Técnicas de Integración de Datos Biomédicos

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN
Instituto de Salud Carlos III

IMPaCT

Infraestructura de Medicina de Precisión asociada a la Ciencia y la Tecnología

Programa	IMPacT: Infraestructura de Medicina de Precisión asociada a la Ciencia y la Tecnología		
Nombre Proyecto	IMPaCT-Data: Programa de Ciencia de Datos de IMPaCT		
Expediente	IMP/00019		
Duración	Enero 2021 – Diciembre 2023		
Página web	impact-data.bsc.es		
Paquete Trabajo	WP5. Integración de Datos		
Tarea	Tarea 5.1 Integración de datos genómicos y radiómicos con datos médicos estructurados para su uso secundario		
Entregable	E5.1. Técnicas de Integración de Datos Biomédicos. Informe con las aproximaciones para la integración de datos genómicos, de imagen médica y médicos estructurados considerando las experiencias existentes dentro del programa de Ciencia de Datos y de IMPaCT		
Versión	1.1.1		
Fecha Entrega	30/06/2022	***Fecha Aprobación***	17/05/2023
Responsable	FPS		
Nivel Diseminación	X	PU	Público
		CO-IMP	Confidencial, sólo participantes de los pilares de IMPaCT, incluyendo la comisión de evaluación de IMPaCT.
		CO-DATA	Confidencial, sólo participantes de IMPaCT-Data, incluyendo la comisión de evaluación de IMPaCT.

Autores		
Organización	**Nombre**	**Rol**
Acrónimo	**Nombre y Apellidos**	**Coordinación / Autor / Revisor**
FPS	Joaquin Dopazo	Coordinador/Autor
FPS	Javier Perez Florido	Coordinador/Autor
CRG	Arcadi Navarro	Revisor
CRG	Jordi Rambla	Revisor
UMIB-FISABIO	Mariam de la Iglesia	Revisora
IIS La Fe	Pedro José Mallol Roselló	Autor
IIS La Fe	Carina Soler Pons	Autora
INiBICA	Irene Bernal Florindo	Autora
IIS La Fe	Maria Eugenia Gas	Autora
UMA	Antonio Garcia Ranea	Autor
UMA-CIBERer	James R. Perkins	Autor
UMA-CIBERer	Pedro Seoane Zonjic	Autor
FJD	Pablo Mínguez	Autor
CNAG-CRG	Sergi Beltran	Autor
FPS	Miguel Angel Armengol	Autor

Historial de versiones			
Nro.	**Fecha**	**Descripción**	**Autor**
v 0.1	26/05/2022	Borrador del índice	JD (FPS) JPF (FPS)
v 0.2	11/10/2022	Apartado 2.2.1	Pedro José Mallol Roselló Carina Soler Pons
v 0.3	13/10/2022	Apartado 2.2.2	Irene Bernal Florindo
v 0.4	13/10/2022	Apartado 5	Maria Eugenia Gas
v 0.5	14/10/2022	Apartado 4.2.1	Antonio Garcia Ranea
v 0.6	25/10/2022	Párrafos en apartados 2.1.1 y 4.2.1 y apartados 4.2.2, 4.2.3	Pablo Mínguez
v 0.7	27/10/2022	Párrafos en apartados 2.1.1 (sugerencia nuevos subapartados), 4.1.1, 4.2.1, 5 (tabla).	Sergi Beltran
v 0.8	27/10/2022	Párrafos en apartado 7	Miguel A. Armengol (FPS)
v 1.0	9/11/2022	Versión final validada por los revisores	JD, JPF (FPS)
v 1.1	17/05/2023	Cambio visibilidad a público y aprobado	Comité Dirección
v 1.1.1	14/06/2023	Cambio de formato para publicar en la Web de IMPaCT	David Velasco (ISCIII)

Contenido

Tablas

Figuras

Resumen Ejecutivo

El documento recopila el proceso completo para el análisis de datos biomédicos de forma integrada, desde el proceso de descubrimiento, siguiendo por su solicitud, y discutiendo las posibilidades de análisis, así como algunas de las plataformas y sistemas de análisis de datos existente. También se discute el análisis federado de datos y otros aspectos importantes como los repositorios de código. Aunque no es el objetivo principal de este entregable, se comentan algunos aspectos éticos también relacionados con la solicitud de los datos y su manejo. El documento no pretende ser un listado exhaustivo de métodos y procedimientos y se centra en aquellos más usados en el contexto de la investigación, como punto de partida para una propuesta de buenas prácticas para el desarrollo de una infraestructura de investigación en salud, en el marco del proyecto de Innovación en Medicina de Precisión a través de la Ciencia de Datos, IMPaCT.

Introducción

Audiencia

Este entregable está dirigido a los participantes del proyecto IMPaCT-Data como referencia de procedimientos y tecnologías para la implementación de sistemas que integren información clínica, genómica y de imagen médica para su uso secundario en investigación clínica sin comprometer la privacidad de los datos. Se proporciona la información básica de los requerimientos técnicos necesarios para ello. El documento también puede ser de utilidad para cualquier grupo de investigación o institución interesada en implementar o participar en la implementación de un sistema que integre información clínica, genómica y de imagen médica de estas características.

Ámbito

Este entregable recoge el trabajo sobre estándares y normas hecho en los paquetes 3 y 4, y puede ser utilizado como una referencia general para la realización de estudios que requieran la integración de dos o más tipos de datos, tanto en los casos de uso propuestos por el paquete 6, como en las infraestructuras provistas por el paquete 2 y en otros proyectos de las convocatorias de medicina personalizada.

Relación con otros Entregables

Al tratar sobre integración de datos, este entregable guarda relación con bastantes otros entregables. Entre ellos, los más relacionados serían los entregables: E4.4 relativo a normas de anotación en imagen médica, el E4.1 sobre normas internacionales de información de HCE, E5.4. Requisitos Técnicos para Puesta en Marcha de Sistemas Beacon, E4.2. Comparación de Técnicas de Gestión de Información de HCE, E4.5. Comparación de Técnicas de Gestión de Información de Imagen Médica, E5.5. Especificaciones Revisadas para la Inclusión de Marcadores Biomédicos de Imagen Médica. E3.2. Descripción de Interfaces de Instancias EGA Comunidad y E3.4. Análisis Genómico en Entornos Sanitarios. El presente entregable guarda también relación con el entregable E6.4. Aspectos de Seguridad en el Manejo de Datos Sensibles.

Estructura Entregable

El entregable comienza con una panorámica general del proceso de descubrimiento integrado de datos, seguido por su solicitud, extracción y análisis. A continuación, en distintas secciones se discuten los pasos en detalle:

2. Descubrimiento de datos: donde se describen técnicas de descubrimiento de datos genómicos, más estandarizadas, así como las maneras de descubrir datos clínicos o de imagen.
3. Acceso a los datos: donde se discuten consideraciones para el acceso a los datos clínicos, genómicos o de imagen médica, todos bajo la regulación general de protección de datos.
4. Extracción de datos y entornos de investigación de confianza: donde se comentan las posibilidades de analizar los datos en distintos entornos, con especial hincapié en entornos de investigación e confianza.
5. Visualización y Análisis de datos: donde se hacen consideraciones generales sobre el análisis de datos y se discuten distintas plataformas y herramientas disponibles.
6. Entornos federados: donde se discute el análisis federado de los datos
7. Repositorios de Código: donde se comenta la necesidad de guardar y documentar el código usado para su reutilización.
8. Por último, se exponen las conclusiones finales del análisis desarrollado para la elaboración del presente documento.

1 Esquema general de integración

La Figura 1, muestra el esquema general de integración de datos biomédicos. Partimos de diversos tipos de datos (clínicos, genómicos y de imagen) que estarán distribuidos en diferentes entidades dentro del sistema de salud y donde cada una de las entidades (hospitales o repositorios específicos) tienen datos clínicos normalmente acompañados de datos de imagen médica y eventualmente datos genómicos. Dentro de cada entidad los datos están referidos a los pacientes de los que se han obtenido. Esta presuposición es importante ya que cuando los datos se anonimicen o pseudonimizen para su descarga y eventualmente su extracción, estos mantendrán su coherencia con respecto a cada individuo.

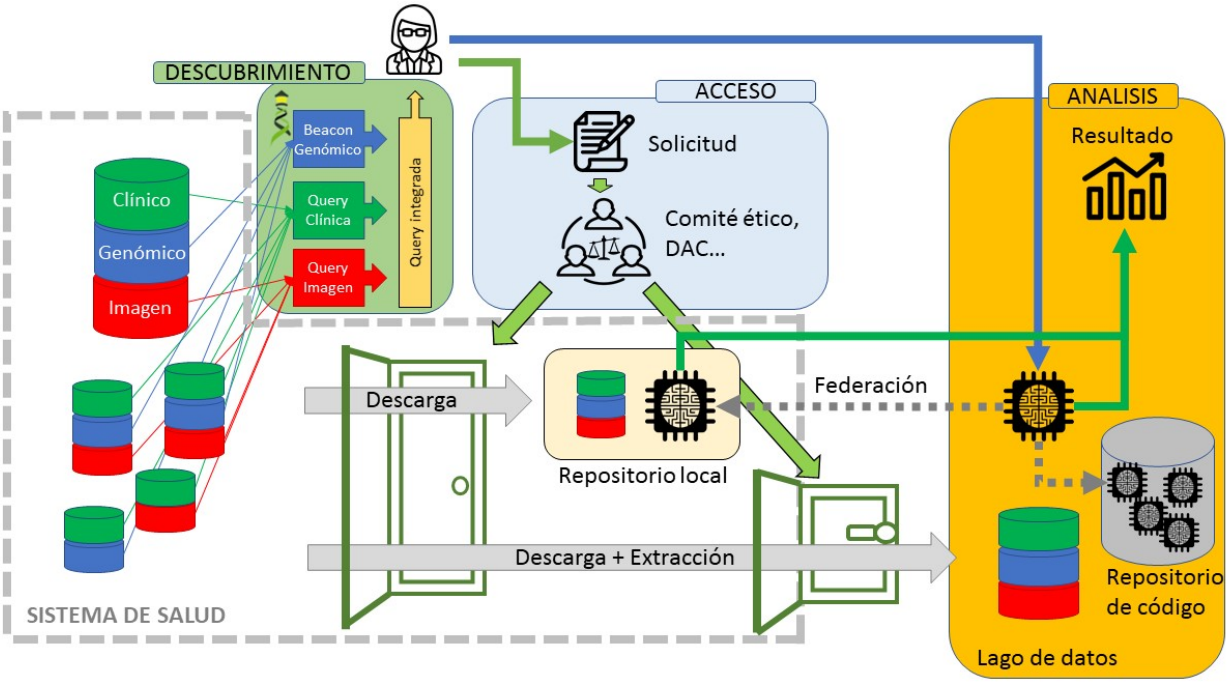

Figura 1. Esquema general del proceso de descubrimiento, acceso y análisis de datos en un entorno seguro de investigación que permite estudios federados con repositorios locales

Distinguimos 3 etapas en el proceso de integración:

- **Descubrimiento**. En esta fase, mediante consultas a datos genómicos (Beacon), clínicos y de imagen, el investigador podrá descubrir si los repositorios de datos de las entidades disponen de los datos necesarios para realizar un determinado estudio. Se permitirán preguntas de tipo AND/OR que permitirán realizar consultas integradas teniendo en cuenta los diferentes tipos de datos.
- **Acceso**. Como los datos están sujetos a Regulación General de Protección de Datos (RGPD) [1], una vez detectados los datos de interés, el investigador deberá solicitar el acceso de acuerdo con lo que marca la regulación y las normativas de acceso a datos

del repositorio. En general, consiste en realizar una solicitud de acceso a los datos de interés al comité correspondiente (Comité ético de investigación -CEI-, *Data Access Committee - DAC*). Estos le autorizarán el acceso a los datos, que puede ser de dos formas posibles: dentro del sistema de salud en un repositorio local de la entidad propietaria de los datos o fuera del sistema de salud, descargando los datos en la entidad definida por IMPACT-Data (lago de datos, WP2)

- **Análisis de los datos y visualización**. Una vez descargados los datos de acuerdo al punto anterior el investigador puede usar herramientas de análisis y visualización para llevar a cabo el estudio autorizado. Cuando además del lago de datos se involucren en el mismo estudio otros repositorios de datos, se requeriría un modelo federado para realizar estudios de los datos. Además del resultado en sí, el código usado para el análisis es un recurso importante que se genera y por tanto es importante disponer de un **repositorio de código** documentado que pueda ser reutilizado para otros estudios, directamente o con modificaciones.

En los siguientes apartados de este documento, desarrollaremos cada una de las etapas mencionadas.

2 Descubrimiento de datos

2.1 Descubrimiento de datos genómicos

Uno de los principales retos a los que se enfrente hoy día la investigación en genética, no solo humana, si no la relativa a otros organismos, es la falta de datos y no porque no se generen los suficientes, sino porque no son compartidos en la mayoría de los casos entre la comunidad científica. Los datos genómicos son identificables y en el caso de datos humanos deben estar protegidos, pero debido a la falta de infraestructura de seguridad sobre ellos y las buenas prácticas en su tratamiento y uso, hace que los investigadores y médicos no los compartan, de forma que los progresos realizados son desconocidos para el resto de la comunidad.

Para el descubrimiento de datos genómicos, describiremos soluciones tipo *Beacon* [2], las cuales son una herramienta que considera, bajo una misma entidad, numerosos conjuntos de datos genómicos. El proyecto *Beacon* se desarrolla bajo la iniciativa GA4GH (*Global Alliance for Genomics and Health*) y de ELIXIR. La API (*Application Programming Interfaces*) de *Beacon* trata de solucionar las limitaciones existentes en la compartición de datos de forma que habilita la búsqueda de variantes genómicas e información asociada sin comprometer la privacidad del conjunto de datos. De esta manera, cualquier institución (de investigación u hospital) puede hacer que sus conjuntos de datos ómicos sean “beaconizados” sin comprometer la privacidad del propietario de los datos. De esta forma, toda la comunidad

científica se ve beneficiada al disponer de un volumen de datos mayor que si no estuvieran accesibles.

En las siguientes secciones, vamos a describir aproximaciones para el descubrimiento de datos genómicos tipo *Beacon*, centrándonos en humano y en patógenos. Comentaremos también el caso particular de la referencia genómica de población española.

2.1.1 Genomas humanos

2.1.1.1 ¿En qué consiste un sistema *Beacon* y por qué es necesario?

En sus orígenes, el protocolo *Beacon* permitía a los usuarios obtener información de la presencia o ausencia de una mutación genómica en un conjunto de individuos, por ejemplo, pacientes de una determinada enfermedad o la población en general (Figura 2).

