ya que tratar con el estado incompleto de las HCD sigue siendo una tarea difícil y estas deficiencias han de tenerse en cuenta a la hora de aplicar las técnicas de extracción de datos.[17]

[a] EGA: European Genome-Phenome Archive. http://ega.crg.eu
[b] EBI-EMBL: The European Bioinformatics Institute – The European Molecular Biology Laboratory
[c] CRG: Centre for Genomic Regulation. https://www.crg.eu/en

Figura 1. Información proporcionada por los datos, fuentes de datos y su implicación en la Medicina Personalizada de Precisión.



Adaptado y traducido de (18).

La salud del individuo puede definirse de una manera holística a través de tres grandes dimensiones: **genoma, fenoma y exposoma**. Cada una de estas tres dimensiones puede conocerse mediante la recolección de diversos datos en salud que proporcionan **información derivada de su análisis** sobre, por ejemplo, el genoma completo o el exoma en el caso del genoma, sobre resultados de laboratorio o procedimientos en el caso del fenoma o sobre la dieta o la actividad física en el caso del exposoma. Estos datos, a su vez, provienen de diferentes fuentes como son las tecnologías ómicas, los sistemas de imagen médica, las bases de datos científicas, los ensayos clínicos, las redes sociales e internet, los sensores ambientales, la HCD, etc. A partir de estos datos se obtiene información que resulta **clave para el desarrollo de diferentes actividades en el marco de la Medicina Personalizada de Precisión:** investigación, prevención, evaluación de los resultados, seguimiento, terapia individualizada, pronóstico, diagnóstico y reclasificación de la enfermedad.

CARACTERÍSTICAS DE LOS DATOS

Con el objetivo de favorecer la reutilización y el intercambio de datos se establecieron una serie de pautas o características que deberían presentar tanto los datos como los metadatos. Estas pautas o características conocidas como principios FAIR[19] (Findable, Accesible, Interoperable and Reusable) se establecieron para favorecer que los datos sean localizables, accesibles, interoperables y reutilizables.

- Localizables (Findable)

  Los datos y metadatos pueden ser encontrados tras su publicación mediante herramientas de búsqueda.

- Accesibles (Accesible)

  Los datos y metadatos están accesibles y por tanto pueden ser utilizados por otros investigadores. Una vez localizados los datos, el usuario debe ser capaz de acceder a ellos.

- Interoperables (Interoperable)

  Los datos y metadatos deben estar descritos siguiendo ciertas reglas, utilizando estándares abiertos con el objetivo de permitir su intercambio y reutilización. Es decir, la interoperabilidad hace referencia a la capacidad de integrar conjuntos de datos de distinta naturaleza, origen, formato, etc., de cara a que puedan ser manejados mediante aplicaciones o incluidos en flujos de trabajo para su análisis, almacenamiento y procesamiento.

- Reutilizables (Reusable)

  Los datos y metadatos pueden ser reutilizados por otros investigadores, al quedar clara su procedencia y las condiciones de reutilización. El objetivo final de los principios FAIR es optimizar la reutilización de los datos, y para esto, los datos y los metadatos deben estar correctamente descritos para que puedan ser reproducidos y/o combinados con otros sets de datos.

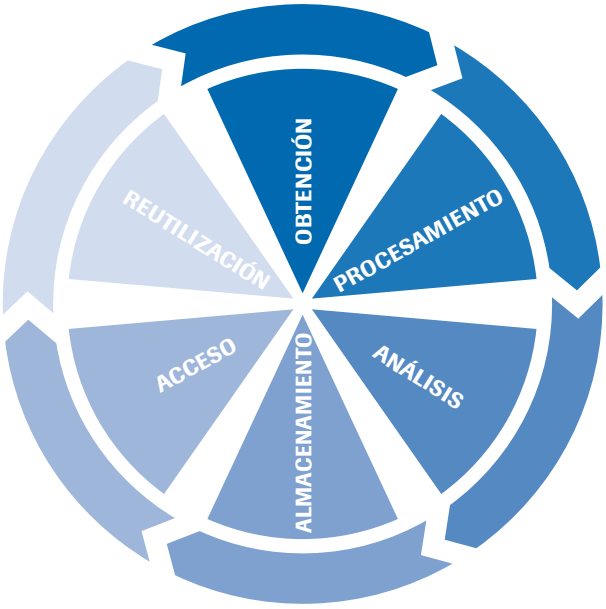
Además de los principios FAIR, al tratarse de datos sensibles de carácter personal, la utilización de los datos en el campo de la salud debe darse bajo estrictas medidas que garanticen su seguridad y, por tanto, otra característica fundamental es que los datos sean confidenciales.

PROCESAMIENTO DE DATOS DE SALUD

La informatización de procesos de todo tipo, como se ha expuesto con anterioridad, ha generado una enorme cantidad de datos[9] a una gran velocidad, lo que ha estimulado el desarrollo de tecnologías y herramientas que gestionan y permiten extraer conocimiento a partir de estos datos complejos.

El denominado ciclo de vida de los datos[20] recoge todas aquellas fases por las que, por lo general, pasan los datos desde su obtención o recogida hasta que son reutilizados comenzando de nuevo el proceso (Figura 2).

Figura 2. Ciclo de vida de los datos.



Tras la **obtención** de los datos es necesario realizar el **procesamiento** que consiste en su manipulación y registro con el objetivo de poder obtener información significativa de los datos tras el **análisis** de los mismos[19]. Este análisis permitirá extraer conclusiones que podrán servir como base en el diseño de nuevas herramientas o enfoques dirigidos a la aplicación de la Medicina Personalizada de Precisión a diferentes niveles. Posteriormente, habrá de garantizarse un correcto **almacenamiento** de los datos y metadatos de manera que se facilite su **acceso** para llevar a cabo nuevas investigaciones y así tanto los datos primarios como los datos secundarios podrán ser **reutilizados**.

Adaptado de (20).

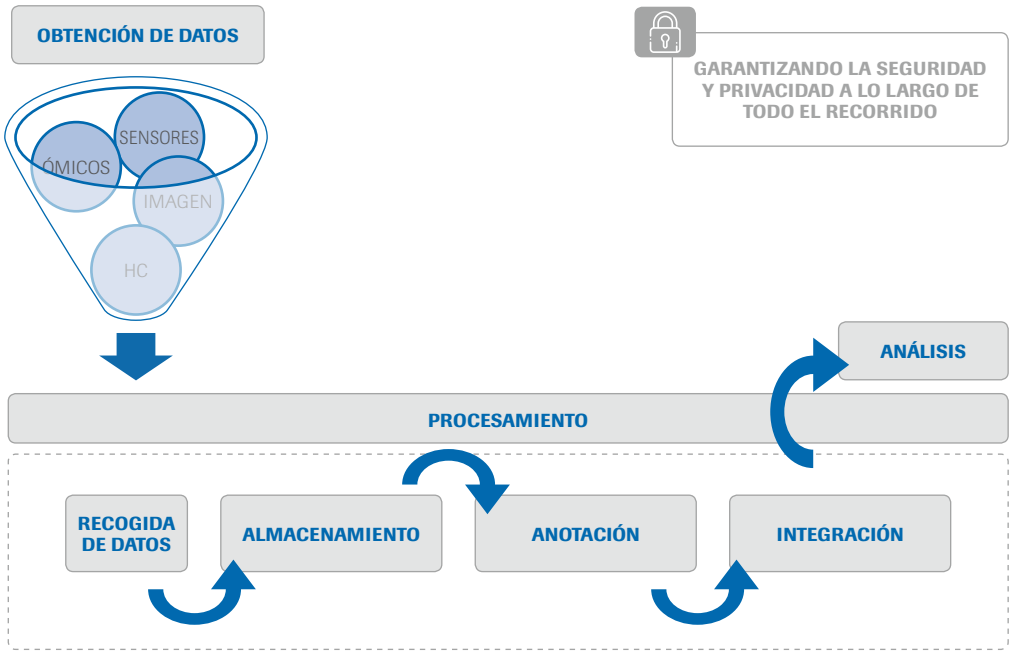
A la hora de abordar el procesamiento de datos de salud, conviene analizar dentro de su ciclo de vida los pasos que deben seguirse (flujo de trabajo) hasta llegar a ser aplicables en el contexto de la Medicina Personalizada de Precisión.

FLUJO DE TRABAJO DEL PROCESAMIENTO DE DATOS PARA SU USO EN SALUD

El flujo de trabajo del procesamiento de datos sigue un esquema general, que puede presentar variaciones en función del ámbito en el que se vayan a aplicar los conocimientos derivados del análisis de los mismos ya sea investigación, práctica clínica o salud pública (Figura 3).

A continuación, se profundiza en cada uno de estos pasos desde la recogida de datos hasta el análisis, que podrán ser distintas según el objetivo que se quiera alcanzar. A lo largo de todo este flujo, como se desarrolla al final del apartado, se habrán de tener en cuenta las medidas necesarias que garanticen la privacidad y la seguridad de los datos.

Figura 3. Flujo de trabajo en el manejo de los datos en el campo de la salud.



Recogida de datos

El primer paso es la recogida de los datos, que dará lugar a conjuntos de datos de diferentes fuentes (ver Figura 1). Estos datos pueden ser generados por distintos grupos de investigación, hospitales o incluso individuos, en los que se siguen distintos protocolos y metodologías, por lo que será necesario realizar un proceso posterior de curación y edición de los datos para asegurar la calidad de los mismos. Es importante que este proceso forme parte de la recogida de datos y durante el mismo se incluyan metadatos completos sobre el origen de las muestras, las características asociadas y los procesos experimentales y computacionales de manera que permitan trazar la procedencia de los datos.

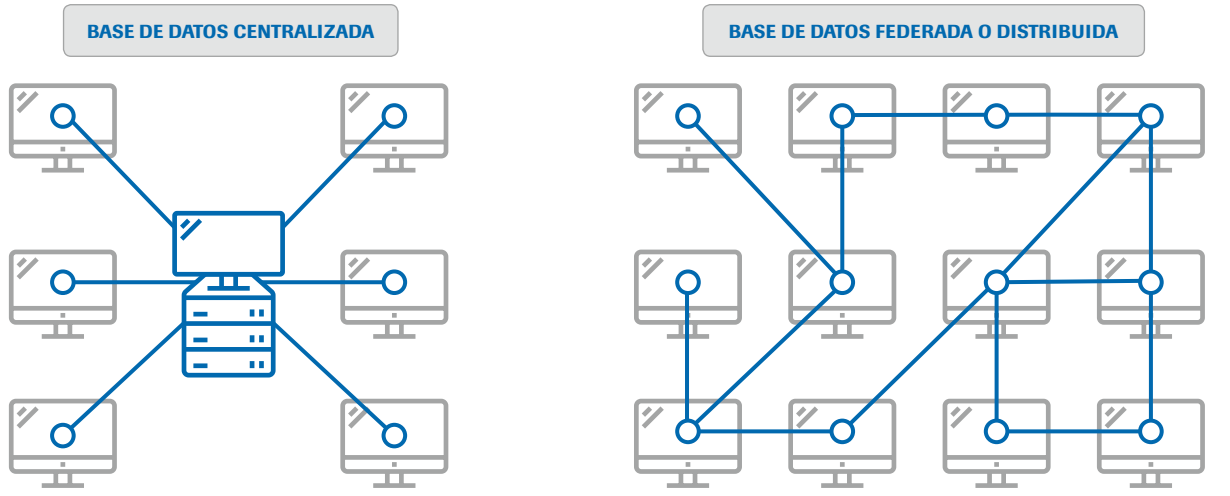
Almacenamiento

Al inicio, la irrupción del *Big Data* supuso un reto para el almacenamiento de datos debido a su volumen, aunque actualmente, de manera general, ha dejado de ser un problema gracias a la existencia de nuevos recursos de almacenamiento físico, nuevas tecnologías de bases de datos y las facilidades de acceso en la nube. En el futuro,

las bases de datos de nueva generación, además de permitir un almacenamiento masivo, responderán a las necesidades actuales de acceso rápido, así como al manejo de las diferentes capas de información.

Actualmente, en cuanto a los tipos de bases de datos, es posible distinguir entre bases de datos centralizadas o federadas según dónde se encuentra la información y cómo se accede a ella (Figura 4).

Figura 4. Base de datos centralizada vs. Base de datos federada o distribuida.



Una **base de datos centralizada** es aquella que se ejecuta en un único sistema informático, es decir, toda la información se encuentra almacenada en una única localización. Mientras que, una **base de datos federada o distribuida**, es aquella en la que existen múltiples bases de datos interconectadas por una red, de manera que cualquier usuario puede acceder a los datos desde cualquier parte de dicha red.

Adaptada de (21).

Anotación

La anotación consiste en aplicar una serie de estándares de normalización y armonización para homogeneizar los datos o referirlos a las mismas magnitudes. Dado que los datos que se emplean en el sector salud son muy heterogéneos, no es posible desarrollar acciones genéricas y cada tipo de dato requiere una estandarización concreta.

Más allá de la digitalización de las historias clínicas, que resulta esencial para trabajar con los datos asociados al fenotipo del paciente, es crucial establecer estándares universales y vocabularios controlados que faciliten la interoperabilidad y el intercambio de datos. En este sentido, a pesar de que existen diversas iniciativas dirigidas a la estandarización de los datos en el campo de salud en diferentes ámbitos (Tabla 2) todavía no se ha alcanzado un consenso a la hora de adoptar unos u otros estándares, a diferencia de otras áreas de la ciencia, como puede ser el campo de la física o la química, en los que sí existen ontologías y estándares reconocidos a nivel mundial.

Tabla 2. Algunas iniciativas relevantes de estandarización en el campo de la salud.

ESTÁNDAR	NOMBRE COMPLETO	ÁMBITO	ORGANIZACIÓN
SNOMED CT	Systematic Nomenclature of Medicine	Terminología clínica integral, multilingüe y codificada	SNOMED Int.
ICD-10	International Classification of Diseases	Clasificación de enfermedades	WHO
ICF	International Classification of Functioning, Disability and Health	Clasificación de función, discapacidad y salud	WHO
ICPC	International Classification of Primary Care	Clasificación de atención primaria	WONCA
HL7 CDA	Health Level 7 – Clinical Document architecture	Codificación, estructura y semántica de documentos clínicos	HL7
FHIR	Fast Healthcare Interoperable Resources	Intercambio de datos de salud	HL7
13606	Norma CEN/ISO EN13606	Interoperabilidad semántica en la comunicación de la Historia Clínica Electrónica	ISO
DICOM	Digital Imaging and Communication On Medicine	Imagen médica y datos asociados	NEMA
LOINC	Logical Observation Identifiers Names and Codes	Terminología para pruebas y resultados de laboratorio	Regenstrief Institute
Continua	Continua Design Guidelines	Recogida de datos dispositivos personales de salud	PCH Alliance
MESH	Medical Subject Headings	Indexar literatura biomédica	NLM-NIH
GO	Gene Ontology	Función de los genes	GO Consortium
UMLS	Unified Medical Language System	Meta-tesauro que incluye múltiples vocabularios y terminologías médicas	NLM-NIH
HPO	Human Phenotype Ontology	Anormalidades fenotípicas en enfermedades humanas	Monarch initiative
ORDO	Orphanet Rare Disease Ontology	Vocabulario estructurado para enfermedades raras	Orphanet and European Bioinformatics Institute
CHEAR	Children's Health Exposure Analysis Resource	Investigación sobre el efecto de las exposiciones ambientales sobre la salud de los niños	NIEHS-NIH

Integración

La integración o *data fusion* es un proceso que consiste en, siempre que se hayan anotado los datos previamente, agregar y complementar los datos procedentes de diferentes fuentes con el objetivo de producir información más consistente, precisa y útil que la que proporciona cualquier fuente de datos de manera individual.[22] Para poder llevar a cabo esta integración es necesario que se realice una digitalización y estructuración de la información, donde juegan un importante papel las iniciativas de procesamiento del lenguaje natural y de traducción automática, que permitirán aumentar la disponibilidad de los datos al convertirlos en componentes interoperables.

Los datos integrados, además de la información de la historia clínica o datos moleculares, comienzan a incluir datos que provienen de dispositivos de monitorización (móvil, reloj o dispositivos creados especialmente para

[d]El internet de las cosas, o IoT por sus siglas en inglés, puede definirse como la agrupación e interconexión de dispositivos y objetos a través de una red (bien sea privada o Internet, la red de redes), dónde todos ellos podrían ser visibles e interaccionar.[53]

obtener datos de salud). Por tanto, el conocido como "internet de las cosas"[d] aplicado al *Big Data* en salud será un factor relevante ya que permitirá compartir información de la interacción de diferentes dispositivos entre médicos y pacientes.

Análisis

El último paso sería, una vez generadas las bases de datos integradas, el análisis de los datos, ya sea de forma masiva en grandes conjuntos de datos procedentes de estudios de investigación para la obtención de patrones o bien, de los datos de cada individuo/paciente en el ámbito de la investigación o la práctica clínica.

La principal diferencia en el análisis de los datos entre el área de investigación y el área asistencial radica en el propósito del análisis. En el caso de la investigación el objetivo final es la búsqueda de patrones complejos en los datos que permitirán definir nuevas hipótesis, encontrar causalidad, desarrollar algoritmos predictivos, etc. Sin embargo, en el caso de la asistencia médica, el objetivo es dar respuesta a las necesidades clínicas de un individuo concreto y orientar al profesional sanitario en la toma de decisiones (en cuanto al diagnóstico, tratamiento, etc.). Ciertamente, la distancia entre ambas aproximaciones es cada vez más pequeña tanto en términos prácticos como en la propia interacción entre los actores en las dos áreas.

En el campo de la salud, existen múltiples herramientas para el análisis de los datos que van desde la estadística clásica, pasando por la minería de datos, hasta las técnicas de Inteligencia Artificial (IA) más modernas. Dentro del campo de la IA, por su aplicabilidad en el campo de la salud, cabe destacar lo que se conoce como aprendizaje automático o *machine learning* (ML) que utiliza las técnicas estadísticas y los algoritmos computacionales para proporcionar a los ordenadores la capacidad de "aprender", es decir, mejorar sus resultados en una tarea específica "aprendiendo" tras procesar datos en suficiente cantidad y sin unas instrucciones explícitas externas (y por tanto potencialmente sesgadas) proporcionadas por el programador.[11] Respecto a la minería de datos o *datamining*, consiste en un conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto.