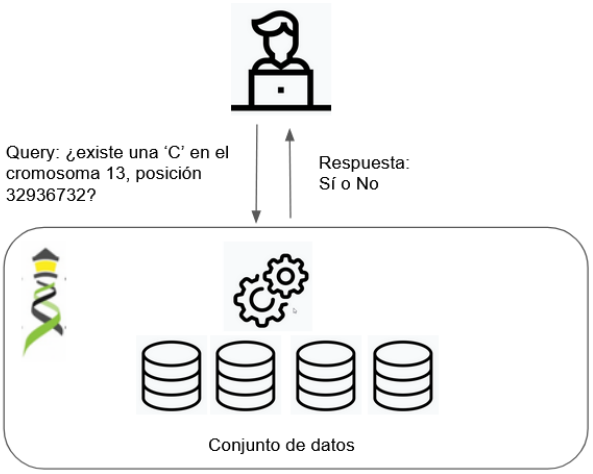

Figura 2. Ejemplo de instancia Beacon v1

Por otro lado, una red de *Beacons* nos permite realizar búsquedas en diferentes instancias *Beacon*, de forma que, con una única consulta sobre una variante genómica, obtenemos una respuesta agregada de todos *Beacons* integrados en la red (Figura 3). Esta estructura, además, favorece la federación de los datos (ver apartado 5 entornos federados)

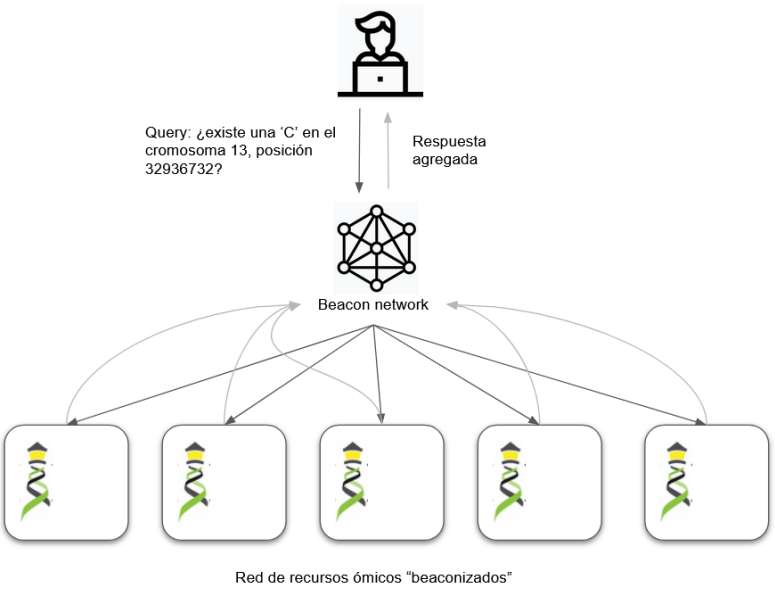

Figura 3. Ejemplo de una red de Beacons con una pregunta y una respuesta agregada

La versión inicial de *Beacon* es algo limitada pues únicamente permite preguntar por la presencia o ausencia de una determinada variante genómica en una instancia Beacon o en una red de Beacons, lo cual hace que su capacidad sea muy limitada

2.1.1.2. *Beacon* v2

Recientemente, la GA4GH ha lanzado la versión 2 de la especificación *Beacon*, la cual define un estándar abierto para el descubrimiento seguro y federado de datos genómicos y fenotípicos en aplicaciones clínicas e investigación biomédica.

Aunque se explica en mayor detalle en el entregable E5.4. Requisitos Técnicos para Puesta en Marcha de Sistemas Beacon, de forma breve la especificación *Beacon v2* consiste en dos componentes principales: el *framework* o entorno y los *modelos*. El entorno define el formato para las peticiones y las respuestas, mientras que los modelos definen la estructura de la respuesta de datos biológicos.

El papel principal de estos componentes es proporcionar las instrucciones para diseñar una API tipo REST (*REpresentational State Transfe*r). Así, únicamente dos elementos básicos son necesarios para implementar una instancia local de Beacon v2: (1) Una base de datos interna donde se almacenan los datos biológicos y (2) una API REST que proporciona una forma estandarizada de recibir peticiones y enviar respuestas. La implementación B2RI *(Beacon v2 Reference Implementation*) desarrollada por el CRG proporciona un conjunto de herramientas software para crear una instancia tipo Beacon [3].

Así pues, la solución B2RI proporciona estos elementos básicos (la base de datos interna y la API REST mencionados), así como un conjunto de herramientas para la transformación de los datos biológicos en un formato de la base de datos interna. B2RI tiene cuatro componentes principales:

- Un conjunto de herramientas para la extracción, transformación y carga de metadatos (metodología de secuenciación, herramientas bioinformáticas), datos fenotípicos y variantes genómicas en una base de datos.
- La base de datos (una instancia de MongoDB)
- El motor de consultas Beacon v2
- Un conjunto de datos de ejemplo con datos sintéticos

El protocolo Beacon v2 cubre diferentes conceptos o entidades y los detalles asociados a ellas. Así, en la actualidad, el modelo incluye las siguientes entidades [4]:

- Conjunto de datos (Dataset): agrupa variantes que tienen algún aspecto en común
- Cohorte (Cohort): un conjunto de características que describen una cohorte, la cual es definida como un conjunto de individuos que pueden pertenecer a uno o más conjuntos de datos
- Individuo (Individual): describe individuos que se almacenan en el repositorio, incluyendo información clínica, como enfermedad, tratamiento o características fenotípicas.
- Biomuestra (Biosample): describe muestras tomadas de individuos, incluyendo detalles sobre el procedimiento de extracción y fechas.
- Experimento (Experiment): incluye detalles del procedimiento utilizado para la secuenciación de una biomuestra.
- Análisis (Analyses): contiene detalles del procedimiento bioinformático para identificar variantes genómicas.
- Variantes genómicas (Genomic Variations): describe cómo una variante está presente en una determinada muestra y si se considera más o menos relevante en el diagnóstico de un caso. Adicionalmente, se incluyen anotaciones sobre el efecto de la variante en un fenotipo dado.

2.1.1.3. Casos de uso existentes tipo *Beacon* para el descubrimiento de datos genómicos

Vamos a describir casos de uso donde se han implementado sistemas Beacon v1 y otros donde se está desarrollando un sistema Beacon v2

CSVS

CSVS (ver sección 2.1.3) es una iniciativa colaborativa que proporciona información acerca de la variabilidad genómica de población española a la comunidad científica. Almacena, en la actualidad, información de más de 2,000 secuencias (entre exomas y genomas) de individuos españoles no relacionados, ya sean sanos o con alguna enfermedad. La base de datos contiene frecuencias alélicas correspondientes a diferentes consorcios y proyectos como *Medical Genome Project*, ENoD del CIBERER y NAGEN 1000 genomas entre otras iniciativas, así como grupos de investigación en España. El repositorio es utilizado como una población pseudo-control para la búsqueda de nuevas variantes y genes responsables de enfermedad. La utilidad de este concepto se demuestra en algoritmos que se basan en pseudo-controles para la priorización de genes [5].

CSVS es descubrible a través de la *Beacon Network* [6] mediante una instancia de Beacon v1 que se comunica con la base de datos de variantes de CSVS. De esta forma, para una variante genómica introducida en la página web de *Beacon Network*, si existe en CSVS, se reportará únicamente su presencia. Si no existe, la instancia Beacon dará la correspondiente repuesta negativa.

Programa ENoD

El programa ENoD (Enfermedades no Diagnosticadas) del CIBERER, nació como un modelo de gestión conjunta para casos sin diagnóstico de origen multicéntrico gracias a la estructura distribuida del Centro de Investigación Biomédica en Red en Enfermedades Raras (CIBERER). Basado en comités transversales, con una herramienta en línea de registro y fenotipado HPO de los casos clínicos y dotado de recursos propios, ofrece orientación al diagnóstico y consejo experto, reinterpretación de datos genómicos previos y nuevas pruebas diagnósticas.

Los datos crudos (ficheros FASTQs) de la secuenciación genómica de estas muestras (exomas clínicos, exomas completos y genomas) son analizados por el Área de Bionformática de la Fundación Progreso y Salud. El análisis de las variantes se realiza en dicha área por herramientas de priorización de variantes desarrolladas para ello y esas variantes genómicas se encuentran almacenadas en una base de datos NoSQL.

Sobre esa base de datos que incluye una cohorte de unas 400 muestras en la actualidad, se está gestionando una instancia *Beacon* v2 donde se implementarán las entidades biomuestra, individuo, variante en muestra y conjunto de datos. Esta instancia *Beacon* permitirá que las variantes recopiladas en el contexto del programa EnoD, sean accesibles por todas las entidades participantes en el mismo y a su vez, si así se considera por parte de los responsables del programa, sea visible a la comunidad científica en un esfuerzo por compartir información que pueda ser útil en el diagnóstico de las enfermedades raras.

RD-Connect GPAP

RD-Connect Genome-Phenome Analysis Platform (GPAP) [7] es una plataforma desarrollada y hospedada en el CNAG-CRG para la colección, análisis e interpretación de datos genómicos y fenotípicos de pacientes de enfermedades raras y familiares. El sistema facilita la compartición de datos y el análisis colaborativo para ayudar en el diagnóstico de enfermedades raras y la búsqueda de nuevas relaciones gene-enfermedad. Actualmente, la RD-Connect GPAP contiene más de 27,000 exomas y genomas integrados con sus datos fenotípicos, mayoritariamente de origen europeo. Se han hecho otras implementaciones de la GPAP para otros proyectos de enfermedades raras y cáncer como la del proyecto PERIS URD-Cat, que contiene más de 1.800 exomas y genomas, incluyendo varios del programa EnoD, u otra instancia local, para proyectos de genómica en Navarra, que contiene más de 2.200 exomas y genomas.

RD-Connect GPAP es parte de la *Virtual Platform de la European Joint Programme of Rare Diseases* (EJP-RD) y ELIXIR, es la puerta de entrada de los datos del proyecto Solve-RD y es un nodo de MatchMaker Exchange. GPAP ha participado en la primera prueba de concepto de 1+MG con el uso de datos sintéticos y está participando en el proyecto Genomed4ALL para permitir Inteligencia Artificial Federada (Federated Learning) para enfermedades hematológicas comunes y raras. RD-Connect GPAP tiene implementado Beacon v1 y es parte de la Beacon Network con las mismas funcionales descritas anteriormente para CSVS. Sin ser parte de la Beacon Network, dos implementaciones adicionales de la GPAP con su propio Beacon v1 forman parte del *Proof of Concept* de enfermedades raras del proyecto europeo *Beyond 1 Million Genomes* B1MG/1+MG [8]. Finalmente, la versión para formación y pruebas de la GPAP [9] dispone de una primera implementación de Beacon v2, incluyendo consultas por genes, fenotipos y enfermedades.

Descubrimiento de datos genómicos en el contexto del proyecto IMPaCT (IMPaCT-Data e IMPaCT-Genómica)

Dentro del contexto IMPaCT, el proyecto IMPaCT-Genómica va a generar alrededor de 2000 WGS de pacientes que serán tomados como prueba de concepto para la compartición de datos genómicos a partir de herramientas como la base de datos *European Genome-Phenome Archive* o los sistemas *Beacon*. La infraestructura creada para tal fin a partir de la colaboración estrecha entre IMPACT-Genómica y Data servirá de ejemplo para la compartición de datos en los centros adscritos y proyectos futuros.

2.1.1.4. Otras consideraciones para el descubrimiento de datos genómicos usando *Beacon*

Privacidad de los datos

Uno de los aspectos a tener en cuenta es la seguridad y privacidad de los datos, ya que con repetidas consultas a un *Beacon* llegaría a ser posible identificar (un) individuo/s en la base de datos [10]. En este sentido, es recomendable evaluar cada una de las posibles distintas consultas y respuestas y estudiar la implementación de técnicas para minimizar estos riesgos a través, por ejemplo, de la ofuscación de datos o, como mínimo, la autenticación y autorización de usuarios.

Autenticación y Autorización

En el caso de implementar un sistema de autenticación y autorización, se recomienda que se acuerde el estándar a utilizar para poderlo federar entre los distintos actores. Un ejemplo de infraestructura de autenticación y autorización (AAI) sería *Life Science Login* [11]. Finalmente, para agilizar la gestión de las autorizaciones, se propone evaluar sistemas cómo GA4GH Passports [12].

Datos sintéticos

Con el objetivo de facilitar las pruebas de concepto e implementaciones sin poner en riesgo datos sensibles, hay datos sintéticos disponibles para la comunidad. Por ejemplo: EGAS00001005702 [13] que consiste en datos sintéticos genómicos y clínicos de 18 genomas completos correspondientes a 6 tríos de enfermedades raras.

2.1.2 Genomas de patógenos

Debido a la grave situación de pandemia del coronavirus SARS-CoV-2, una de las herramientas recomendadas para la vigilancia epidemiológica del virus propuesta por la Organización Mundial de la Salud (OMS) y secundada por numerosas autoridades públicas ha sido la secuenciación genómica del patógeno.

En España, la Comisión de Salud Pública del Consejo Interterritorial hizo la recomendación de integrar la secuenciación del genoma en la vigilancia epidemiológica en enero de 2021. Numerosas comunidades autónomas adoptaron esta recomendación, entre ellas la comunidad andaluza en la que, desde enero de 2021, se puso en marcha un circuito clínico para la secuenciación de genomas SARS-CoV-2 a lo largo de toda la geografía andaluza y reportar las variantes con información sobre su potencial riesgo para la salud (variantes de preocupación o VOCs y de sospecha o VOIs, así como las mutaciones de preocupación.