Los datos en salud en el ámbito de la Medicina Personalizada de Precisión presentan múltiples dimensiones que pueden abordarse mediante las técnicas de ML. Esta multidimensionalidad de los datos hace referencia a la gran cantidad de atributos o variables de las que es posible obtener información en cada individuo, dificultando las estrategias de análisis mediante estadística clásica de este tipo de datasets.[13] Hay dos formas de hacer frente a este problema de la multidimensionalidad. Una de ellas consiste en la "selección de dimensiones", es decir, se eliminan del estudio aquellas variables que no interesen para el estudio concreto que se realiza. La otra es la "extracción de dimensiones", que consiste en asociar dimensiones para crear nuevas variables que aglutinen información de múltiples variables. En este sentido, las técnicas de ML permiten, por ejemplo, analizar de manera simultánea diferentes subtipos de una enfermedad y diferentes subtipos de pacientes permitiendo desarrollar modelos de predicción en función de múltiples variables.

Resulta interesante destacar, por su gran potencial a la hora de diseñar la Medicina Personalizada de Precisión del futuro, el aprendizaje profundo o *deep learning*, que representa una variante del *machine learning* más compleja, sofisticada y autónoma. La principal diferencia entre el *deep learning* y los sistemas de ML radica en que, debido a su estructura, con más capas y unidades de proceso intermedias, los sistemas de *deep learning* no necesitan que el programador introduzca las características (*features*) que son necesarias para discriminar los datos (por ejemplo, elementos presentes en una imagen), sino que el sistema es capaz de identificarlos por sí mismo, en la fase de entrenamiento. Esta técnica reduce el margen de error y, por tanto, aumenta la precisión de las conclusiones. Hay que referirse en este punto a la necesidad de avanzar en lo que se conoce como ML interpretable, que trata de reducir los problemas prácticos y éticos (sesgos implícitos) asociados a la utilización de ML como una caja negra, sin capacidad para explicar por qué ha llegado a una conclusión a partir de los datos.

El procesamiento del lenguaje natural es otro campo de conocimiento de la IA que está avanzando rápidamente y cuya aplicación en el contexto de la medicina resulta fundamental dado que la mayoría de la información recogida por los profesionales en la HC está en formato de texto libre. Estas técnicas, más allá de contribuir a la anotación e integración de los datos, tienen como objetivo principal la transformación de manera automática del texto libre en información estructurada que complete o enriquezca los datos previos, generando nuevas capas

de conocimiento, de manera que se facilite la clasificación o fenotipado de los pacientes en distintos grupos en función del contenido escrito por los profesionales.

A pesar de que aún se ha de trabajar en perfeccionar y potenciar las herramientas y algoritmos de análisis con el objetivo de mejorar la capacidad de predicción, de simulación y de modelado, las herramientas de IA aplicadas al análisis de los datos en salud suponen una gran oportunidad para extraer conocimientos dirigidos a mejorar la salud de los pacientes y de la población en general. La interpretación de los datos y la mayor capacidad de inferencia por parte de estas herramientas se asocian a la necesidad de que los datos estén correctamente anotados y sean adecuados en cuanto a su calidad, cuyo nivel ha de aumentar de manera directamente proporcional a la complejidad de los datos. Como se desarrolla más adelante en este informe, la capacidad de generar modelos que permitan realizar simulaciones a distintos niveles hace posible la utilización de estas herramientas en los sistemas de soporte a la decisión clínica, en el descubrimiento de biomarcadores o en el reposicionamiento de fármacos. Por tanto, el uso conjunto de nuevas fuentes de datos y nuevas técnicas de ML está cambiando ya la práctica médica con la introducción de sistemas de apoyo a la decisión clínica que mejoran la atención y contribuyen a la sostenibilidad del Sistema Sanitario.

Sin embargo, no hay que dejar de lado el hecho de que, para generar un conocimiento valioso, el *big data* en salud debe provenir de datos individuales de alta calidad (se podría decir que no hay *"big data"* sin *"small data"*). La importancia del *small data* en el campo de la biomedicina hace necesario el desarrollo de herramientas estadísticas y de ML adaptadas a su análisis dentro del contexto de *big data* en salud.[23]

PRIVACIDAD Y SEGURIDAD DE LOS DATOS

A lo largo de todo el ciclo de la vida de los datos es fundamental tomar las medidas necesarias para mantener el anonimato de los individuos y asegurar la confidencialidad y la protección de los datos siguiendo la legislación europea GDPR (*General Data Protection Regulation*) aplicable a los proyectos de sanidad en los que se trabaja con datos clínicos de pacientes.

En este sentido, existen numerosas estrategias destinadas a garantizar la seguridad y privacidad de datos como son las herramientas de criptografía o el aislamiento de *firewalls* [e], ampliamente implantadas en sectores como la banca y potencialmente aplicables a los datos de salud. Recientemente ha surgido la metodología del *blockchain* que se basa en la estructuración de conjuntos de datos en paquetes de datos cifrados. Estos paquetes son como eslabones que están enlazados entre sí a través de un sistema de códigos identificadores, formando una "cadena". Si se produce una variación de la información de cualquier bloque, cambia el identificador del bloque y el eslabón ya no encajará en la cadena, de manera que puede ser detectada por el resto de los componentes de la misma. A pesar de que este tipo de metodología está en auge, su implementación resulta controvertida en el ámbito de salud.

El conflicto surgido entre la necesidad de compartir datos y el mantenimiento de la seguridad de los mismos ha estimulado la investigación activa de nuevas implementaciones criptográficas que sirven tanto para la atención médica del paciente como para proteger la información genómica sensible y la privacidad del paciente.[24]

Más allá de la tecnología usada para mantener la privacidad de los datos, de la potencia de las bases de datos para almacenar la información en tiempo real y de los recursos de supercomputación disponibles, el desafío real consiste en cómo conectar estos recursos. El paradigma actual es acercar la computación a los datos, dada la obvia dificultad técnica para mover grandes cantidades de datos hasta los recursos de computación. Con el objetivo de facilitar este acercamiento, garantizando la privacidad y seguridad de los datos, el "Dosier del Paciente" y "Continuo de Computación" son dos conceptos emergentes claves.[25] Por un lado, el "Dosier del Paciente" hace referencia a la creación de un documento virtual descentralizado que contiene toda la información necesaria para responder a una pregunta clínica, incluyendo localizar y gestionar el acceso a toda la información distribuida en distintos repositorios y la interacción con el "Continuo de Computación" para la gestión de los análisis necesarios para cada tipo de dato. Por otro lado, el "Continuo de Computación" hace referencia a la creación de sistemas inteligentes capaces de asignar a cada tipo de datos el recurso computacional adecuado desde dispositivos personales a recursos de supercomputación, gestionando en tiempo real los permisos necesarios en cada sistema.

[e] La criptografía es una forma de codificación de la información mediante códigos alfanuméricos; mientras que la estrategia de firewalls consiste en impedir el acceso o la salida de paquetes de datos en una red que no cumplan una serie de normas de seguridad. La supercomputación consiste en la realización en paralelo de cálculos complejos con grandes cantidades de datos.

INICIATIVAS EN MANEJO DE DATOS EN MEDICINA PERSONALIZADA DE PRECISIÓN

Dada la importancia de los datos en salud en el avance de la Medicina Personalizada de Precisión, existen numerosas iniciativas de diferente índole, tanto a nivel nacional como internacional, cuyo objetivo general es la recopilación de este tipo de datos para su posterior análisis, de manera que ofrezcan nuevos conocimientos aplicables en el campo de la salud.

A continuación, se recogen ejemplos de iniciativas* que se están llevando a cabo en el manejo de datos en Medicina Personalizada de Precisión:

Tabla 3. Ejemplos de iniciativas en el manejo de datos en Medicina Personalizada de Precisión.

INICIATIVA	ÁMBITO	ORGANISMO IMPULSOR	BREVE DESCRIPCIÓN
GA4GH (Global Alliance for Genomics and Health)	Mundial	50 grupos internacionales	Su objetivo es acelerar el progreso en la investigación genómica y la salud humana mediante el diseño de un marco común de estándares y enfoques armonizados para el intercambio de datos genómicos y relacionados con la salud que sean efectivos y responsables.
ICGC ARGO Project	Mundial	ICGC	ICGC se estableció para desentrañar los cambios genómicos presentes en muchas formas de cáncer. En su segunda fase, (Pan Analysis of Whole Genomes- PCAWG), definió similitudes y diferencias entre miles de genomas completos. La tercera fase, ahora en marcha, tiene como objetivo producir marcadores de interés clínico para al menos 100,000 pacientes de cáncer, aunando genómica y datos clínicos de alta calidad.
All of Us	EE.UU.	Instituto Nacional de Salud (NIH) Americano	Iniciativa dirigida a recopilar datos de más de un millón personas con el objetivo de conocer las diferencias individuales en el estilo de vida, el medio ambiente y la biología, para brindar una medicina de precisión.
eMERGE (Electronic Medical Records and Genomics)	EE.UU.	Instituto Nacional de Salud (NIH) Americano	El objetivo es desarrollar, difundir y aplicar enfoques en investigación para la combinación de repositorios biológicos con los registros médicos (HCD) y así contribuir al descubrimiento genómico y a la investigación de la implementación de la medicina genómica.

* No se incluyen iniciativas de tipo comercial privado

Tabla 3. Ejemplos de iniciativas en el manejo de datos en Medicina Personalizada de Precisión.

INICIATIVA	ÁMBITO	ORGANISMO IMPULSOR	BREVE DESCRIPCIÓN
Proyecto InSite	Europa	TriNetX Company	Plataforma que permite la reutilización de manera fiable de las HCD orientada a la investigación, facilitando la colaboración entre personal clínico e investigadores, con el objetivo de maximizar los resultados de la investigación clínica.
1 million genomes	Europa	Comisión Europea	Iniciativa de 20 estados miembros de la Unión Europea y Noruega, cuyo objetivo es recopilar en una base de datos genómicos la secuencia de al menos 1 millón de genomas antes del 2020 de manera que estén disponibles para la investigación que dará lugar a una medicina personalizada.
Biobank	Reino Unido	National Health Service (NHS)	Repositorio nacional e internacional de muestras biológicas de 500000 participantes con el objetivo de mejorar la prevención, el diagnóstico y el tratamiento de un amplio rango de enfermedades.
100000 genomes	Reino Unido	National Health Service (NHS)	Secuenciación genómica a gran escala para el estudio del cáncer y de enfermedades raras.
ELIXIR	Europa	ESFRI (Foro Europeo de Estrategias de Investigación en Infrastructuras)	Organización intergubernamental que reúne recursos de ciencias de la vida de toda Europa cuyo objetivo es coordinar estos recursos para que formen una infraestructura única en la que los científicos encuentren y compartan datos, intercambien experiencias y se pongan de acuerdo sobre las mejores prácticas.
EOSC (European Open Science Cloud)	Europa	Comisión Europea	El objetivo es iniciar la federación de las infraestructuras de datos científicos existentes para que el acceso a los datos científicos sea más fácil y más eficiente.
HARMONY	Europa / España*	Innovative Medicines Initiative (IMI) con el apoyo de Horizonte 2020 de la Unión Europea y la Federación Europea de Industrias y Asociaciones Farmacéuticas (EFPIA)	Se trata de una red de excelencia europea público-privada cuya misión es desbloquear y difundir valiosos conocimientos sobre tumores malignos hematológicos entre un gran número de partes interesadas, con el objetivo de aprovechar y extraer Big Data para acelerar el desarrollo de tratamientos mejorados para pacientes y estrategias de tratamiento más efectivas. * Proyecto coordinado por el Instituto de Investigaciones Biomédicas de Salamanca (IBSAL).
Plan de Impulso de las Tecnologías del Lenguaje	España	Secretaría de Estado para el Avance Digital	Entre otros, tiene como objetivos fomentar el desarrollo del procesamiento del lenguaje natural y la traducción automática. En el campo de la salud esto permitirá entre otras cosas analizar la información contenida en las HC independientemente del idioma en el que esté registrada lo que, en un futuro, supondrá un avance a la hora de intercambiar datos a nivel mundial en el campo de la Medicina Personalizada de Precisión.
NAGEN 1000 (Proyecto Genoma 1000 Navarra)	España	Centro de investigación biomédica Navarrabiomed	El objetivo es trasladar el uso de la tecnología más vanguardista de análisis de genoma humano completo a la red sanitaria pública de Navarra. Para ello, se está acometiendo el estudio de 1.000 genomas de pacientes y sus familiares con enfermedades raras y algunos tipos de cáncer del Servicio Navarro de Salud-Osasunbidea (SNS-O).

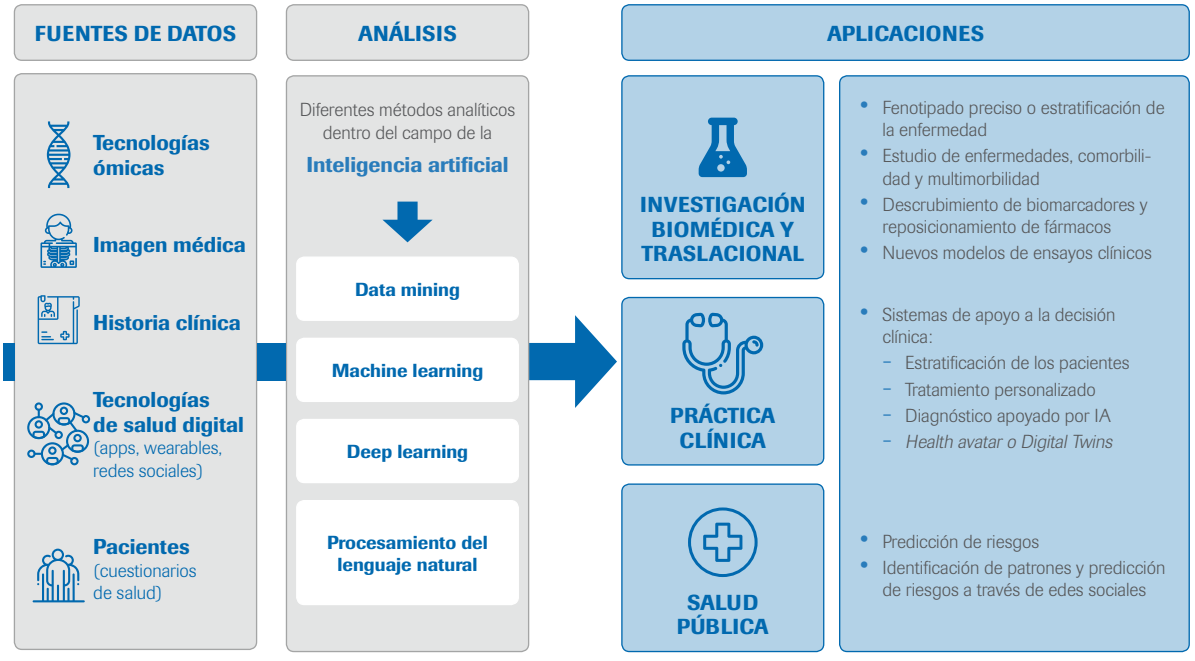
APLICACIONES EN LA MEDICINA DEL FUTURO

El estudio exhaustivo de datos en salud se traduce en un amplio abanico de nuevas oportunidades y aplicaciones en todos los campos de la medicina desde la investigación biomédica y traslacional hasta la práctica clínica y salud pública.

Estas nuevas aproximaciones mejorarán la prevención, el diagnóstico y el tratamiento contribuyendo a mantener los estados de salud y a evitar el desarrollo de determinadas enfermedades, en áreas como el cáncer o las enfermedades raras,[26] pero también en otras como psiquiatría, enfermedades respiratorias o cardiología.

Ante la imposibilidad de tratar en profundidad todas estas aplicaciones, a continuación, se desarrollan algunos de los ejemplos más relevantes en Medicina Personalizada de Precisión.

Figura 5. Visión general de las aplicaciones más relevantes de los datos en salud



APLICACIONES EN INVESTIGACIÓN BIOMÉDICA Y TRASLACIONAL

Fenotipado preciso o estratificación de la enfermedad

Una de las aplicaciones del conocimiento derivado de los datos en salud es la estratificación de enfermedades basada en datos que va más allá del enfoque clásico sustentado en "signos y síntomas" e incluye la identificación de "rasgos tratables"[f], de manera que estos subgrupos de enfermedad pueden abordarse con mayor exactitud gracias a una mejor comprensión de las causas de la patología y la subsiguiente caracterización fenotípica más precisa.[27]

Un factor clave que hace posible una correcta estratificación es la descripción de subgrupos de patologías basados en datos genómicos, transcriptómicos, epigenómicos y clínicos[28] mediante la aplicación de herramientas de análisis y algoritmos a los conjuntos de datos correspondientes. Un ejemplo de ello, es el desarrollo de herramientas que de manera automática o semiautomática realizan el fenotipado. Estas herramientas se basan en algoritmos para, por ejemplo, interpretar el lenguaje natural, como sucede en el marco del proyecto estadounidense eMERGE[29] o en el Plan Nacional de Impulso a las Tecnologías del Lenguaje.[30] Las nuevas aproximaciones de IA representan una segunda revolución en la estratificación de enfermedades.

Estudio de enfermedades, comorbilidad y multimorbilidad.

El análisis de los grandes data sets existentes en salud generados con estas tecnologías permiten plantear hipótesis que establecen nuevas asociaciones con nuevas variables no contempladas en un primer momento, o en nuevos patrones identificados al analizar estos conjuntos de datos, permitiendo llevar a cabo proyectos de investigación que siguen una aproximación guiada por los propios datos (*data-driven research*).[31]

La inferencia de conocimiento a partir de la combinación de datos genómicos, fenómicos y exposómicos es especialmente relevante en este caso, ya que sólo con una aproximación holística será posible comprender los mecanismos intrínsecos y extrínsecos subyacentes a las enfermedades multifactoriales, como por ejemplo las enfermedades neurodegenerativas,[32] las enfermedades cardiovasculares[14] o el cáncer.[33] Las continuas mejoras en el manejo de toda la información disponible permitirán establecer las relaciones entre patologías y describir los patrones de asociación y las comorbilidades, mejorando el tratamiento de los pacientes con multimorbilidades.