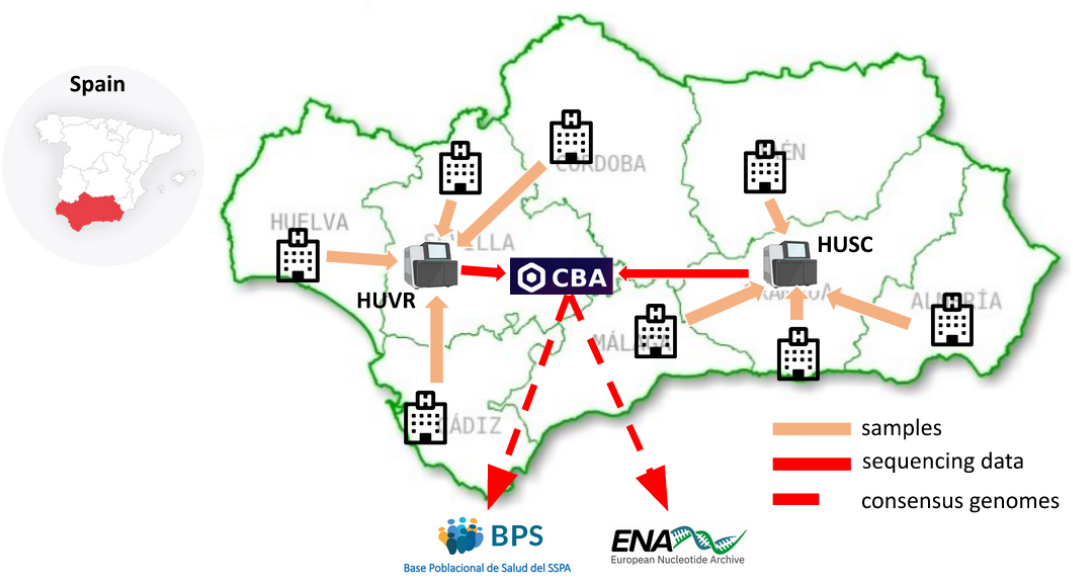

Figura 4. Circuito de secuenciación del genoma del SARS-CoV-2 en Andalucía

En este circuito (Figura 4), los hospitales de cada provincia de la comunidad andaluza mandan una selección de muestras recogidas de pacientes positivos a los hospitales de referencia (Virgen del Rocío de Sevilla para el oeste de Andalucía y San Cecilio de Granada para el este), donde se secuencian con un protocolo común. Los datos se transfieren a la Plataforma de Medicina Computacional (antigua Área de Bioinformática) donde se procesan y se analizan. Los resultados se devuelven a los hospitales de referencia, desde donde además se reportan al Ministerio de Sanidad. Las secuencias se utilizan para obtener una filogenia de los virus, que se actualiza cada semana en un servidor *Nextstrain* local. Adicionalmente, la Plataforma de Medicina Computacional envía todos los genomas virales a la Base Poblacional de Salud (BPS) [14], donde quedan integrados con el resto de datos clínicos y de imagen médica que el sistema Andaluz Público de Salud obtiene de sus usuarios y a la base de datos internacional *European Nucleotide Archive* (ENA), para ponerlas a disposición de la comunidad científica.

La BPS alberga datos clínicos y del uso de recursos sanitarios de todas las personas que reciben asistencia sanitaria en el Servicio Andaluz de Salud (su contenido, acumulado desde 2001, es de más de 13 millones de registros). La principal ventaja de almacenar el genoma viral en BPS es que éste se almacena de forma permanente y vinculado a la información clínica de los pacientes como una prueba clínica más. Esto no solo permite tener un repositorio centralizado y accesible, si no que permite acceder, en ese mismo repositorio, a los datos clínicos de los pacientes junto al dato genómico, lo que permite desarrollar estrategias de descubrimiento de ambos tipos de datos dentro del sistema sanitario. En este

sentido, se está desarrollando un sistema de descubrimiento de datos tipo *Beacon* v2 para los datos genómicos de SARS-CoV-2

Por otro lado, cabe mencionar aproximaciones ya desarrolladas en el contexto de descubrimiento de datos genómicos de SARS-CoV-2 como COVID-19 Viral Beacon [15]. COVID-19 Viral Beacon permite a los usuarios buscar determinadas variantes genómicas del patógeno y explorar ciertos metadatos asociados con una aproximación tipo *Beacon* como las descritas más arriba. Por ejemplo, se pueden realizar búsquedas de ciertas cepas virales de una región geográfica concreta. O, por ejemplo, permitir realizar búsquedas de ciertas variantes genómicas que están presente en una proporción baja en la población viral de cada individuo.

2.1.3 Referencia genómica de la población española

En la actualidad, más de 4500 enfermedades monogénicas pueden ser diagnosticadas directamente por la genómica personalizada [16], una posibilidad que pronto podría extenderse a todo el espectro de enfermedades raras de base genética [17]. Aun así, en la actualidad más de un 50 de los pacientes con enfermedades raras quedan sin un diagnóstico conclusivo por no mostrar ninguna mutación conocida de enfermedad. Entre las estrategias utilizadas para descubrir nuevas variantes de enfermedades, especialmente en los trastornos monogénicos, el filtrado heurístico de variantes basado en la frecuencia ha demostrado ser una herramienta muy útil [18]. El fundamento es el siguiente: las variantes que son relativamente comunes en una población de control (variación común) son probablemente benignas [19], mientras que las variantes raras (especialmente si tienen consecuencias funcionales) que se encuentran en múltiples casos afectados pero que están ausentes en la población de control son probablemente causantes de la enfermedad [20].Estos filtros buscan genes o variantes presentes en todos (o la mayoría) de los individuos afectados, pero en ninguno (o muy pocos) de los individuos de control no afectados. Por lo tanto, la disponibilidad de controles sanos es un factor decisivo para el progreso del descubrimiento de nuevos determinantes de la enfermedad [21].

Desde una perspectiva histórica, el Proyecto 1000 Genomas produjo el primer catálogo completo de la variación genética humana común [22]. Sin embargo, se sabe que las variantes de baja frecuencia (con frecuencias alélicas menores, MAF, por debajo del 5%) y raras (MAF por debajo del 0,5%), típicamente específicas de la población, están poco representadas en dicho catálogo o en otros ampliamente usados, como gnomAD [23]. De hecho, estudios recientes han descrito un componente local importante y un alto nivel de estratificación [24] en muchas variantes raras con consecuencias funcionales inciertas. Como consecuencia de ello, el riesgo de muchas enfermedades difiere en distintas poblaciones humanas según sus antecedentes genéticos [25]. Lo que sugiere que el conocimiento de la variabilidad genética de la población local es un factor crítico para el descubrimiento de nuevas variantes de enfermedades [26]. Todas estas observaciones ponen de manifiesto la necesidad de disponer

de catálogos de variación genética específicos para cada población [27]. Sin embargo, hasta la fecha sólo se han llevado a cabo unas pocas iniciativas para estudiar la variación genética a nivel poblacional [28], entre ellas CSVS [21], ampliamente usado en proyectos nacionales y especialmente dentro del contexto del CIBERER.

Desde 2017, la información genómica de CSVS es descubrible a través de la red GA4GH Beacon [6]. Aunque CSVS almacena los datos en 1-base, puede responder a consultas tanto en 1-base como en 0-base (Beacon solicita los datos en 0-base). También está disponible un formulario para realizar directamente consultas al estilo de Bacon [29].

2.2 Descubrimiento de datos de imagen médica

2.2.1 Datos de radiología

Al contrario que los datos genómicos, la imagen médica es un tipo de dato cuyo uso está bien consolidado dentro del sistema sanitario. Al igual que el resto de datos, la imagen médica se puede encontrar en dos escenarios básicos: en un entorno clínico asistencial, formado por todas las exploraciones realizadas a los pacientes a lo largo de su proceso diagnóstico (su uso primario), y en un entorno de investigación (su uso secundario). En los siguientes apartados se va a explicar el descubrimiento de las imágenes médicas, así como las características de los sistemas de gestión de la información. Por último, se incluye una descripción del procesamiento mínimo requerido en caso de querer trabajar con datos de imagen radiológica para fines de investigación.

2.2.1.1 Entorno clínico asistencial: producción de los datos

Los hospitales y centros sanitarios utilizan los sistemas PACS (*Picture Archiving and Communication System*) para gestionar la información de las imágenes médicas. Concretamente, los PACS se encargan del almacenamiento, gestión, visualización y distribución de las imágenes médicas y su información relacionada (metadatos).

Las tres etapas asociadas a las imágenes que se relacionan con los PACS hospitalarios son las siguientes:

- **Adquisición:** se obtienen las imágenes a partir de las diferentes modalidades diagnósticas (entendiendo por modalidad cada una de las técnicas usadas para la obtención de la imagen) en formato digital usando el protocolo DICOM (*Digital Imaging and Communication in Medicine*) como estándar de transmisión. A su vez, DICOM dispone de diferentes funcionalidades y servicios, entre ellos: almacenamiento o archivo (*Storage*), consulta y recuperación (*Query/Retrieve*), impresión (*Print Management*) y gestión de la lista de trabajo (*Basic Worklist Management*).

- **Almacenamiento:** las imágenes se almacenan en servidores de datos que gestionan los diferentes dispositivos de almacenamiento, tanto internos como externos, normalmente basados en discos duros redundantes RAID (*Redundant Array od Independent Disks*). Adicionalmente, puede ampliarse la capacidad de almacenamiento añadiendo discos duros o, incluso, almacenamiento en la nube gracias a servicios como *Amazon Web Services* (AWS) o Azure. Estos servidores suelen contar con Sistemas de Alimentación Ininterrumpida (SAI) que estabilizan la corriente e incluso pueden proporcionar energía por un tiempo limitado de forma controlada en caso de cortes prolongados de corriente.
- **Visualización y procesamiento:** las imágenes pueden estar accesibles desde estaciones de trabajo diferentes, lo que proporciona una gran flexibilidad para trabajar con ellas. Además, estos sistemas incorporan aplicaciones para la visualización, edición e interpretación posterior de las imágenes, pudiendo así obtener un informe final de la exploración realizada.

El PACS no es un sistema aislado que recibe y distribuye imágenes. La interacción con los Sistemas de Información Radiológica (RIS) y los Sistemas de Información Hospitalaria (HIS) es fundamental para el mejor aprovechamiento de las capacidades del PACS. El RIS recopila, controla y explora todos los datos que se obtienen del servicio de radiodiagnóstico, desde la solicitud de la prueba hasta su informe asociado. En el HIS se almacenan y procesan los datos médicos y administrativos de los pacientes. Para realizar todo este intercambio de información se utilizan diferentes estándares de intercambio de datos electrónicos como *Health Level 7* (HL7).

Además, existen PACS comunitarios que permiten la intercomunicación de los sistemas en los diferentes centros, para visualizar y transmitir las imágenes entre hospitales. Existen muchos modelos ya implementados a lo largo de las diferentes comunidades autónomas que facilitan la gestión de la imagen médica en la comunidad, como por ejemplo el sistema de Gestión de Imagen Médica Digital de la Comunidad Valenciana (GIMD) o el PACS Regional del Servicio Andaluz de Salud.

2.2.1.2 Entorno de investigación

Los repositorios son los principales sistemas de gestión de la imagen médica en el entorno de investigación fuera de su uso asistencial. Según el grupo de trabajo de la Sociedad Europea de Radiología [30], estos repositorios se definen como una base organizada de datos de imágenes médicas y biomarcadores de imagen asociados, compartida entre investigadores y ligada a otros biorepostorios. Existen diferentes tipos de repositorios de imagen médica en función de sus características, pero, en general, pueden agruparse en tres áreas:

- **Repositorios para la investigación clínica:** infraestructura para archivar, compartir y difundir (para uso secundario) los datos de imágenes que se utilizaron originalmente en el contexto de proyectos de investigación clínica, como los ensayos clínicos. En este caso, las imágenes pueden haber sido sometidas a algún tipo de procesamiento para extraer uno o más biomarcadores de imagen relevantes para la cuestión científica que motivó el estudio.
- **Repositorios para enfermedades específicas:** recursos para recibir, archivar, compartir y difundir imágenes en ámbitos clínicos específicos, por ejemplo, la enfermedad de Alzheimer o la esclerosis múltiple. Dichos sistemas tienen como objetivo recopilar las imágenes clínicas de pacientes con una determinada patología a gran escala (por ejemplo, a escala nacional). Estos biobancos de imágenes también pueden basarse en una iniciativa de cribado regional o nacional que recoja datos de imágenes de un grupo de personas con características específicas.
- **Repositorios con datos de población general:** colecciones de datos obtenidos de la población asintomática. Estos repositorios cuentan con la mayor cantidad de datos recogidos del mayor número de sujetos posible, usualmente con fines epidemiológicos.

Un ejemplo de estos tipos de repositorios son el Banco de Imagen de la Comunidad Valenciana (BIMCV) [31]. Por otra parte, tal y como se explicó en el entregable E4.4 (Normas Internacionales de Anotación de Información de Imagen Médica), se puede hacer una diferenciación entre repositorios centralizados, federados o mixtos en función de su arquitectura y el lugar de almacenamiento de los datos. Además, cabe considerar que el acceso a estas plataformas puede ser abierto, restringido o de pago, según las características propias de cada repositorio. En el Anexo A del entregable E4.5 (Comparación de Técnicas de Gestión de Información de Imagen Médica) se puede consultar más información sobre algunos de los repositorios más usados en imagen radiológica, así como ejemplos concretos de colecciones de datos que pueden encontrarse en ellos.

2.2.1.3 Procesamiento mínimo requerido

Como se ha comentado anteriormente, en la mayoría de los casos existe un sistema PACS el que se gestiona el flujo de las imágenes, por lo que la extracción se puede realizar mediante la funcionalidad de consulta y recuperación (Query/Retrieve) que dispone el sistema. Ahora bien, se debe realizar un procesamiento mínimo requerido, siguiendo unas pautas que garanticen la calidad de las imágenes y la privacidad de los pacientes.

Para asegurar que se protege la identidad de los pacientes, se aplican técnicas de **anonimización** sobre las imágenes. Es necesario anonimizar los perfiles DICOM, los metadatos y también cualquier característica identificable de las imágenes. Estos procesos de anonimización están descritos en detalle en el documento E6.4 (Aspectos de seguridad

en el manejo de datos sensibles) y en el documento E4.4 (Normas Internacionales de Anotación de Información de Imagen Médica).

Además, para disponer de un conjunto de datos de calidad para uso secundario es necesario aplicar procesos de **homogeneización** y **armonización** de los datos. En el contexto de la imagen médica, estos procesos incluyen la **normalización** de las imágenes, es decir su el re-escalado a un mismo tamaño, un mismo rango de valores de los píxeles, y la eliminación de imágenes atípicas. Esta homogeneización de las imágenes es necesaria para eliminar posibles sesgos que produzcan un sobreajuste de los modelos creados con estos datos. Previamente a la aplicación de estos procesos de homogeneización, es necesario realizar un estudio de las imágenes para determinar las técnicas a aplicar más adecuadas.

Por último, es también necesario hacer un **control de la calidad** de las imágenes. De forma automática se puede, por ejemplo, detectar imágenes con tamaños atípicos (demasiado grandes o pequeñas) o rangos de valores extremos. No obstante, es recomendable realizar también un control manual de las imágenes para detectar casos como imágenes que no se corresponden a la parte del cuerpo deseada o protocolos de adquisición de las imágenes erróneos.