Descubrimiento de biomarcadores y reposicionamiento de fármacos

De la mano de la estratificación de pacientes, y a través del análisis de datos moleculares es posible el descubrimiento de nuevos biomarcadores que sirvan para el diagnóstico.[27] Para ello, es necesario el procesamiento de los datos primarios obtenidos por las diferentes técnicas existentes, permitiendo la extracción de perfiles moleculares y la identificación de moléculas sobre o infrarrepresentadas de manera estadísticamente significativa[34] que pueden ser utilizadas como biomarcadores.

En cuanto al reposicionamiento de fármacos consiste en la identificación y el desarrollo de nuevos usos para fármacos existentes a través de herramientas de integración y análisis, que va de la mano de una reducción en los costes respecto al descubrimiento y desarrollo de novo de los mismos.[28]

Por ejemplo, de cara a inferir otros usos para un determinado fármaco los diferentes métodos computacionales, tanto derivados del estudio de redes (biología de sistemas) como de la IA (ML) pueden seguir una aproximación basada en similitudes con otros fármacos (*drug-based*), en el tipo de enfermedad al que van dirigidos (*disease-based*) o en la diana sobre la que actúan (*target-based*); o pueden basarse en la simulación del acoplamiento molecular, en la que por modelado 3D de las dianas es posible predecir el sitio de unión con fármacos.[28] Independientemente de los métodos utilizados, estas estrategias permitirán la identificación de nuevas indicaciones para los fármacos ya desarrollados de una manera más rápida.

Nuevos modelos de ensayos clínicos

La disponibilidad de grandes conjuntos de datos proporciona información que permite generar hipótesis que

[f]Es decir, aquellos rasgos o características de la patología que presenta cada individuo.

se incorporen en el diseño de nuevos modelos de ensayos clínicos. Por ejemplo, marcadores como la presencia de ciertas mutaciones asociadas significativamente a algunos pacientes con distinto tipos de cáncer, lleva al desarrollo de un nuevo tipo de ensayo clínico, *basket trial*, donde participantes con distinto tipo de cáncer se analizan en un único ensayo clínico usando como criterio de inclusión la presencia de marcadores genómicos susceptibles de ser atacados por un determinado fármaco, o *umbrella trials*, donde se trata un solo tipo de tumor adaptando el protocolo del ensayo a los marcadores presentes en distintos participantes en el ensayo.[35-37]

Todos estos nuevos tipos de ensayos plantean a su vez una nueva problemática de análisis tanto para la incorporación de las hipótesis generadas a partir del análisis de grandes cantidades de datos al diseño de los ensayos con un número necesariamente pequeño de participantes, como en el análisis de los resultados, que pueden ser necesario realizar para la toma de decisiones durante el ensayo.[38,39] Este es el caso de los conocidos como *adaptive trials*, donde el protocolo del ensayo, como pueden ser la dosis que se administra, el tamaño de la muestra, el fármaco en estudio, los criterios de selección del paciente o la combinación de alguno de estos, se adapta durante su ejecución a la respuesta de los participantes en forma de expresión de distintos marcadores o aparición de efectos secundarios.

Un caso extremo son los llamados ensayos clínicos *N-of-1*, diseñados para determinar la terapia óptima para cada persona, alternando la toma de medicamentos con periodos de descanso, de manera que como resultado final se obtengan series de datos sobre el resultado del tratamiento en un solo individuo y, a partir de ellos, se pueda alcanzar una terapia individualizada y personalizada.

A pesar de que todas estas nuevas modalidades de ensayo clínico parece contrapuestas al concepto de *big data*, lo cierto es que no solo su diseño proviene de los resultados de *big data*, sino que además se pueden integrar múltiples ensayos de este tipo[40] de cara a realizar un meta-análisis de los datos derivados de los mismos, y así poder obtener conclusiones generales basadas en las características de pacientes en distintos ensayos.

APLICACIONES EN PRÁCTICA CLÍNICA

A pesar de que aún no se cuenta con la capacidad para procesar e interpretar la gran cantidad de datos de los que se dispone, el análisis de los mismos podría ayudar en la toma de decisiones por parte de los profesionales clínicos, pero también por parte de los gestores de los centros sanitarios, mejorando así el servicio a los pacientes. Y no sólo eso, sino que contribuirían al desarrollo de modelos predictivos que permitirán anticipar las necesidades sanitarias.

Sistemas de Apoyo a la Decisión Clínica (SADCs)

Los Sistemas de Apoyo a la Decisión Clínica (SADCs) son herramientas que buscan mejorar la eficiencia y la calidad de la asistencia clínica, ayudando a los profesionales de la salud en el proceso de toma de decisiones, ya que permiten el acceso a la información referente a salud del paciente, facilitan información previa al diagnóstico y validan y corrigen los datos proporcionados,[41] aunque todavía se encuentran en etapas iniciales de desarrollo e implantación.

Estos sistemas, que pueden encontrarse en distinto formato, incluyen, por ejemplo, analizadores automatizados de imágenes médicas, herramientas de *text mining* de textos médicos,[13] guías de práctica clínica o clinical pathways, sistemas de alerta, recordatorios, etc.

Hasta el momento los SADCs han ayudado a mejorar la adherencia a protocolos, en recordatorios para la toma de fármacos, en la mejora del *screening* de pacientes y en la predicción de reingresos en hospitales. Estos sistemas tienen múltiples aplicaciones en la práctica clínica, algunas de las cuales se desarrollan a continuación:

- Una de las principales aplicaciones de estos sistemas es la posibilidad de estratificar a los pacientes, es decir, situar al paciente en el subgrupo que más se ajuste en función de sus características biológicas, de la progresión de su enfermedad o de la respuesta al tratamiento, en lo que se llama similaridad de pacientes[42] (*patient similarity*[9]). A partir de todos los datos de salud de un paciente, que pueden provenir de distintas fuentes, los SA-

[9] *Patient similarity:* consiste en agrupar o clasificar a los pacientes según su grado de similitud en diversas características, incluidos los perfiles genómicos. Se trata de una estrategia análoga al diagnóstico médico estándar, pero con un mayor rendimiento, y que preserva la privacidad del paciente.

DCs se encargarán, de manera automatizada, de consultar bases de datos de población para posicionar al paciente dentro de un subgrupo de población y dentro de un subtipo de patología. Este abordaje permitirá un diagnóstico personalizado y la elección del tratamiento más adecuado en función de las evidencias generadas a partir del análisis de grandes cantidades de datos.

- La posibilidad de integrar la información tanto del genoma, como del fenoma y el exposoma, permitirá a los clínicos disponer de una visión más completa del paciente y en base a esta información tomar decisiones dirigidas a la personalización de los tratamientos atendiendo no sólo al perfil farmacogenómico del paciente, sino incluyendo otros factores en la ecuación. Por ejemplo, el análisis molecular se está convirtiendo en parte del protocolo rutinario en la asistencia clínica, permitiendo adecuar los tratamientos a las características de cada paciente lo que mejora las expectativas de supervivencia y reduce los efectos secundarios de los medicamentos.[43]

  Además, la integración de datos obtenidos a partir de dispositivos de monitorización, puede ser útil para introducir cambios en los hábitos de vida de los individuos, y en concreto en el caso de los pacientes, para ajustar los tratamientos a su estilo de vida.

- A partir de los datos derivados de las imágenes médicas, en el campo de la radiología o de la dermatología, se han logrado avances relevantes aplicando la IA en la identificación de anomalías o signos radiológicos sugestivos de determinados procesos patológicos. Además, permiten la generación de bases de datos en las que automáticamente sea posible comparar una imagen de un paciente con miles de imágenes previas, ayudando al médico a realizar el diagnóstico.[10]

- Por último, y con aplicación en la práctica clínica de la medicina del futuro con el objetivo de predecir riesgos y diseñar estrategias terapéuticas personalizadas, se está avanzando en el desarrollo de los denominados Avatares de Salud (*Health Avatar*) o Gemelos digitales (*Digital Twins*). Estas herramientas son un paso más en relación con los SADC y consisten en una representación virtual de cada individuo en la que se almacena toda su información referente a salud, que permite aplicar una serie de algoritmos y herramientas médicas de análisis de cara a predecir el futuro estado de salud del individuo[44] en función de diferentes variables que pueden ir modificándose dado la posibilidad de ensayar diferentes escenarios y predecir las consecuencias derivadas de la toma de una decisión u otra.