2.2.2 Datos de patología digital

2.2.2.1 Entorno clínico en Anatomía Patológica

La patología digital es la aplicación de la patología usando imágenes digitales en lugar de la observación en el microscopio convencional. El término patología digital incluye todos los procesos relacionados con maximizar la utilidad de las imágenes digitales, incluyendo la adquisición, almacenamiento, visualización, anotaciones y sus usos en aplicaciones tanto clínicas como no clínicas (investigación y educación). En esta disciplina destaca la técnica *Whole Slide Imaging* (WSI), modalidad de adquisición de imagen que consiste en el escaneado completo de laminillas histológicas y citológicas de cristal, generando las correspondientes preparaciones digitales.

Para que se mantenga el flujo de información, el Sistema de Información de Anatomía Patológica (SIPAT) debe estar integrado con el resto de sistemas de información clínica. Dentro de estos sistemas se encuentra el Sistema de Información Hospitalaria (HIS), que realiza la petición electrónica con los parámetros relevantes para el estudio. Existen dos sistemas principales de información externos con los que el SIPAT tiene que relacionarse. Por un lado, se encuentran los servidores de terminología, generalmente basados en SNOMED-CT, que mantienen actualizados los vocabularios controlados de muestras, procedimientos quirúrgicos, servicios, observaciones y diagnósticos clínicos y patológicos. Por otro lado, se encuentran los servidores de imagen PACS, ya que las imágenes generadas

en el departamento de Anatomía Patológica (macroscópicas, microscópicas y preparaciones digitales) deben estar disponibles en el servidor central.

El SIPAT son los sistemas que gestionan los datos y las imágenes de forma eficiente para la generación de los informes anatomopatológicos finales y su incorporación al historial médico del paciente. Las funciones del SIPAT en algunos casos son equivalentes a los del Sistema de Información del Laboratorio (LIS), pero incluye otras funciones similares a las del RIS. Todas estas funciones se pueden agrupar globalmente en tres grupos:

- Sistemas encargados de generar la información e integración de los datos clínicos, radiológicos y patológicos. Para ello, es necesaria una colección de información bien estructurada y una comunicación bidireccional para poder mantener los datos actualizados.
- Sistemas encargados de la gestión de las imágenes generadas. Las fuentes de adquisición de las imágenes son variadas.
- Los sistemas encargados del control del proceso técnico y la evaluación de la información que permita una mejora continua del proceso.

2.2.2.2 Entorno de investigación en Anatomía Patológica

En Anatomía Patológica el repositorio más usado es *The Cancer Genome Atlas* (TCGA) [32]. Este repositorio ha caracterizado molecularmente más de 20.000 muestras de tumores primarios abarcando 33 tipos de cánceres. Esta iniciativa entre el Instituto Nacional de Cáncer (NCI) en EEUU y el Instituto de Investigación del Genoma Humano comenzó en el año 2006, reuniendo a diferentes investigadores de varias disciplinas e instituciones. Se puede acceder públicamente a los datos clínicos, de imagen y genómicos a través de su portal de datos TCGA. A nivel europeo estaría el repositorio Eurobioimaging [33], un recurso open, centralizado, de imágenes biomédicas.

2.2.2.3 Procesamiento mínimo requerido para uso secundario

Las imágenes digitales de Anatomía Patológica se tienen que procesar previamente a su uso en investigación mediante unas técnicas para garantizar la calidad de las imágenes y la privacidad de los pacientes. Primero se tienen que someter a un proceso de anonimización, de manera que no se pueda identificar la identidad de la persona a la que pertenece esa imagen. Una vez que las preparaciones digitales se encuentran anonimizadas, se les somete a un control de calidad para asegurarse que aquellas que no cumplan alguno de los criterios establecidos y requeridos (resolución suficiente, áreas desenfocadas, regiones anatómicas indeseadas) sean excluidas del estudio.

En el Departamento de Anatomía Patológica del Hospital Puerta del Mar (Cádiz), investigadores del Instituto de Investigación Biomédica de Cádiz (INiBICA) llevaron a cabo

dos proyectos aplicando la tecnología de la patología digital. En uno de ellos, el proceso de anonimización de las imágenes se realizó eliminando la etiqueta identificativa de la preparación digital mediante el visor SlideViewer 2.5 (3D Histech Ltd., Budapest, Hungría). Las imágenes anonimizadas fueron sometidas a un control de calidad manual, asegurándose de que tuvieran suficiente resolución y no hubiera zonas desenfocadas, proceso que se realizó también mediante el mismo visor mencionado anteriormente. En el otro proyecto, para anonimizar las preparaciones digitales se usó un desarrollo propio basado en librerías de *OpenSlide* [34]. El control de calidad fue manual, para evitar realizar los análisis de imágenes posteriores con alguna de las preparaciones desenfocadas o con algún error en el proceso de escaneo, usando un visor web propio desarrollado durante el proyecto [35].

2.2.2.4 Flujo de procesado de la imagen de Anatomía Patológica

Tras los procesos mínimos requeridos de anonimización y control de calidad, las preparaciones digitales se someten a un flujo de procesado en el que se encuentran diferentes métodos (Figura 5).

Primero es necesario realizar a las imágenes una normalización del color. La gestión del color permite la estandarización de las distribuciones del color a través de la imagen de entrada. La mayoría de las aplicaciones de la inteligencia artificial en patología digital se centran en el campo del cáncer; la normalización del color se ha basado hasta ahora en preparaciones teñidas con hematoxilina y eosina.

Para el análisis de las WSI se requieren pasos adicionales, debido a limitaciones computacionales producidas por el gran tamaño de las imágenes. Para solventar esta situación, la imagen WSI se divide en porciones más pequeñas denominadas teselas.

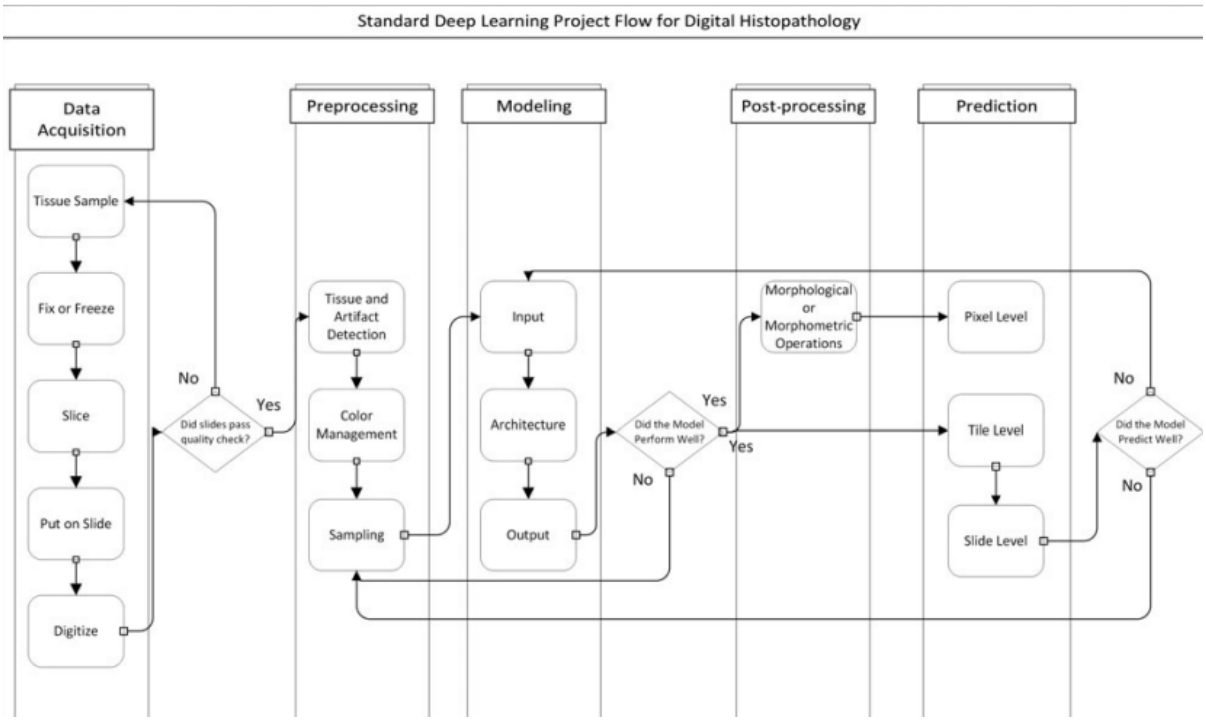

Figura 5. Flujo de procesado en la imagen de Anatomía Patológica, extraída de Smith et al. (2021) [36]

Posteriormente, se procede a la cuantificación de los objetos segmentados y medidas morfológicas (tamaño, forma, redondez, distancias) en las preparaciones digitales. Estas características cuantificadas pueden tener una utilidad clínica en sí mismas (como por ejemplo el tamaño de un tumor) o pueden ser usadas en algoritmos más complejos de inteligencia artificial para generar modelos predictivos y de clasificación de objetos o imágenes.

Por último, se encuentra la predicción, en el que hay varios escenarios que plantean en qué nivel de la unidad experimental hay que centrarse para realizar la predicción; a nivel de tesela o a nivel de preparación. El Entregable E4.5 (Comparación de Técnicas de Gestión de Información de Imagen Médica) contiene información adicional a este respecto.

2.3 Descubrimiento de datos clínicos

Una gran cantidad de los sistemas de información asistenciales o administrativos que van a ser fuente de datos para los repositorios o lagos de datos de uso secundario responden a modelos de datos desarrollados ad hoc por parte de los servicios técnicos de los centros y servicios sanitarios, o bien son soluciones comerciales apoyadas en modelos de datos propietarios de la empresa desarrolladora, y que no responden a ningún tipo de estandarización en la estructura de los datos. Esta disparidad en los modelos de datos de uso primario va a tener una serie de implicaciones a la hora de extraer información de los

mismos para su uso secundario (diseño específico de procesos de extracción, imposibilidad de reutilizar procesos de ETL, requerimiento del conocimiento de la estructura interna de los modelos de datos, las relaciones y las dependencias entre entidades de información, necesidad de participación de técnicos TIC y de la empresa desarrolladora para la extracción).

Algunos sistemas de Historia Clínica Electrónica (cada vez más) están construidos sobre "modelos duales" o "modelos basados en arquetipos". Estos modelos permiten modelar con mucha más flexibilidad el conocimiento del dominio sanitario, y por tanto facilitan la interoperabilidad semántica entre sistemas de uso primario y secundario [37]. La arquitectura de modelo dual, propuesta en 2000 por Thomas Beale [38], plantea una separación clara entre los modelos de información y de conocimiento. Esto lo hace a partir de un modelo de referencia (RM) y uno de arquetipos (AM). El (RM) contiene las entidades básicas capaces de representar cualquier información contenida en la Historia Clínica Electrónica, mientras que el conocimiento de dominio se implementa mediante el AM. Los arquetipos son combinaciones estructuradas y restringidas de entidades del modelo de referencia, que formalizan los conceptos del dominio clínico. Son la unidad básica de conocimiento en el sistema de información.

Por último, una de las infraestructuras de datos que, cada vez con más frecuencia, están desplegando instituciones y servicios sanitarios son los almacenes y lagos de datos para uso secundario. Un lago de datos, o *datalake*, es un sistema o repositorio de datos almacenados en formato natural, es decir, sin tener en cuenta su nivel de estructuración, pudiendo incluir datos totalmente desestructurados tales como textos en lenguaje natural, imágenes o vídeos. El guardado de datos en bruto permite un almacenamiento de datos masivo y muy rápido; no obstante, obliga a la aplicación de un procesamiento posterior sobre estos datos para que sean utilizables.

El descubrimiento de datos clínicos se centra en el desarrollo de métodos y herramientas que actúan sobre las Historias de Salud Electrónicas, en tanto que son la fuente principal de datos sobre la que trabajar.

La importante inversión que se está realizando en las Historias de Salud Electrónicas (EHR siglas en inglés aceptadas en la literatura) debe aprovecharse para acelerar de forma rentable la investigación en medicina de precisión a escala poblacional y para reducir los costes. El descubrimiento de datos clínicos requiere tanto de los datos codificados disponibles en la EHR como las caracterizaciones fenotípicas enterradas en el texto narrativo del registro médico mediante el procesamiento del lenguaje natural (PLN). Entre las ventajas del descubrimiento de datos clínicos frente a los estudios de cohorte convencionales se encuentran la relevancia clínica oportuna y la escalabilidad rentable. Los biobancos y los estudios de cohortes existentes utilizan cada vez más los datos derivados de la EHR para aumentar las caracterizaciones fenotípicas obtenidas.

En este sentido son destacables las iniciativas para el uso secundario de los datos de las EHR, como i2b2 orientado al descubrimiento y normalización de datos de cohortes de pacientes [39] con largo recorrido y aceptación en España [40,41]. Más avanzadas son las iniciativas de extracción y "meta-datación" descriptiva de datos de las EHR que incluso permiten el cumplimiento de los principios FAIR [42,43].

3 Acceso a los datos

3.1 Consideraciones

Hay que tener en cuenta que todos los datos contemplados están sometidos a regulación de protección de datos. El entregable E6.4. Aspectos de Seguridad en el Manejo de Datos Sensibles, amplia aspectos sobre el manejo de estos datos. En general, en el contexto de uso secundario de datos para la investigación clínica, una vez se ha verificado, mediante las técnicas de descubrimiento, que los datos de interés existen, se procede a la petición de autorización al comité ético de investigación correspondiente. En función de la gobernanza de los datos, este CEI puede ser el del propio hospital, o en casos de repositorios construidos reuniendo datos de distintos hospitales puede ser uno de mayor nivel. Por ejemplo, las iniciativas PADRIS [44] o iRWD [45] requieren de CEIS de más alto rango. El caso de iRWD, que usa datos de la BPS, requiere el CEI regional andaluz [46].

En particular, los datos genómicos, si están en el estándar de almacenamiento EGA, disponen de Comité de Acceso a los Datos (DACs) [47], que son organismos formados por una o más personas nombradas que son responsables de la divulgación de los datos a los solicitantes externos sobre la base del consentimiento y/o de los términos de la ética de la investigación nacional. Un DAC suele estar formado, aunque no necesariamente, por la misma organización que recogió las muestras y generó cualquier análisis asociado. En general, el uso de DACs, que simplemente verifican que el estudio solicitado es compatible con los consentimientos para los que los datos pedidos se obtuvieron y autorizan su uso, acelera el proceso de solicitud de datos en comparación con CEIs, aunque depende de los DACs y los CEIs específicos. Los entregables E3.2. Descripción de Interfaces de Instancias EGA Comunidad y E3.4. Análisis Genómico en Entornos Sanitarios contienen más detalle sobre este aspecto.

3.2 Solicitud y descarga de los datos

En el caso de los datos clínicos, el proceso de solicitud de los datos diferirá dependiendo de cómo estén modelados y almacenados.

En los casos en que los datos están modelados bajo estándares que adoptan el modelo dual (openEHR, EN/ISO 13606), se puede realizar una solicitud utilizando el *Archetype Query Language* (AQL), un lenguaje declarativo, independiente de las aplicaciones, del sistema y del modelo de almacenamiento y persistencia. Este lenguaje de consulta lo que hace es expresar las consultas a nivel semántico (de los arquetipos) y no a nivel de los datos.

El requerimiento mínimo para que los datos puedan ser consultados con AQL es que estén basados en arquetipos, conteniendo marcas semánticas con elevada granularidad, bajo la forma de códigos de arquetipos y terminologías. Esto puede corresponder tanto a datos nativos de algún modelo de referencia (openEHR, EN/ISO 13606) o datos provenientes de un sistema legacy al cual se le agregaron los marcadores semánticos relevantes.

Consecuentemente, AQL expresa consultas bajo la forma de combinaciones de elementos semánticos de los arquetipos y elementos de estructura de datos de los modelos de referencia sobre los que están basados esos arquetipos. La estructura de resultado de una consulta AQL, en su forma cruda, es una tabla de dos dimensiones, conceptualmente similar a la proyección tabular generada por una consulta SQL.

Disponer de los datos de salud en un lago de datos propietario, es decir, que no sigue un modelo de datos estándar, requiere que el responsable del mismo, conocedor de los metadatos y la estructura del repositorio, defina para cada caso de uso las consultas necesarias para extraer los datos requeridos en un estudio concreto, así como los criterios o atributos que definen a la población de estudio. Estos factores que identifican la cohorte de estudio pueden ser múltiples, aunque generalmente estarán presentes las dimensiones geográficas y temporales -en sus escalas de edad y calendario-, además de otros criterios clínicos estructurados, tales como diagnósticos codificados siguiendo vocabularios de enfermedades (CIE9, CIE10 OMS, CIE 10-ES, CIAP, SNOMED), o pautas de prescripción de fármacos, en múltiples sistemas de clasificación.

Otra de las configuraciones posibles que nos podemos encontrar, o que podemos plantearnos a la hora de construir una infraestructura de datos para uso secundario, en la configuración basada en almacenes de datos definidos sobre un modelo común de datos. Es decir, tendremos todos los datos estructurados recogidos de los sistemas de información primarios (Historia Clínica Electrónica y otros sistemas asistenciales) dentro de un modelo común de datos, como puede ser OMOP-CDM. La principal ventaja es que, a la hora de definir un subconjunto de datos para compartir con otros participantes mediante un modelo común de datos, no será preciso mapear ningún dato entre dos modelos de datos distintos, ya que los datos se encuentran estructurados según el modelo de destino. Será necesario, simplemente, realizar una selección de los datos del almacén de acuerdo con los criterios de inclusión en la cohorte, y el conjunto de datos complementarios necesarios para el análisis. Además, al partir de modelos de datos estandarizados y ampliamente utilizados, existe una gran cantidad

de herramientas y procesos ya desarrollados para trabajar con los mismos, y que podemos utilizar sobre nuestro repositorio.

4 Extracción de datos y entornos de investigación de confianza

La Figura 6 ilustra las maneras en las que se puede acceder a datos clínicos siguiendo el circuito de solicitud. Como se ha comentado, se requiere la aprobación de un CEI y en muchos casos un estudio de evaluación de impacto en protección de datos [48].

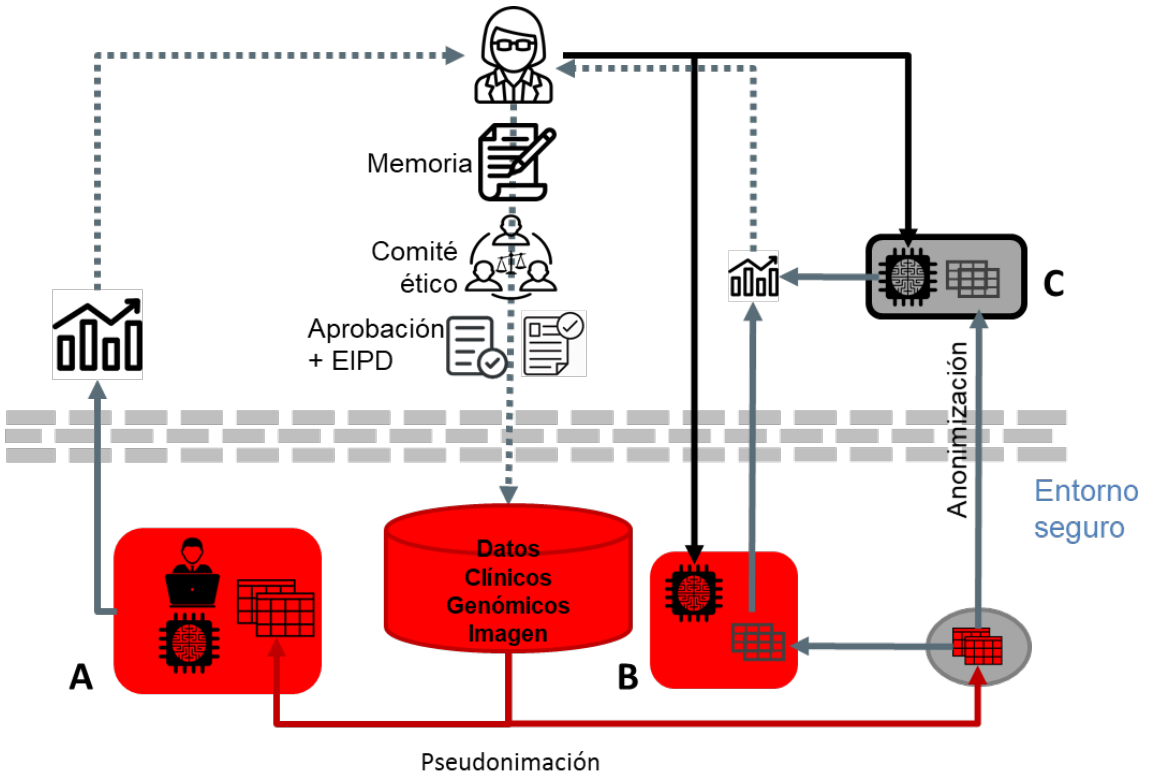

Figura 6. Distintos métodos de extracción de datos: A) TRE manual, B) TRE automático, C) método convencional de extracción de la información a un entorno que no es de confianza, vía anonimización.

La vía más frecuente implica la extracción de datos previa anonimización (Figura 6C) para evitar una serie de riesgos que van dese la pérdida de datos hasta la posible re-identificación de pacientes [49]. Las EIPD suelen penalizar este tipo de estudios. El documento "10 malentendidos relacionados con la anonimización" de la Agencia Española de protección de Datos [50] muestra como la anonimización es un proceso que se hace ad hoc para cada estudio específico y además no garantiza en el tiempo la privacidad, ya que van apareciendo cada vez más metodologías y datos open que permiten re-identificar individuos.

Los entornos de investigación de confianza (*Trusted Research Environment* - TRE), también conocidos como "refugios de datos" o "entornos de datos seguros", son entornos informáticos de alta seguridad que proporcionan acceso remoto a datos protegidos por RGPD para que los investigadores autorizados los utilicen en sus investigaciones [51].

Los TREs ofrecen a los investigadores una ubicación para acceder a conjuntos de datos clínicos. Dentro del TRE los datos y las herramientas de análisis se encuentran en un solo lugar. En vez de "liberar" de datos clínicos a un entorno menos seguro, los TRE proporcionan acceso a un entorno analítico seguro ("entorno seguro") para que los investigadores aporten análisis (como algoritmos) a los datos. La puesta a disposición de los datos a través de una TRE proporciona a los pacientes y al público la confianza de que se accede a sus datos sanitarios personales de forma segura y se protege su privacidad, cumpliendo con la RGPD. Los TREs pueden usarse individualmente o en federación (ver sección 5 Entornos Federados).

Los TRE también contribuyen a que la investigación sea eficiente, colaborativa y rentable, ya que evitan los costes de transferencia y almacenamiento de duplicados de conjuntos de datos cada vez más grandes, especialmente de imágenes y datos genómicos. También evita a los investigadores la responsabilidad de tener que garantizar la seguridad de los conjuntos de datos descargados.

En un TRE tanto manual (Figura 6A), en el que los datos sean analizados por personal del TRE, como automático (Figura 6B), en el que investigadores externos podrían analizar datos sin tener acceso a ellos ni capacidad de copiarlos fuera. La EIPD en estos casos es muy favorable ya que los riesgos de pérdida de datos o de re-identificación de pacientes se reducen drásticamente.

Un ejemplo de TRE es la Plataforma para la generación segura de conocimiento biomédico a partir del big data clínico de la base poblacional de salud [45], que correspondería al caso de TRE manual (actualmente en transición al automático). .

5 Visualización y Análisis de datos

5.1.1 Consideraciones generales

Existen dos tipos de obstáculos para el análisis integrado de datos: i) **obstáculos técnicos**, debidos a la falta de herramientas informáticas que permitan estudiar de forma integrada datos de distinta naturaleza, y ii) **obstáculos culturales**, derivados de la forma en la que tradicionalmente, debido a las dificultades de integración de los datos, se han analizado estos. Simplificando la situación, podemos distinguir tres dominios principales de datos: datos **clínicos**, datos de **imagen médica** y datos **genómicos**. Los expertos en datos clínicos

normalmente se bastan sólo con dichos datos. Los expertos en datos de imagen médica los estudian en conjunción con datos clínicos, como covariables en el estudio. Finalmente, los expertos en genómica estudian los datos genómicos en conjunción con un subconjunto muy específico de datos clínicos codificados en la ontología de fenotipos humanos (HPO). Aunque hay algunos casos en los que se estudian datos genómicos y datos de imagen (p. ej. transcriptómica espacial [52]) son estudios más específicos y tampoco hemos encontrado ejemplos claros de uso simultáneo integrado de datos de los tres dominios.

5.1.2. Herramientas de visualización y análisis de datos

En general las herramientas de visualización y análisis de datos tienden a estar bastante mezcladas ya que la visualización es parte del análisis inicial y de la representación de resultados. Un aspecto importante en la selección del software que se va a usar para hacer análisis es la transparencia de las metodologías, la posibilidad de hacer comparaciones (benchmarking), además de la reproducibilidad requerida por los principios FAIR. Esto implica, por un lado la necesidad de que el software sea disponible y open [53]. Por otra parte, incluso en el caso de software abierto, es recomendable evitar algoritmos de "caja negra", en los que no es fácil deducir que variables son las que se están usando para tomar decisiones. En particular, en algoritmos de inteligencia artificial hay una tendencia a usar algoritmos explicables (XAI) [54].

5.1.2.1. Entornos de ejecución de *workflows*

Una solución genérica a cualquier tipo de análisis consiste en el uso de entornos de ejecución de *workflows*. Estos entornos permiten ejecutar *workflows* en máquinas locales de escritorio o a través de una infraestructura mayor (como superordenadores, Grids o entornos en la nube). Estos entornos de ejecución de *workflows* vienen con acceso a varios miles de herramientas y recursos diferentes que están disponibles gratuitamente en una gran variedad de instituciones de ciencias de la vida. Una vez construidos, los *workflows* son protocolos bioinformáticos ejecutables que pueden compartirse, reutilizarse y reingenierizarse. Ejemplos de entornos de ejecución de *workflows* son Taverna [55], Galaxy [56] o Kepler [57]. Otro entorno muy popular es Bioconductor [58]. Aunque facilitan las tareas de programación, estos entornos requieren de ciertos conocimientos de programación.

5.1.2.2. Plataformas de visualización y análisis de datos genómicos

Existen aplicaciones específicas para para la visualización de datos genómicos y plataformas más completas para el análisis de datos genómicos, orientadas al diagnóstico de variantes de enfermedad, y que ofrecen una solución tanto para el almacenamiento de dichos datos como para su visualización y manejo. Entre las soluciones *open-source* se citan a continuación algunos ejemplos, sin ánimo de hacer una lista exhaustiva.

La visualización de datos genómicos suele requerir del uso de un ***Genome Browser*** (Navegador de Genoma). Los *Genome Browsers* necesitan tener acceso, local o remoto, a una secuencia genómica de referencia (ej: una versión específica del genoma de referencia humano versión) y a sus anotaciones, y a uno o más fichero de alineamientos genómicos (generalmente en formato BAM [59] o CRAM [60]). Opcionalmente, se pueden cargar ficheros adicionales como por ejemplo VCF [61] (de variantes). Las opciones de visualización y filtros a seleccionar pueden ser variadas en función de los datos cargados (p.ej: datos de genoma, o de RNA-Seq, variantes estructurales, etc.). Para permitir un acceso rápido y eficiente, y la visualización de datos de alineamientos genómicos, GA4GH ha definido el estándar htsget [62], que permite acceder a secciones de BAM/CRAM alojados en servidores remotos, permitiendo su visualización sin tener que descargar todo el fichero.

Aunque se han usado distintos *Genome Browsers* a lo largo del tiempo, como *Genome Maps* [59], el estándar actual claramente es el ***Integrated Genome Viewer*** (IGV) [63], una aplicación de código abierto (bajo licencia MIT) disponible en varias formas (aplicación de escritorio, integración en aplicación web, generación de informes desde línea de comandos, etc.), lo que facilita su uso dependiendo de la casuística. IGV soporta el uso de GA4GH htsget. EGA dispone de un servicio htsget, al cual se pueden conectar distintas herramientas para permitir a sus usuarios visualizar alineamientos genómicos allí almacenados. Por ejemplo, EGA y RD-Connect GPAP han puesto a punto un servicio de visualización remoto con IGV, que está disponible, de momento, para los datos del proyecto Solve-RD.

La plataforma **cBio Cancer Genomics Portal** [60] es un recurso abierto para la exploración, visualización y análisis de datos multidimensionales relacionados con la genómica del cáncer. La interfaz de la herramienta combinada con un almacenamiento de datos específico permite la exploración de datos genómicos permitiendo la visualización y análisis a lo largo de genes, muestras y tipos de datos. Los usuarios pueden visualizar patrones de alteraciones genómicas a lo largo de muestras en un estudio de cáncer, comparar frecuencias de mutaciones a lo largo de varios estudios de cáncer o resumir todas las alteraciones genómicas relevantes en una muestra tumoral concreta. El portal también proporciona exploración biológica de *pathways*, análisis de supervivencia, análisis de exclusividad mutua entre alteraciones genómicas, descarga de datos y una API para un acceso programático.