En la atención sanitaria del futuro, teniendo en cuenta todo lo anteriormente expuesto, toda la información relativa a la salud del individuo estará disponible a través de un sistema de información integrado con acceso y capacidad de análisis de todos los datos de salud de cada individuo en el punto de asistencia sanitaria. De esta manera, probablemente el clínico ante la visita de un paciente dispondrá de un SADC que le permita clasificar el problema clínico del paciente correctamente (fenotipado) asociándole una entidad clínica concreta. Posteriormente, el sistema tendrá que ser capaz de caracterizar e identificar los parámetros relevantes del perfil molecular y de riesgo ambiental del paciente. Una vez hecho esto, el propio sistema consultará las grandes bases de datos de investigación y conocimiento biomédico, puesto que puede haber aspectos poblacionales que han de tenerse en cuenta como, por ejemplo, que la incidencia de la enfermedad sea distinta en función del área geográfica. En función del perfil clínico, del perfil molecular, del perfil de riesgo debido a exposición a factores ambientales y del conocimiento biomédico, el sistema podrá proporcionar recomendaciones terapéuticas personalizadas y adaptadas a la entidad clínica y molecular propia del paciente. Además, estos sistemas incorporarán las preferencias del paciente y sus hábitos de vida transformándose en SADC compartida. Este modelo permitirá que todo el conocimiento derivado de la investigación basada en el análisis de grandes datos en salud, mediante el uso de herramientas de IA, esté accesible en el punto de atención sanitaria para su aplicación en la práctica clínica, reduciendo la carga cognitiva del profesional sanitario de manera que el médico no tenga que memorizar e interpretar toda la información disponible sino que le permita tomar decisiones soportadas por el conocimiento

y la evidencia, dando lugar a un nuevo escenario en la atención sanitaria personalizada de precisión.

APLICACIONES EN SALUD PÚBLICA

A nivel poblacional, el uso de los datos guía a la Medicina Personalizada de Precisión hacia el diseño y la implementación de estrategias preventivas que reduzcan la incidencia de enfermedades, evitando el desarrollo de determinados problemas de salud y/o, en caso de haberse manifestado una enfermedad determinada, disminuyan su morbi-mortalidad minimizando su impacto sobre la vida activa del paciente en un grupo de individuos. Se trata de diseñar una salud pública "inteligente", teniendo en consideración los datos genómicos, ambientales, de conducta, socio-económicos, epidemiológicos, etc., de manera que permitan caracterizar mejor a los individuos y diseñar estrategias que irán dirigidas a grupos similares en los que realmente serán efectivas. Además, el conocimiento derivado de estos datos permitirá no solo que los programas de prevención sean más específicos, sino que sean más rentables desde el punto de vista económico contribuyendo a la sostenibilidad del Sistema Sanitario.

Predicción de riesgos

Actualmente, muchas herramientas de análisis de *big data* basadas en IA y concretamente en *machine learning* están dirigidas a la generación de modelos predictivos para detectar subgrupos de población con alto riesgo de padecer determinadas enfermedades, puesto que la identificación precisa y rápida de estos individuos podría facilitar una atención más eficaz.[14]

Existen numerosos factores predictivos de enfermedades, como puede ser el marcador HER-2 en el caso de cáncer de mama, pero lo ideal sería identificar factores de riesgo con la suficiente antelación de manera que sea posible diseñar intervenciones dirigidas a disminuir el riesgo de cada individuo.[44] Por ello, la información derivada de la integración de los datos que proporcionan los *wearables* y *apps*, por ejemplo sobre la dieta, el consumo de alcohol, el tabaquismo o la práctica de ejercicio físico, con datos genómicos pueden servir no sólo para el establecimiento de tratamientos personalizados, sino que a partir del conocimiento derivado de su análisis podrían surgir nuevas oportunidades de cara a identificar patrones combinados de factores de riego para determinados subgrupos de individuos en relación con determinadas enfermedades.[45] Es importante que estos modelos de predicción del riesgo sean dinámicos e incluyan factores de riesgo modificables, permitiendo modificar las probabilidades de aparición de la enfermedad en el individuo.[44]

Desde una perspectiva más global, la integración de los datos en salud junto con los datos de exposición derivados de, por ejemplo, sensores medioambientales, pondrá de manifiesto la necesidad de diseñar y establecer medidas preventivas no solo desde una perspectiva de hábitos saludables a nivel individual, sino que posiblemente serán necesarias medidas de gestión medioambientales. Un ejemplo de ello es la relación existente entre la salud de los individuos a nivel respiratorio y la densidad de tráfico en las grandes ciudades, en este caso la solución a largo plazo de estas patologías respiratorias no pasa por la administración de fármacos específicos, sino por medidas de prevención supeditadas a la adopción de medidas de gestión medioambiental de las ciudades. Por tanto, los datos abren una puerta hacia la gestión del bienestar planetario a través de la integración de los datos de salud individuales, los datos de salud pública y los datos medioambientales.

Identificación de patrones y predicción de riesgos a través de redes sociales

Dentro del entorno de la salud, las redes sociales han emergido como una herramienta para compartir información de salud y conectar a enfermos, familiares o profesionales de la salud.[10] Además, en combinación con los datos generados por los pacientes con sus *smartphones* y los *wearables*, las redes sociales permiten conocer información sobre el entorno social y los comportamientos en salud que podrían considerarse en el contexto de la Medicina Personalizada de Precisión y la salud pública de precisión, para crear "fenotipos y expotipos digitales" de enfermedad.[44] Por ejemplo, las imágenes de Instagram se han utilizado para establecer hábitos dietéticos[46] o para identificar marcadores predictivos de depresión.[47]

En esta línea, las terapias digitales surgen como herramientas digitales, como pueden ser apps, que pueden ayudar a los pacientes a gestionar su enfermedad,

por ejemplo, al hacerles recordatorios sobre la toma o la dosis de medicamentos. Pero también pueden emplearse como alternativa al tratamiento con medicamentos, ya que a través de estímulos sensoriales emitidos desde dispositivos electrónicos pueden ayudar a gestionar el insomnio o la depresión.[48] Algunos ejemplos de estas terapias de reemplazo clínicamente aprobadas son las apps diseñadas para el tratamiento de adicciones,[49] o los videojuegos como tratamiento para afecciones neurológicas y psiquiátricas (como el Alzheimer, el TDAH o la depresión).[48] En el sentido opuesto, también se ha acuñado recientemente el término exposoma digital,[50] para hacer referencia al impacto que está teniendo sobre la salud de las personas el uso continuado y excesivo de sistemas electrónicos (redes sociales, videojuegos, etc.).

Por último, señalar que se están desarrollando y perfeccionando otras aplicaciones que, a través del análisis de estos datos, tienen como objetivo la detección temprana de posibles brotes epidémicos o condiciones patológicas,[10] como es el caso de GoogleFluTrends[51] que supuso una primera aproximación para la predicción de brotes de gripe a partir del análisis y la geolocalización de búsquedas relacionadas con esta infección.[14,52]

RETOS

Como ha quedado de manifiesto a lo largo del informe, la enorme cantidad de datos que se generan de manera continua en el ámbito de la biomedicina precisa de tecnologías y herramientas dirigidas a gestionar y extraer información útil y conocimientos aplicables en el campo de la salud. Este fenómeno está relacionado, entre otros, con la extensión de la utilización de las tecnologías ómicas, especialmente de la genómica, tanto en el campo de la investigación como en la práctica clínica, con la implantación de la HCD, y con el hecho de que la mayoría de datos que se generan en este campo no son estructurados.[10]

La extensión en la aplicación de los datos en el campo de la Medicina Personalizada de Precisión como vía para su completo desarrollo e implantación todavía tiene que afrontar diversos retos relacionados, entre otros, con el procesamiento de los datos, la formación, educación y difusión de los avances producidos en este campo, así como de carácter organizativo, algunos de cuales, los más relevantes, se enumeran a continuación.

RETOS RELACIONADOS CON EL PROCESAMIENTO DE LOS DATOS

- Escasa adopción de criterios de estandarización de los datos que faciliten el intercambio de datos y la interoperabilidad entre sistemas de información.
- Existencia de limitaciones en el cumplimiento de los criterios FAIR, especialmente en relación con la capacidad de localizar los datos, y, en consecuencia, desarrollo de las infraestructuras necesarias.
- Ausencia de procesos que permitan, a posteriori, establecer relaciones entre los datos disponibles y la interoperabilidad para trasladar la información de unos sistemas a otros (falta de expresividad de la información asociada a los datos).
- Necesidad de sistemas para evaluar y validar los sistemas de gestión y análisis de datos de forma sistemática, abierta y continua, facilitando tanto el trabajo de los desarrolladores de sistemas como la información necesaria para los usuarios.
- Necesidad de evaluar y validar formalmente los resultados de los proyectos de investigación o sistemas implementados (muchas veces es un proceso costoso o poco factible y además puede existir una fuente de variabilidad debida a las características de los propios centros, como puede ser la formación del personal, los pacientes que acuden al centro, etc.).
- Dificultades en la aplicación de un marco legal (GDPR) que permita compartir datos entre instituciones garantizando la seguridad y privacidad de los datos, así como la utilización de sistemas de claves de seguridad (como por ejemplo la Infraestructura de Autenticación y Autorización que emplea ELIXIR, que permite a los investigadores acceder empleando las credenciales de la organización, comunidad o entidad comercial a la que pertenezcan), y garantice los procesos de anonimización (p. ej. normativa HIPAA[h] de Estados Unidos) como paso necesario para facilitar el acceso y la reutilización de los datos.
- Mayor concienciación sobre la necesidad de diseñar estrategias que promuevan la publicación y accesibilidad de los datos de salud obtenidos tanto en proyectos realizados con

[h]HIPAA: Health Insurance Portability and Accountability Act

financiación pública como los obtenidos en empresas privadas fomentando el hecho de que sean reutilizables de manera gratuita.

RETOS FORMATIVOS, EDUCATIVOS Y DE DIFUSIÓN

- Necesidad de recursos suficientes destinados a la formación y capacitación del personal clínico, científico y de gestión de manera que puedan ser operativos a medio plazo, por ejemplo, a través de iniciativas de formación en perfiles múltiples. Esta formación permitirá asegurar que todos estos profesionales adquieran las competencias necesarias para la correcta interpretación y manejo de datos de salud en Medicina Personalizada de Precisión y puedan conocer que pueden esperar y hasta donde pueden llegar las herramientas de procesamiento y análisis de datos.
- Falta de capacitación en estos temas (procesamiento digital de información) de los clínicos y gestores que participan de manera ordinaria en los comités de los hospitales y en los órganos de decisión de los mismos, dado que participarán en la toma de decisiones relacionadas con la instauración de servicios de informática en los hospitales, entre otras.
- Falta de concienciación en el ámbito científico y asistencial de la importancia de recoger los datos de una manera adecuada y con calidad. Se percibe cierta reticencia al intercambio de datos entre centros/ investigadores/ clínicos por lo que parece necesario reiterar la importancia de la accesibilidad de los datos.
- Garantizar la difusión responsable y rigurosa de la información relacionada con temas como la IA de manera que se evite sobredimensionar los logros alcanzados y alcanzables por estas tecnologías. Parece necesario hacer un esfuerzo dirigido a comunicar adecuadamente a la sociedad los beneficios que se pueden obtener con los datos asegurando la protección de la identidad y de los datos personales.
- Falta de visibilidad del papel, relevancia e impacto de la (bio)informática aplicada a la práctica clínica.

RETOS ORGANIZATIVOS

- Falta de financiación destinada a dotar a los centros de estructuras computacionales suficientemente potentes, personal entrenado y mecanismos adecuados de mantenimiento y actualización de software que permitan asegurar la sostenibilidad de los datos y sus repositorios.
- Dar soporte a los procesos encaminados a la interoperabilidad de datos clínicos (anotación, metadatos, etc.).
- Generar estructuras organizativas que permitan compartir grandes cantidades datos confidenciales, acceder a las bases de datos de referencia y utilizar las herramientas necesarias para el análisis en los recursos computacionales adecuados, todo ello en entornos que faciliten el acceso a los usuarios finales, bien biomédicos o clínicos.
- Insertar los desarrollos locales en el ámbito internacional, incluyendo los recursos e infraestructuras necesarias para la participación en consorcios internacionales de medicina personalizada.
- Necesidad de incorporar perfiles profesionales en el área de datos biomédicos, que actualmente no se encuentran suficientemente integrados en las plantillas de los hospitales, y carecen de un papel formal en la toma de decisiones relacionadas con la salud de los pacientes y la asistencia sanitaria.

CONCLUSIONES Y RECOMENDACIONES

Los avances acontecidos en torno a la generación, el almacenamiento, el procesamiento y el análisis de los datos en salud suponen un impulso para el desarrollo completo y la traslación a la práctica clínica de la Medicina Personalizada de Precisión y en gran medida están contribuyendo a configurar la medicina del futuro. Sin embargo, la total incorporación de estas nuevas estrategias basadas en la información obtenida a partir de los diversos datos ómicos y de salud está sujeta a retos relacionados con el propio procesamiento de los datos, con la formación de los profesionales, con la difusión de la importancia del uso de los datos a la población y los decisores, así como retos de carácter organizativo que tendrán que solventarse.

RECOMENDACIONES:

- Aunar y coordinar esfuerzos dirigidos a la estandarización de los diferentes sistemas informáticos utilizados en todos los centros del Sistema Sanitario atendiendo a las características FAIR de manera que sea posible compartir los datos en salud entre los distintos centros a lo largo de todo el territorio nacional, como primer paso en la implantación de una HCD universal y estandarizada, utilizable como fuente de datos para la investigación y la no menos importante sistematización de la inclusión de los datos ómicos en la HCD.
- Incidir desde el ámbito científico y clínico en la importancia de la recogida de datos genómicos en los ensayos clínicos como paso indispensable para alcanzar una Medicina Personalizada de Precisión efectiva.
- Promover y apoyar la formación mixta entre los diferentes perfiles profesionales implicados en el campo de la atención sanitaria, de la biomedicina y de la salud en general. Esta formación mixta permitirá la adquisición de competencias, por un lado, en cuanto a la extracción de información biomédica a partir de los datos y la gestión de esta información resaltando la importancia de la fiabilidad de los datos, de la curación y edición de los mismos, y, por otro, conocimientos en biomedicina entendiendo el significado y la información que pueden aportar los datos. Además, asegurar una formación continuada de calidad que dé respuesta a los avances que vayan aconteciendo en este campo servirá como herramienta para la traslación de los resultados de la investigación a la práctica clínica.
- Integrar la formación en informática aplicada dentro de los programas educativos de los grados en ciencias de la salud, especialmente en los grados de medicina y enfermería, para asegurar un nivel adecuado de conocimientos en esta área que serán clave en la medicina del futuro.
- Incluir en los programas formativos a nivel de enseñanza primaria y secundaria herramientas dirigidas a incorporar el pensamiento computacional y la programación, de manera que en un futuro se convierta en una habilidad universal en lugar de ser algo exclusivo de los expertos en computación, como medida para asegurar una correcta formación a las generaciones futuras.
- Concienciar, tanto a la comunidad científica como a los ciudadanos, de la importancia de compartir

los datos en salud y la información derivada de los mismos siempre que se garantice su confidencialidad y seguridad, haciendo posible el trabajo en red y el desarrollo de una investigación participativa que facilite el avance de la Medicina Personalizada de Precisión.

- Definir guías hospitalarias y protocolos estandarizados para la interpretación de los datos (p. ej. genómicos) que contribuya a configurar una atención personalizada del paciente.
- Fomentar el diseño e incorporación dentro de las partidas presupuestarias, por parte de las autoridades, de medidas dirigidas a asegurar que los datos sean localizables de una manera rápida y automatizada.
- Impulsar espacios de debate multidisciplinares en los que participen profesionales sanitarios, representantes de las administraciones públicas, representantes del sector privado y representantes de pacientes, con el objetivo de evaluar el coste-beneficio de las aplicaciones derivadas del procesamiento de los datos en salud y, posteriormente, diseñar las posibles estrategias y políticas en salud orientadas a su incorporación en la asistencia sanitaria, teniendo en cuenta la visión del ciudadano.
- Asegurar una completa alineación entre las estrategias nacionales, autonómicas y locales a la hora de trabajar con datos en salud, partiendo de las experiencias exitosas que ya han sido implantadas a pequeña escala a lo largo del territorio nacional.
- Promover una estrecha colaboración entre los diferentes agentes involucrados en la obtención y procesamiento de los datos en salud (administración pública, investigadores, profesionales sanitarios y representantes de pacientes) con el objetivo final de diseñar marcos normativos y éticos que, sin frenar el avance tecnológico, permitan un buen desarrollo de las herramientas basadas en inteligencia artificial.