Otra potente herramienta *open-source* es el ***Interactive Variant Analysis***, IVA [61]. IVA es una herramienta web derivada de prototipos previos [62,63] que se usaron en el CIBERER para la búsqueda de genes de enfermedad que permite la búsqueda, el filtrado, el análisis y la interpretación de datos genómicos. A grandes rasgos, IVA permite realizar varios tipos de análisis: a) exploratorio de variantes genómicas en un estudio en concreto pudiendo realizar numerosos filtros de las variantes en base a su anotación, b) un análisis más clínico de un conjunto de muestras (trio, casos individuales), con el objetivo encontrar la variante diagnóstica de enfermedad, donde también se pueden aplicar numerosos filtros en base a la anotación y generar informes clínicos, c) análisis de variantes somáticas de cáncer

Para realizar los diferentes tipos de análisis, IVA trabaja con la información de variantes almacenada en OpenCGA [64], el cual es un software para el almacenamiento y recuperación de datos genómicos y datos clínicos asociados. Tanto IVA como openCGA son soluciones basadas en el proyecto OpenCB [65], un conjunto integrado de herramientas de alto rendimiento para el manejo y el análisis de datos genómicos a nivel poblacional que ha sido usado con éxito en el proyecto de los 100.000 genomas de UK.

Otra solución *open-source* es ***PriorR***, un interfaz gráfico para el diagnóstico de enfermedades genéticas a partir de los resultados de secuenciación masiva. Es posible instalarlo tanto localmente como en un servidor para su acceso remoto y/o comunitario. Es capaz de analizar tanto paneles de genes como WES y WGS. Aunque la versión distribuida de PriorR está acoplada a la anotación del pipeline de detección de variantes utilizada en el IIS-FJD, puede configurarse a cualquier otro tipo de anotación mediante un archivo de configuración. Ofrece un sistema de priorización y filtrado de variantes para SNVs y CNVs con diferentes facilidades. PriorR está disponible en *github* [66] y como imagen de *docker* para su fácil instalación [67].

Otra herramienta a destacar es **GPAP** (*Genome-Phenome Analysis Platform*). GPAP es un sistema escalable e interoperable que permite la recopilación, análisis, interpretación y compartición de conjuntos de datos genómicos junto con su información fenotípica. La herramienta está especialmente enfocada al diagnóstico de enfermedades raras y el descubrimiento de nuevos genes, aunque existe un módulo para el análisis e interpretación de mutaciones somáticas.

Existe también un gran número de herramientas *open-source* para la priorización de variantes de enfermedad cuando estas no aparecen en los genes esperados y que pueden facilitar las tareas de búsqueda. Estas herramientas pueden ser usadas dentro de entornos de *workflow* también. Una de las más populares es **exomizer** [68], y merece la pena citar alguna otra desarrollada dentro de este consorcio, como **GLOWgenes** [69], que utiliza diferentes redes de asociación funcional para integrar su información y generar un ranking de genes que puede usarse como valor de priorización en el proceso de análisis genético. Hay múltiples variantes a este modelo, con más o menos granularidad, más o menos automatización, y distintos requerimientos, tanto iniciales cómo de computación o almacenamiento. Por ejemplo, en el proyecto europeo Solve-RD se ha realizado el análisis programático a través de una API de miles de exomas y genomas a través de la interrogación de una base de datos tipo Big Data que contiene todas las variantes y anotaciones de los individuos y familiares a analizar [70]. Este tipo de aproximaciones permite los reanálisis periódicos con un mínimo esfuerzo.

6 Entornos federados

El tratamiento y la explotación de grandes volúmenes de registros clínicos pueden ofrecer múltiples beneficios a la sociedad siempre que se mantenga el respeto por los derechos de las personas, su privacidad y la protección de sus datos personales. En la actualidad en función de dónde se almacene la información y cómo se accede a ella se distinguen dos tipos de bases de datos (Figura 7). Las bases de datos centralizadas son aquellas en las que la información se encuentra almacenada en una única localización. Mientras que, las redes de datos federada o distribuidas, albergan múltiples bases de datos interconectadas.

En función de la administración y el grado de integración de los nodos que conforman la red, las redes federadas se clasifican en sistemas acoplados débil o estrechamente. En el primer supuesto no existe ningún control impuesto por el sistema federado y sus administradores y la responsabilidad de formar y mantener la red reside en cada uno de los nodos locales. Mientras que, en el segundo, la responsabilidad reside en su administrador (o administradores).

El aspecto más interesante de las bases de datos federadas es que los datos pueden permanecer en sus lugares de origen y ser analizados allí, lo que en caso de datos sujetos a RGPD es una gran ventaja.

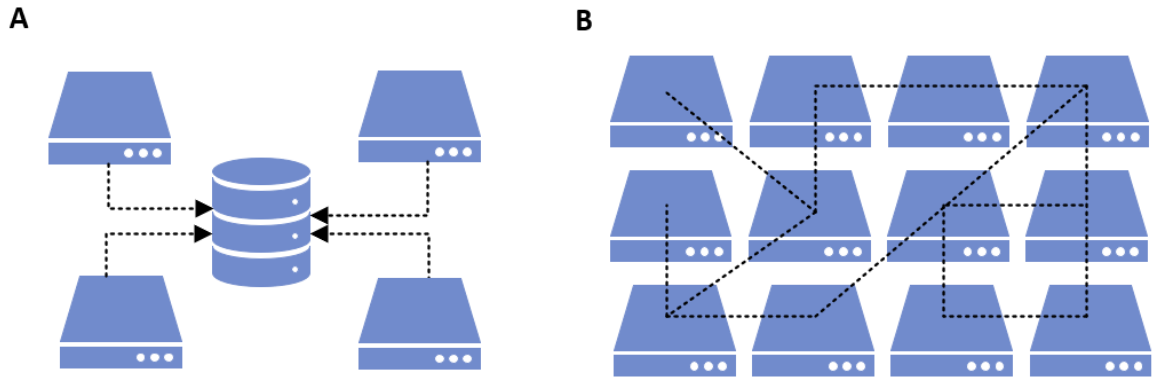

Figura 7. Representación esquemática de una base de datos centralizada (A) versus una base de datos federada o distribuida (B).

El desarrollo e implementación de entornos federados de datos hace necesario superar una serie de dificultades tecnológicas y organizativas entre las que destacan: i) garantizar el acceso eficiente a cada uno de los nodos de la red y la fidelidad de la información almacenada en los sistemas locales y globales; ii) integrar los diferentes tipos de procesamientos que tienen lugar entre los nodos de la red; iii) controlar el acceso a los datos; iv) hacer frente a los errores que puedan producirse en los diferentes módulos del sistema de manera segura y

eficiente; v) estimar la capacidad y el tráfico dentro de la red; y vi) considerar la competencia de recursos entre los nodos que conforman la red.

Pero a pesar de las barreras mencionadas anteriormente, las ventajas competitivas de dichos entornos frente a los sistemas centralizados explican el crecimiento y aceptación de los entornos federados de datos especialmente en el ámbito de la salud. Siendo la capacidad de atender consultas globales, sin interferir con el funcionamiento de los nodos locales, la implementación de soluciones de protección de datos desde el diseño, así como el establecimiento de políticas controladas de reutilización de datos uno de los principales atractivos de los entornos federados de datos.

En la tabla 1, se describen diversas iniciativas de trabajo en entornos federados en el ámbito de compartición de datos clínicos y genómicos que están implementadas actualmente. Estas propuestas permiten la utilización de datos de forma federada para la ejecución de estudios en colaboración en el ámbito nacional e internacional. Se caracterizan por la utilización de modelos comunes de datos (Common Data Model) en los que es necesaria la transformación de los datos de origen mediante vocabularios estándares internacionales.

Tabla 1. Ejemplo de iniciativas dirigidas a la implementación de entornos federados

Inicitativa	***Ámbito***	***Descripción***
TriNetX	Europeo	La plataforma TriNetX permite la reutilización segura los registros de las historias clínicas electrónicas estandarizadas al formato i2b2, facilitando la colaboración entre personal clínico e investigadores, con el objetivo de maximizar los resultados de la investigación clínica.
ELIXIR	Europeo	Ecosistema federado de servicios interoperables que proporcionan un marco para la presentación, archivo, difusión y análisis seguros de los datos genómicos y biomoleculares a nivel poblacional europeo. Acelerando la investigación y mejorando la salud de las personas residentes en toda Europa
European Health Data & Evidence Network (EDHEN)	Europeo	Red de excelencia pública-privada cuyo objetivo es desarrollar un ecosistema federado de instituciones que armonicen sus datos clínicos al modelo común de datos OMOP y creen una red de tecnología para la investigación en el mundo real.
Observational Health Data. Sciences and Informatics (OHDSI)	Internacional	Red internacional de investigadores y bases de datos de salud estandarizadas al modelo común de datos OMOP. El centro coordinador de la red está ubicado en la Universidad de Columbia
European Open Science Cloud (EOSC)	Europeo	Su ambición es proporcionar a los investigadores, innovadores, empresas y ciudadanos europeos un entorno multidisciplinar federado y abierto donde puedan publicar, encontrar y reutilizar datos, herramientas y servicios con fines de investigación, innovación y educación
1+ Million Genomes	Europeo	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

		de manera que estén disponibles para la investigación que dará lugar a una medicina personalizada.
The Federated Tumor Segmentation (FeTS) initiative		Red de aprendizaje federado formada por 30 institutos sanitarios que trabajan para mejorar la detección de los límites de los tumores
Data Analysis and Real World Interrogation Network (DARWIN)	Europeo	El objetivo de esta iniciativa, impulsada por la Agencia Europea del Medicamento (EMA), es facilitar a todos los reguladores de medicamentos acceso a pruebas fiables del mundo real, sobre enfermedades, poblaciones de pacientes y el uso, la seguridad y la eficacia de los medicamentos y desarrollar y gestionar una red de fuentes de datos sanitarios del mundo real en toda la Unión Europea para realizar estudios científicos solicitados por los reguladores de medicamentos y, en una fase posterior, solicitados por otras partes interesadas
European Joint Programme on Rare Diseases Virtual Platform (EJP-RD VP)	Europeo	La Virtual Platform es un ecosistema federado que tiene por objetivo abrir una puerta única a la descubierta, interrogación y, eventualmente, acceso a registros de pacientes, biobancos, repositorios genómicos y multi-ómicos, plataformas de análisis de datos, bases de datos de conocimiento, recursos animales y celulares, materiales de soporte y servicios translacionales y de investigación clínica, etc.
Genomed4ALL	Europeo	Genomed4ALL representa un salto quántico en medicina personalizada avanzada, reuniendo datos genómicos / ómicos de salud a través de una plataforma federadada de aprendizaje (Federated Learning) para enfermedades hematológicas comunes y raras.
MatchMaker Exchange	Internacional	Red federada de plataformas con datos genéticos y/o fenotípicos que permite interrogación recíproca de recursos para hallar casos de enfermedades raras similares en cuanto a fenotipo y/o gen causal candidato. https://www.matchmakerexchange.org/

Una de las ventajas de los entornos federados de datos es que permiten la implementación de entornos de aprendizaje distribuido (Figura 8), los cuales posibilitan entrenar modelos predictivos de manera iterativa o ejecutar códigos analíticos localmente y compartir sólo resultados agregados.

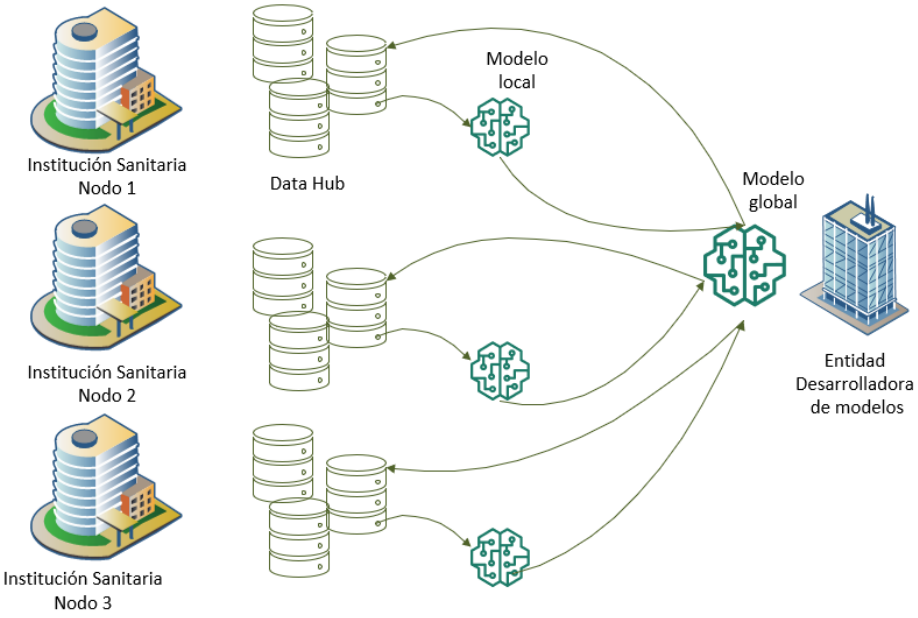

Figura 8. Representación esquemática de un sistema de aprendizaje federado.

El aprendizaje federado es un enfoque genérico de aprendizaje automático cuyo objetivo es entrenar modelos de alta calidad con datos distribuidos entre varios proveedores de forma independiente preservando de este modo la privacidad. Es decir, en lugar de reunir los datos en un único servidor central, los datos permanecen almacenados en los proveedores de datos, mientras que los algoritmos y los modelos predictivos se intercambian entre ellos. Al superar el cuello de botella que supone el intercambio de datos, este novedoso paradigma puede contribuir a que el aprendizaje automático alcance todo su potencial, sobre todo en el ámbito de la sanidad, donde los datos son especialmente sensibles. De hecho, se espera que la adopción del aprendizaje federado conduzca a modelos entrenados en conjuntos de datos de un tamaño sin precedentes, lo que tendrá un impacto catalizador hacia la medicina de precisión/personalizada.

En este contexto, la anonimización de los datos personales adquiere un valor especial como una fórmula que puede garantizar el avance de la sociedad de la información sin menoscabar el respeto a la protección de datos. Para alcanzar este objetivo es preciso garantizar la irreversibilidad de la anonimización. El avance de la tecnología y la información disponible hacen difícil garantizar el anonimato absoluto, especialmente a lo largo del tiempo. Para solventar esta limitación, la comunidad científica en los últimos años está investigando en el uso de técnicas conocido como *Privacy Enhancing Technologies (PET*), basada en métodos criptográficos avanzados que permiten realizar operaciones sobre datos cifrados bajo unos protocolos que garantizan que los datos nunca salgan del perímetro de seguridad del responsable de los datos.

Una de las tecnologías desarrolladas para implementar e incrementar la seguridad de este tipo de entornos de aprendizaje federado es la Computación Segura Multi-parte o SMPC (por sus siglas en inglés, *Secure Multiparty Computation*) en la cual diferentes entidades/elementos computacionales se unen para ejecutar operaciones cada uno sobre una parte de los datos, manteniendo así la privacidad del conjunto total de los datos. Esta aproximación demanda altas capacidades de computación, algo que las hace inviables con las soluciones computaciones empleadas hasta la fecha. Sin embargo, se está estudiando la aplicación de las técnicas de Big Data, con el objetivo de conseguir que la computación sobre información cifrada sea práctica.

En la figura 9 se representa esquemáticamente la comparación entre los sistemas de aprendizaje centralizado (panel A) frente a un sistema de aprendizaje federado en el que se ha implementado protocolos de SMPC (panel B). En la segunda opción se siguen los siguientes pasos: i) cada entidad ha de generar de un modo aleatorio N valores teniendo presente que la suma de los valores generados por cada uno de ellos ha de coincidir con el valor real; ii) cada entidad comparten utilizando canales seguros todos los valores generados excepto uno, el cual conserva para el siguiente paso; iii) cada entidad dentro de su entorno suma todos los valores recibidos con el no compartido obteniendo así una suma parcial; iv) todas las entidades comparten de un modo seguro sus resultados parciales, los cuales se suman para obtener el resultado global sin que ninguno de ellos haya revelado sus propios datos. Es de destacar que la privacidad de la información queda garantizada sólo si el número de entidades implicadas en el proceso es superior a dos, ya que de lo contrario cada una de las dos entidades podría inferir los registros de la otra.

Alguno de los productos disponibles que ofrecen capacidad de federación de datos son: Teradata con Query Grid, IBM Pure Data Systems con Fluid Query, GMV con uTile y SAP HANA con Smart Data Services, Open Federated Learning, FedAUX

Ejemplos de iniciativas que usan datos federados son EHDEN [71], el proyecto *European Health Data & Evidence Network* que aspira a ser el ecosistema de investigación observacional de confianza que permita mejorar las decisiones, los resultados y la atención sanitaria. Su misión es proporcionar un nuevo paradigma para el descubrimiento y el análisis de datos sanitarios en Europa, mediante la construcción de una red federada a gran escala de fuentes de datos estandarizadas según el modelo de datos común OMOP. Otra iniciativa similar es OHDSI [72], *The Observational Health Data Sciences and Informatics*, con un carácter más internacional.

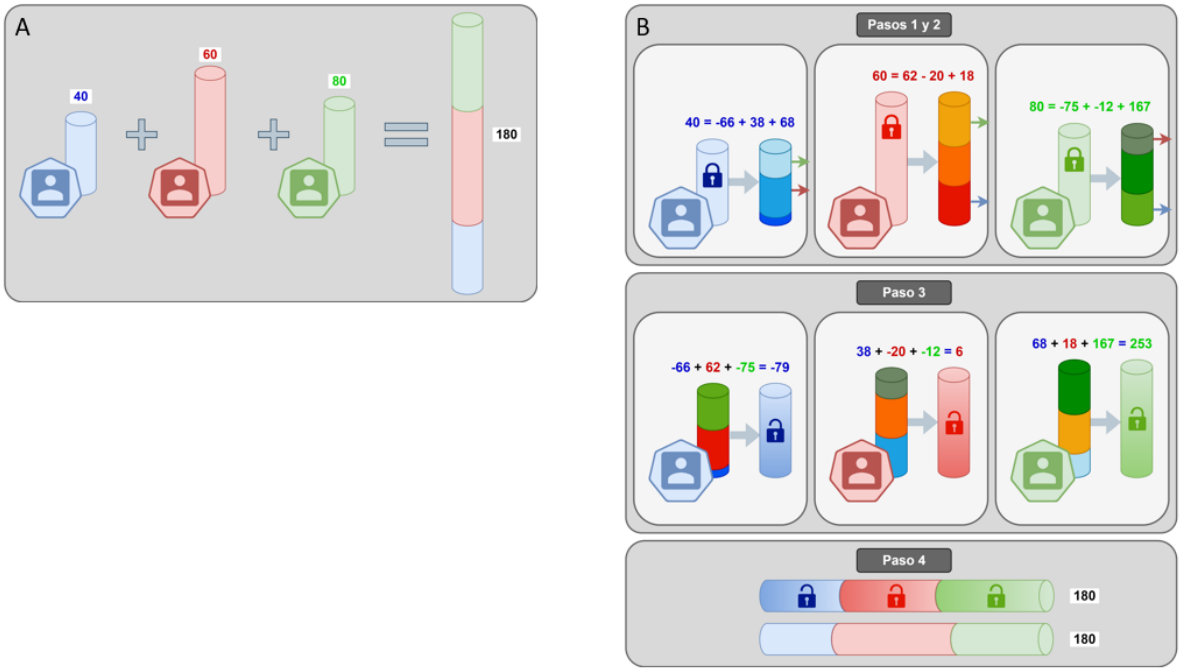

Figura 9. Esquema de funcionamiento de un sistema de aprendizaje centralizado (A) y el de la Compartición Aditiva de Secretos para llevar a cabo la suma de los datos de los participantes (B). Imágenes cortesía de GMV Soluciones Globales Internet, S.A.U

7 Repositorios de Código

La investigación moderna basada en el análisis de los registros clínicos de salud depende cada vez más del análisis cuantitativo de los mismos. Pero a pesar de esta realidad todavía no se han desarrollado ni adoptado mecanismos eficaces que garanticen la transparencia y la reproducibilidad de los resultados generados ni de los análisis realizados. La importancia y los beneficios derivados de la difusión, comunicación y compartición de los resultados y desarrollos generados en el marco de los proyectos de investigación ha sido ampliamente reconocido por la comunidad científica. De hecho, en la investigación biomédica, no sólo es imprescindible publicar una descripción detallada del diseño, metodología, resultados e interpretación del estudio, sino que existe la necesidad apremiante de ponerlos a disposición de la comunidad científica para incrementar la transparencia y reproducibilidad de las investigaciones. Pero a pesar de los beneficios derivados de esta aproximación, estudios recientes han puesto de manifiesto la falta creciente de reproducibilidad en todas las disciplinas científicas, es decir, que los resultados publicados contienen a menudo análisis que no se reproducen debido a la falta de documentación, código y datos necesarios para verificar el análisis.

Ante esta necesidad desde hace unos años se ha trabajado en la creación de repositorios de código los cuales posibilitan su almacenamiento y su distribución. Dichos repositorios están dotados de sistemas de control que permiten llevar un control de las versiones y modificaciones realizadas tanto sobre la versión original como sobre cada versión generada con posterioridad. Así mismo, permiten la reversión de los cambios realizados, así como el trabajo colaborativo. A continuación, se mencionan algunos de los beneficios de estos tipos de repositorios:

- Posibilitar trabajar colaborativamente entre dos o más usuarios de una aplicación o programa sin que se den por válidas versiones con elementos que entren en conflicto entre sí.
- Facilitar un entorno de alta seguridad garantizada mediante diferentes métodos avanzados de ciberseguridad y la creación constante de copias de seguridad.
- Facilitar la gestión y monitorización de los proyectos, así como la optimización de los recursos al facilitar un historial de cambios y requerir la inclusión de explicaciones de los cada uno de los cambios implementados. Esta práctica permite evitar duplicidades y facilitar la corrección de errores.

El código depositado en el repositorio debe seguir los principios de la buena programación científica, tal y como señalan [73], incluido el desarrollo incremental con un sistema de control de versiones distribuido, pruebas unitarias y un rastreador de problemas público.

De entre los repositorios de código más empleados, destacan Git y SVN. El primero de ellos proporciona un sistema de administración de código fuente distribuido de código abierto. Sistema que permite trabajar a varios desarrolladores simultáneamente y que no requiere de permisos especiales de lectura y escritura para los diferentes directorios. Por contra SVN, no permite que varios usuarios puedan acceder simultáneamente a un mismo archivo y tan sólo incluye las últimas versiones de cada código.

La elección de un entorno u otro deberá ajustarse a las necesidades de cada proyecto ya que mientras Git ofrece un entorno más flexible y ágil que permite trabajar en local con la seguridad de disponer los archivos en un almacén central. SVN por contra se emplea habitualmente cuando se trabaja con archivos de gran tamaño y se quiere agrupar todo el proyecto en un único lugar.

En un entorno Git, se podría animar o requerir que los usuarios de la plataforma compartiesen el código que utilizan para el procesamiento y el análisis de datos. Como hemos dicho, compartir el código ayuda a que los estudios sean reproducibles y promueve la investigación colaborativa. La manera de contribuir podría estructurarse de la siguiente manera:

- Hacer un *fork* del repositorio.
- Confirmar los cambios en el repositorio desde el cual se hizo el *fork*.

- Enviar una solicitud de subida de código a los administradores del repositorio de código.

Los conceptos son abstracciones útiles de los datos brutos que son ampliamente aplicables a las preguntas de investigación y se organizan en carpetas. Se podría fomentar que los usuarios compartiesen estos 'conceptos' que han extraído escribiendo un código que genere una vista materializada. Estas vistas materializadas pueden ser utilizadas por todos los investigadores con acceso a las bases de datos permitiendo acelerar la extracción de datos de toda la comunidad.

8 Conclusiones

A lo largo de las secciones de este documento se presenta una recopilación del proceso que se sigue para el análisis de datos biomédicos de una forma integrada, desde el descubrimiento inicial de los datos de interés (genómicos y/o clínicos, y/o de imagen), seguida de los procedimientos para su solicitud, y finalmente el análisis de los datos, donde se discute as posibilidades distintas posibilidades de análisis, así como algunas de las plataformas y sistemas de análisis de datos existente.

Otro aspecto de gran interés actualmente es el análisis federado de datos, que permite reunir cohortes mayores sin necesidad de disponer de repositorios centralizados. Se comenta el concepto de entornos de investigación de confianza. También se discuten otros aspectos importantes como los repositorios de código.

El documento pretende ofrecer un marco general para el análisis integrado de combinaciones de datos clínicos, genómicos y de imagen explorando las herramientas disponibles y sugiriendo un protocolo de actuación dentro del proyecto IMPaCT-data. Es importante tener en cuenta que se trata de un campo muy dinámico y cambiante, y continuamente aparecen nuevos algoritmos de análisis de datos, así como nuevas aplicaciones y plataformas. Por ello, las ideas generales de cómo se procede con los datos serán razonablemente válidas en los próximos años, pero las aplicaciones específicas que se mencionan podrían no ser de uso común en poco tiempo, o aparecer nuevas aplicaciones que resuelvan problemas aquí planteados de formas innovadoras. Es evidente que no nos vamos a aburrir en un futuro próximo.

Referencias

1. Reglamento General de Protección de Datos. Available online: https://rgpd.es/ (accessed on 31-10-2022).
2. GA4GH Beacon project. Available online: https://beacon-project.io/ (accessed on 31-10-2022).
3. Rueda, M.; Ariosa, R.; Moldes, M.; Rambla, J. Beacon v2 Reference Implementation: a toolkit to enable federated sharing of genomic and phenotypic data. *Bioinformatics* **2022**, *38*, 4656-4657.
4. Beacon v2 Documentation. Available online: http://docs.genomebeacons.org/models/#introduction (accessed on 31-10-2022).
5. Iancu, I.F.; Perea-Romero, I.; Núñez-Moreno, G.; de la Fuente, L.; Romero, R.; Ávila-Fernandez, A.; Trujillo-Tiebas, M.J.; Riveiro-Álvarez, R.; Almoguera, B.; Martín-Mérida, I.; et al. Aggregated Genomic Data as Cohort-Specific Allelic Frequencies can Boost Variants and Genes Prioritization in Non-Solved Cases of Inherited Retinal Dystrophies. *Int J Mol Sci* **2022**, *23*, doi:10.3390/ijms23158431.
6. Beacon network. A global search engine for genetic mutations. Available online: https://beacon-network.org/ (accessed on 31-10-2022).
7. GPAP. Available online: https://platform.rd-connect.eu/ (accessed on 31-10-2022).
8. Beyond 1 Million Genomes. Available online: https://b1mg-project.eu/ (accessed on 31-10-2022).
9. GPAP playground. Available online: https://playground.rd-connect (accessed on 31-10-2022).
10. Shringarpure, S.S.; Bustamante, C.D. Privacy risks from genomic data-sharing beacons. *The American Journal of Human Genetics* **2015**, *97*, 631-646.
11. Life Science Login. Available online: https://lifescience-ri.eu/ls-login/ls-aai-aup.html (accessed on 31-10-2022).
12. GA4GH Passports. Available online: https://github.com/ga4gh-duri/ga4gh-duri.github.io/blob/master/researcher_ids/ga4gh_passport_v1.md#ga4gh-passport (accessed on 31-10-2022).
13. Human genomic and phenotypic synthetic data for the study of rare diseases. Available online: https://ega-archive.org/studies/EGAS00001005702 (accessed on 31/10/2022).
14. Muñoyerro-Muñiz, D.; Goicoechea-Salazar, J.A.; García-León, F.J.; Laguna-Téllez, A.; Larrocha-Mata, D.; Cardero-Rivas, M. Conexión de registros sanitarios: base poblacional de salud de Andalucía. *Gaceta Sanitaria* **2019**, *34*, 105-113, doi:https://doi.org/10.1016/j.gaceta.2019.03.003.
15. Viral Beacon. Available online: https://inb-elixir.es/news/viral-beacon-beacon-ocean-sars-cov-2-data (accessed on 31-10-2022).
16. Wenger, A.M.; Guturu, H.; Bernstein, J.A.; Bejerano, G. Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. *Genetics in Medicine* **2017**, *19*, 209.
17. Boycott, K.M.; Hartley, T.; Biesecker, L.G.; Gibbs, R.A.; Innes, A.M.; Riess, O.; Belmont, J.; Dunwoodie, S.L.; Jojic, N.; Lassmann, T.; et al. A Diagnosis for All Rare Genetic Diseases: The Horizon and the Next Frontiers. *Cell* **2019**, *177*, 32-37.