BIBLIOGRAFÍA

1. Ledley RS. Report on the Use of Computers in Biology and Medicine. Washington DC; 1960. https://books.google.es/books/about/Report_on_the_Use_of_Computers_in_Biolog.html?id=J5grAAAAYAAJ&redir_esc=y. Accessed July 30, 2019
2. Kim K-N, Hong Y-C. The exposome and the future of epidemiology: a vision and prospect. Environ Health Toxicol. 2017;32:e2017009. doi:10.5620/eht.e2017009
3. Li J, Li X, Zhang S, Snyder M. Gene-Environment Interaction in the Era of Precision Medicine. Cell. 2019;177(1):38-44. doi:10.1016/j.cell.2019.03.004
4. Hulsen T, Jamuar SS, Moody AR, et al. From Big Data to Precision Medicine. Front Med. 2019;6(March):1-14. doi:10.3389/fmed.2019.00034
5. The HCA Consortium. The Human Cell Atlas White Paper.; 2017. https://www.humancellatlas.org/files/HCA_WhitePaper_18Oct2017.pdf. Accessed September 5, 2019
6. Adams S, Petersen C. Precision medicine: opportunities, possibilities, and challenges for patients and providers. J Am Med Informatics Assoc Med Info. 2016;23(4):787-790
7. Gilliland AJ. Setting the Stage. In: Baca M, ed. Introduction to Metadata. 2nd ed. The Getty Research Institute; 2008:1-19. http://www.getty.edu/research/publications/electronic_publications/intrometadata/setting.pdf. Accessed September 26, 2019
8. Hekler EB, Klasnja P, Chevance G, Golaszewski NM, Lewis D, Sim I. Why we need a small data paradigm. BMC Med. 2019;17(133):1-9
9. Andreu-Pérez J, Poon CCY, Merrifield RD, Wong STC, Yang GZ. Big Data for Health. IEEE J Biomed Heal Informatics. 2015;19(4):1193-1208. doi:10.1109/JBHI.2015.2450362
10. Menasalvas E, Gonzalo C, Rodríguez-González A. Big Data En Salud: Retos Y Oportunidades. Econ Ind. 2017;(405):87-97. doi:10.1016/S0142-694X(99)00011-3
11. Núñez A, Armengol MA, Sánchez M. Big Data Analysis and Machine Learning in Intensive Care Units. Med Intensiva. 2018;1293. doi:10.1016/j.medin.2018.10.007
12. Durinx C, McEntyre J, Appel R, et al. Identifying ELIXIR Core Data Resources. F1000Research. 2017;5:2422. doi:10.12688/f1000research.9656.2
13. Lee CH, Yoon H-J. Medical big data: promise and challenges. Kidney Res Clin Pract. 2017;36(1):3-11. doi:10.23876/j.krcp.2017.36.1.3
14. Rumsfeld JS, Joynt KE, Maddox TM. Big data analytics to improve cardiovascular care: Promise and challenges. Nat Rev Cardiol. 2016;13(6):350-359. doi:10.1038/nrcardio.2016.42
15. Graham GN. Why Your ZIP Code Matters More Than Your Genetic Code: Promoting Healthy Outcomes from Mother to Child. Breastfeed Med. 2016;11(8):396-397. doi:10.1089/bfm.2016.0113
16. Ministerio de Sanidad Consumo y Bienestar Social. Historia Clínica Digital del SNS. https://www.mscbs.gob.es/organizacion/sns/planCalidadSNS/docs/HCDSNS_Castellano.pdf
17. Ristevski B, Chen M. Big Data Analytics in Medicine and Healthcare. J Integr Bioinform. 2018;15(3):1-5. doi:10.1515/jib-2017-0030
18. Martín-Sánchez F. Big Data Challenges from an Integrative Exposome/Expotype Perspective. In: Househ M, Kushniruk A, Borycki E, eds. Big Data, Big Challenges: A Healthcare Perspective. Lecture Notes in Bioengineering.: Springer, Cham; 2019:127-141. doi:10.1007/978-3-030-06109-8_11
19. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. https://doi.org/10.1038/sdata.2016.18
20. Surkis A, Read K. Research data management. J Med Libr Assoc. 2015;103(3):154-156. doi:10.3163/1536-5050.103.3.011
21. Tecnologías Información. Bases de Datos Distribuidas: Popularidad, Uso y Tipos. https://www.tecnologias-informacion.com/distribuidas.html. Published 2018. Accessed October 2, 2019
22. Castanedo F. A Review of Data Fusion Techniques. Sci World J. 2013;2013:1-19. doi:10.1155/2013/704504
23. Sacristán JA, Dilla T. No big data without small data: learning health care systems begin and end with the individual patient. J Eval Clin Pract. 2015;21(6):1014-1017. doi:10.1111/jep.12350
24. Berger B, Cho H. Emerging technologies towards enhancing privacy in genomic data sharing. Genome Biol. 2019;20(1). doi:10.1186/s13059-019-1741-0

25. Vazquez M, Valencia A. Patient Dossier: Healthcare queries over distributed resources. Panchenko AR, ed. PLOS Comput Biol. 2019;15(10):e1007291. doi:10.1371/journal.pcbi.1007291

26. Jang Y, Choi T, Kim J, et al. An integrated clinical and genomic information system for cancer precision medicine. BMC Med Genomics. 2018;11(Suppl 2). doi:10.1186/s12920-018-0347-9

27. König IR, Fuchs O, Hansen G, von Mutius E, Kopp M V. What is precision medicine? Eur Respir J. 2017;50(4):1-12. doi:10.1183/13993003.00391-2017

28. Gligorijević V, Malod-Dognin N, Pržulj N. Integrative methods for analyzing big data in precision medicine. Proteomics. 2015;16(5):741-758. doi:10.1002/pmic.201500396

29. National Human Genome Research Institute. Electronic Medical Records and Genomics (eMERGE) Network. https://www.genome.gov/Funded-Programs-Projects/Electronic-Medical-Records-and-Genomics-Network-eMERGE. Accessed June 19, 2019

30. Plan de Tecnologías del Lenguaje - Página principal del Plan de Impulso de las tecnologías del Lenguaje. https://www.plantl.gob.es/Paginas/index.aspx. Accessed October 24, 2019

31. Baumgartner C. The Era of Big Data: From Data-Driven Research to Data-Driven Clinical Care. In: Al. XW et, ed. Application of Clinical Bioinformatics. Vol 11. ; 2016:1-22. doi:10.1007/978-94-017-7543-4

32. Miller JB, Shan G, Lombardo J, Jimenez-Maggoria G. Biomedical informatics applications for precision management of neurodegenerative diseases. Alzheimer's Dement Transl Res Clin Interv. 2018;4:357-365. doi:10.1016/j.trci.2018.03.007

33. Xu J, Yang P, Xue S, et al. Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives. Hum Genet. 2019;0(0):0. doi:10.1007/s00439-019-01970-5

34. Wu PY, Cheng CW, Kaddi CD, Venugopalan J, Hoffman R, Wang MD. -Omic and Electronic Health Record Big Data Analytics for Precision Medicine. IEEE Trans Biomed Eng. 2017;64(2):263-273. doi:10.1109/TBME.2016.2573285

35. Doroshow JH. Introduction by the Guest Editor: Oncologic Precision Medicine and the Use of Basket and Umbrella Clinical Trials. Cancer J. 2019;25(4):243-244. doi:10.1097/PPO.0000000000000394

36. Burd A, Schilsky RL, Byrd JC, et al. Challenges and approaches to implementing master/basket trials in oncology. Blood Adv. 2019;3(14):2237-2243. doi:10.1182/BLOODADVANCES.2019031229

37. Sudhop T, Brun NC, Riedel C, Rosso A, Broich K, Senderovitz T. Master protocols in clinical trials: a universal Swiss Army knife? Lancet Oncol. 2019;20(6):e336-e342. doi:10.1016/S1470-2045(19)30271-2

38. Renfro LA, Sargent DJ. Statistical controversies in clinical research: basket trials, umbrella trials, and other master protocols: a review and examples. Ann Oncol. 2016;28(1):mdw413. doi:10.1093/annonc/mdw413

39. Yee LM, McShane LM, Freidlin B, Mooney MM, Korn EL. Biostatistical and Logistical Considerations in the Development of Basket and Umbrella Clinical Trials. Cancer J. 2019;25(4):254-263. doi:10.1097/PPO.0000000000000384

40. Lillie EO, Patay B, Diamant J, et al. The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? Per Med. 2012;8(2):161-173. doi:10.2217/pme.11.7.The

41. Cavanillas JM, Curry E, Wahlster W. Big Data in the Health Sector. In: Cavanillas JM, Curry E, Wahlster W, eds. New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe. Springer Berlin / Heidelberg; 2016:1-303. doi:10.1007/978-3-319-21569-3

42. Pai S, Bader GD. Patient Similarity Networks for Precision Medicine. J Mol Biol. 2018;430(18):2924-2938. doi:10.1016/j.jmb.2018.05.037

43. Kim W-J. Knowledge-based diagnosis and prediction using big data and deep learning in precision medicine. Investig Clin Urol. 2018;59(2):69. doi:10.4111/icu.2018.59.2.69

44. Prosperi M, Min JS, Bian J, Modave F. Big data hurdles in precision medicine and precision public health. BMC Med Inform Decis Mak. 2018;18(1):1-15. doi:10.1186/s12911-018-0719-2

45. Cyrcadia Health. Circadian Thermal Sensing to Detect Breast Disease - Full Text View - ClinicalTrials.gov. https://clinicaltrials.gov/ct2/show/NCT02511301. Published 2015. Accessed June 17, 2019

46. Holmberg C, E. Chaplin J, Hillman T, Berg C. Adolescents' presentation of food in social media: An explorative study. Appetite. 2016;99:121-129. doi:10.1016/j.appet.2016.01.009

47. Reece AG, Danforth CM. Instagram photos reveal predictive markers of depression. EPJ Data Sci. 2017;6(1):15. doi:10.1140/epjds/s13688-017-0110-z

48. M. Joyce, O. Leclerc, K. Westhues HX. Digital Therapeutics: Preparing for Takeoff.; 2018. https://www.mckinsey.com/industries/pharmaceuticals-and-medical-products/our-insights/digital-therapeutics-preparing-for-takeoff

49. Mukherjee S. FDA Approves First Addiction Treatment Mobile App | Fortune. Fortune. https://fortune.com/2017/09/14/fda-alcohol-marijuana-cocaine-mobile-app/. Published 2017. Accessed August 1, 2019

50. Lopez-Campos, Guillermo Merolli M, Martin-Sanchez F. Biomedical Informatics and the Digital Component of the Exposome. In: International Medical Informatics Association, IOS Press, eds. Medinfo 2017: Precision Healthcare Through Informatics. Vol 245. ; 2017:496-500. doi:10.3233/978-1-61499-830-3-496

51. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457(7232):1012-1014. doi:10.1038/nature07634

52. Lazer D, Kennedy R, King G, Vespignani A. The Parable of Google Flu: Traps in Big Data Analysis. Science (80-). 2014;343(6176):1203-1205. doi:10.1126/science.1248506

53. Deloitte. ¿Qué es IoT (Internet Of Things)? https://www2.deloitte.com/es/es/pages/technology/articles/IoT-internet-of-things.html. Accessed June 19, 2019

NOTAS

NOTAS

NOTAS

Fundación InstitutoRoche 15 aniversario

www.institutoroche.es