18. Rehm, H.L.; Bale, S.J.; Bayrak-Toydemir, P.; Berg, J.S.; Brown, K.K.; Deignan, J.L.; Friez, M.J.; Funke, B.H.; Hegde, M.R.; Lyon, E. ACMG clinical laboratory standards for next-generation sequencing. *Genetics in medicine* **2013**, *15*, 733.
19. Lek, M.; Karczewski, K.J.; Minikel, E.V.; Samocha, K.E.; Banks, E.; Fennell, T.; O'Donnell-Luria, A.H.; Ware, J.S.; Hill, A.J.; Cummings, B.B.; et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **2016**, *536*, 285, doi:10.1038/nature19057.
20. Ng, S.B.; Buckingham, K.J.; Lee, C.; Bigham, A.W.; Tabor, H.K.; Dent, K.M.; Huff, C.D.; Shannon, P.T.; Jabs, E.W.; Nickerson, D.A.; et al. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **2010**, *42*, 30-35, doi:10.1038/ng.499.
21. Peña-Chilet, M.; Roldán, G.; Perez-Florido, J.; Ortuño, F.M.; Carmona, R.; Aquino, V.; Lopez-Lopez, D.; Loucera, C.; Fernandez-Rueda, J.L.; Gallego, A.; et al. CSVS, a crowdsourcing database of the Spanish population genetic variability. *Nucleic Acids Research* **2020**, *49*, D1130-D1137, doi:10.1093/nar/gkaa794.
22. Abecasis, G.R.; Auton, A.; Brooks, L.D.; DePristo, M.A.; Durbin, R.M.; Handsaker, R.E.; Kang, H.M.; Marth, G.T.; McVean, G.A. An integrated map of genetic variation from 1,092 human genomes. *Nature* **2012**, *491*, 56-65, doi:10.1038/nature11632.
23. Nelson, M.R.; Wegmann, D.; Ehm, M.G.; Kessner, D.; St Jean, P.; Verzilli, C.; Shen, J.; Tang, Z.; Bacanu, S.A.; Fraser, D.; et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **2012**, *337*, 100-104, doi:10.1126/science.1217876.
24. Mathieson, I.; McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* **2012**, *44*, 243-246, doi:10.1038/ng.1074.
25. Corona, E.; Chen, R.; Sikora, M.; Morgan, A.A.; Patel, C.J.; Ramesh, A.; Bustamante, C.D.; Butte, A.J. Analysis of the genetic basis of disease in the context of worldwide human relationships and migration. *PLoS Genet* **2013**, *9*, e1003447, doi:10.1371/journal.pgen.1003447.
26. Dopazo, J.; Amadoz, A.; Bleda, M.; Garcia-Alonso, L.; Alemán, A.; García-García, F.; Rodriguez, J.A.; Daub, J.T.; Muntané, G.; Rueda, A.; et al. 267 Spanish Exomes Reveal Population-Specific Differences in Disease-Related Genetic Variation. *Molecular Biology and Evolution* **2016**, *33*, 1205-1218, doi:10.1093/molbev/msw005.
27. Bustamante, C.D.; Burchard, E.G.; De la Vega, F.M. Genomics for the world. *Nature* **2011**, *475*, 163-165, doi:10.1038/475163a.
28. Smetana, J.; Brož, P. National Genome Initiatives in Europe and the United Kingdom in the Era of Whole-Genome Sequencing: A Comprehensive Review. *Genes* **2022**, *13*, 556.
29. UCSC Beacon. Available online: http://ucscbeacon.clinbioinfosspa.es (accessed on 31-10-2022).
30. European Society of Radiology. ESR position paper on imaging biobanks. *Insights into imaging* **2015**, *6*, 403-410.
31. Banco de Imagen de la Comunidad Valenciana Available online: https://bimcv.cipf.es/ (accessed on 31-10-2022).
32. Tomczak, K.; Czerwińska, P.; Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology* **2015**, *19*, A68.
33. EuroBioimaging. Available online: https://www.eurobioimaging.eu/data/bia (accessed on 31-10-2022).
34. OpenSlide. Available online: https://openslide.org/ (accessed on 31-10-2022).

35. Visor WSI INIBICA. Available online: https://github.com/INiBICA-CO07/Visualization-WSI-tool (accessed on 31-10-2022).
36. Smith, B.; Hermsen, M.; Lesser, E.; Ravichandar, D.; Kremers, W. Developing image analysis pipelines of whole-slide images: Pre-and post-processing. *Journal of Clinical and Translational Science* **2021**, *5*.
37. Beales, T. Constraint-based domain models for future-proof information systems. *The Good Electronic Health Record Project (www gehr org), Australia* **2001**.
38. Kalra, D.; Beale, T.; Heard, S. The openEHR foundation. *Studies in health technology and informatics* **2005**, *115*, 153-173.
39. 12b2 Informatics for Integrating Biology & the Bedside. Available online: https://www.i2b2.org (accessed on 31-10-2022).
40. Conde, J.M.; Moreno-Conde, A.; Salas-Fernández, S.; Parra-Calderón, C.L. ITCBio, a Clinical and Translational Research Platform. *AMIA Annu Symp Proc* **2019**, *2019*, 673-680.
41. Pedrera-Jimenez, M.; Garcia-Barrio, N.; Hernandez-Ibarburu, G.; Baselga, B.; Blanco, A.; Calvo-Boyero, F.; Gutierrez-Sacristan, A.; Quiros, V.; Cruz-Bermudez, J.L.; Bernal, J.L.; et al. Building an i2b2-Based Population Repository for COVID-19 Research. *Stud Health Technol Inform* **2022**, *294*, 287-291, doi:10.3233/shti220460.
42. Carmona-Pírez, J.; Poblador-Plou, B.; Poncel-Falcó, A.; Rochat, J.; Alvarez-Romero, C.; Martínez-García, A.; Angioletti, C.; Almada, M.; Gencturk, M.; Sinaci, A.A.; et al. Applying the FAIR4Health Solution to Identify Multimorbidity Patterns and Their Association with Mortality through a Frequent Pattern Growth Association Algorithm. *Int J Environ Res Public Health* **2022**, *19*, doi:10.3390/ijerph19042040.
43. Jiang, G.; Kiefer, R.C.; Rasmussen, L.V.; Solbrig, H.R.; Mo, H.; Pacheco, J.A.; Xu, J.; Montague, E.; Thompson, W.K.; Denny, J.C.; et al. Developing a data element repository to support EHR-driven phenotype algorithm authoring and execution. *J Biomed Inform* **2016**, *62*, 232-242, doi:10.1016/j.jbi.2016.07.008.
44. Programa de analítica de datos para la investigación y la innovación en salud (PADRIS). Available online: https://aquas.gencat.cat/ca/ambits/analitica-dades/padris/ (accessed on 31-10-2022).
45. Plataforma para la generación segura de conocimiento biomédico a partir del big data clínico de la base poblacional de salud. Available online: https://www.clinbioinfosspa.es/projects/iRWD/indexEsp.html (accessed on 31-10-2022).
46. CCEIBA. Available online: http://si.easp.es/eticaysalud/content/comite-coordinador-etica-investigacion-biomedica-andalucia (accessed on 31-10-2022).
47. Data Access Committee. Available online: https://ega-archive.org/submission/data_access_committee (accessed on 31-10-2022).
48. García-León, F.; Villegas-Portero, R.; Goicoechea-Salazar, J.; Muñoyerro-Muñiz, D.; Dopazo, J. Impact assessment on data protection in research projects. *Gaceta sanitaria* **2020**, *34*, 521-523.
49. El Emam, K.; Jonker, E.; Arbuckle, L.; Malin, B. A systematic review of re-identification attacks on health data. *PloS one* **2011**, *6*, e28071.
50. 10 malentendidos relacionados con la anonimización. Available online: https://www.aepd.es/es/documento/10-malentendidos-anonimizacion.pdf (accessed on

51. Building Trusted Research Environments - Principles and Best Practices; Towards TRE ecosystems. Available online: https://zenodo.org/record/5767586#.Y2PvmHbMKUk (accessed on 31-10-2022).
52. Burgess, D.J. Spatial transcriptomics coming of age. *Nature Reviews Genetics* **2019**, *20*, 317-317.
53. Vihinen, M. No more hidden solutions in bioinformatics. *Nature* **2015**, *521*, 261-261, doi:10.1038/521261a.
54. Murdoch, W.J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* **2019**, *116*, 22071-22080.
55. Wolstencroft, K.; Haines, R.; Fellows, D.; Williams, A.; Withers, D.; Owen, S.; Soiland-Reyes, S.; Dunlop, I.; Nenadic, A.; Fisher, P.; et al. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Res* **2013**, *41*, W557-561, doi:10.1093/nar/gkt328.
56. Goecks, J.; Nekrutenko, A.; Taylor, J. T.(2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol 11*, R86.
57. Stropp, T.; McPhillips, T.; Ludäscher, B.; Bieda, M. Workflows for microarray data processing in the Kepler environment. *BMC bioinformatics* **2012**, *13*, 1-15.
58. Gentleman, R.C.; Carey, V.J.; Bates, D.M.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* **2004**, *5*, 1-16.
59. Medina, I.; Salavert, F.; Sanchez, R.; de Maria, A.; Alonso, R.; Escobar, P.; Bleda, M.; Dopazo, J. Genome Maps, a new generation genome browser. *Nucleic Acids Res* **2013**, *41*, W41-46, doi:10.1093/nar/gkt530.
60. cBioPortal. Available online: http://cbioportal.org (accessed on 31-10-2022).
61. Interactive Variant Analysis (IVA). Available online: http://docs.opencb.org/display/iva (accessed on 31-10-2022).
62. Aleman, A.; Garcia-Garcia, F.; Medina, I.; Dopazo, J. A web tool for the design and management of panels of genes for targeted enrichment and massive sequencing for clinical applications. *Nucleic Acids Res* **2014**, *42*, W83-87, doi:10.1093/nar/gku472.
63. Aleman, A.; Garcia-Garcia, F.; Salavert, F.; Medina, I.; Dopazo, J. A web-based interactive framework to assist in the prioritization of disease candidate genes in whole-exome sequencing studies. *Nucleic Acids Res* **2014**, *42*, W88-93, doi:10.1093/nar/gku407.
64. OpenCGA. Available online: http://docs.opencb.org/display/opencga (accessed on 31-10-2022).
65. OpenCB project. Available online: http://docs.opencb.org/ (accessed on 31-10-2022).
66. PrioR prioritization tool. Available online: https://github.com/TBLabFJD/PriorR (accessed on 31-10-2022).
67. PriorR. Variant Filtering and Prioritization. Available online: https://hub.docker.com/r/tblabfjd/priorr (accessed on 31-10-2022).
68. Smedley, D.; Jacobsen, J.O.B.; Jäger, M.; Köhler, S.; Holtgrewe, M.; Schubach, M.; Siragusa, E.; Zemojtel, T.; Buske, O.J.; Washington, N.L.; et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nature Protocols* **2015**, *10*, 2004, doi:10.1038/nprot.2015.124.

69. de la Fuente, L.; Del Pozo-Valero, M.; Perea-Romero, I.; Blanco-Kelly, F.; Ayuso, C.; Mínguez, P. Prioritization of new candidate genes for rare genetic diseases by a disease-aware evaluation of heterogeneous molecular networks. *medRxiv* **2022**, 2022.2010.2007.22280759, doi:10.1101/2022.10.07.22280759.
70. Matalonga, L.; Hernández-Ferrer, C.; Piscia, D.; Schüle, R.; Synofzik, M.; Töpf, A.; Vissers, L.; de Voer, R.; Tonda, R.; Laurie, S.; et al. Solving patients with rare diseases through programmatic reanalysis of genome-phenome data. *Eur J Hum Genet* **2021**, *29*, 1337-1347, doi:10.1038/s41431-021-00852-7.
71. The European Health Data & Evidence Network Portal. Available online: https://portal.ehden.eu/ (accessed on 31-10-2022).
72. The Observational Health Data Sciences and Informatics. Available online: https://www.ohdsi.org/ (accessed on 31-10-2022).
73. Wilson, G.; Aruliah, D.A.; Brown, C.T.; Hong, N.P.C.; Davis, M.; Guy, R.T.; Haddock, S.H.; Huff, K.D.; Mitchell, I.M.; Plumbley, M.D. Best practices for scientific computing. *PLoS biology* **2014**, *12*, e1001745.

Acrónimos y Abreviaturas

En la siguiente tabla se incluyen los acrónimos y abreviaturas utilizadas en el entregable, los términos deben estar ordenados alfabéticamente por la primera columna.

AAI	Infraestructura de Autenticación y Autorización
AM	Modelo de arquetipos
API	Application Programming Interfaces
AQL	Archetype Query Language
AWS	Amazon Web Services
BAM	Binary Alignment Map
B1MG	Beyond 1 Million Genomes
BPS	Base poblacional de Salud
CEI	Comité ético de investigación
CIBERER	Centro de investigación Biomédica En Red – Enfermedades Raras
CSVS	Collaborative Spanish Variant Server
DAC	Data Access Committee
DICOM	Digital Imaging and Communication in Medicine
EGA	European Genome Archive
ENA	European Nucleotide Archive
EHR	Historias de Salud Electrónicas
EIPD	Evaluación de Impacto en protección de Datos
EnoD	Programa de enfermedades raras no diagnosticadas del CIBERER
ESR	Sociedad Europea de Radiología
GA4GH	Global Alliance for Genomics and Health
GIMD	Gestión de Imagen Médica Digital de la Comunidad Valenciana
GPAP	Genome-Phenome Analysis Platform
HIS	Sistema de Información Hospitalaria
HL7	Health Level 7
HPO	Human Phenotype Ontology
IMPaCT	Infraestructura de Medicina de Precisión asociada a la Ciencia y la Tecnología
IMPaCT-Data	Programa de ciencia de datos de IMPaCT
IMPaCT-genoma	Programa de análisis genómico de IMPact
IVA	Interactive Variant Analysis
LIS	Sistema de Información de Laboratorio
MAF	Frecuencia del alelo minoritario
OMOP	Observational Medical Outcomes Partnership
PACS	Picture Archiving and Communication System
PLN	Procesamiento del Lenguaje Natural

RAID	Redundant Array of Independent Disks
RGPD	Regulación General de Protección de Datos
RIS	Radiology Information Service
RM	Modelo de referencia
SAI	Sistemas de Alimentación Ininterrumpida
SIPAT	Sistema de Información de Anatomía Patológica
SNOMED	Systematized Nomenclature of Medicine Clinical Terms
SVN	Subversion: repositorio de código
TCGA	The Cancer Genome Atlas
TRE	Trusted Research Environment
VOC	Variante de preocupación
VOI	Variante de sospecha
WGS	Whole Genome Sequening
WSI	Whole Slide Imaging
XAI	eXplainable Artificial Intelligence