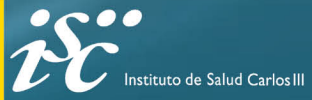
Comparación de Técnicas de Gestión de Información de HCE



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	https://impact-data.bsc.es/			
Paquete Trabajo	WP4 – Datos Médicos e Imagen			
Tarea	Tarea 4.1 – Adaptación, instalación y uso de software de código abierto para la extracción de variables a partir de HCE			
Entregable	E4.2 Comparación de Técnicas de Gestión de Información de HCE.			
Versión	1.1.1			
Fecha Entrega	30/09/2022		***Fecha Aprobación***	17/05/2023
Responsable	IACS			
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**
IACS	Javier Gómez-Arrue Azpiazu	Coordinación / Autor
IACS	Carlos Tellería Orriols	Coordinación / Autor
FPS	Joaquín Dopazo Blázquez	Revisor
H12O	Pablo Serrano Balazote	Revisor
ISS La Fe	María Eugenia Gas López	Autor
FISABIO-UMIB	Silvia Nadal Almela	Autor
FISABIO-UMIB	Mariam de la Iglesia	Autor
H12O	Miguel Pedrera Jiménez	Autor
HCB	Santiago Frid	Autor
ISS La Fe	Felix Francisco Enríquez Romero	Autor
SNS-O	Javier Gorricho Mendívil	Autor
Navarrabiomed	Julián Librero López	Autor
HUVR	Carlos Parra Calderón	Autor

Historial de versiones			
Nro.	**Fecha**	**Descripción**	**Autor**
v 0.1	26/05/2022	Borrador del índice	JGA (IACS) CTO (IACS)
v 0.2	02/09/2022	Borrador de documento para envío a revisores	Todos los autores
v 0.3	20/09/2022	Documento revisado	Joaquín Dopazo (FPS) Pablo Serrano (H12O)
v 1.0	26/09/2022	Versión final para envío a coordinación	Todos los autores
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

Figuras

Resumen Ejecutivo

El uso secundario de datos sanitarios, independientemente del uso final que se le vaya a

dar, exige acceder a los sistemas de información primarios, asistenciales y administrativos, seleccionar la información relevante, extraerla y persistirla en almacenes o lagos de datos diferenciados de los sistemas primarios, y optimizados para el uso analítico y masivo que se les va a dar. La utilidad de estos repositorios secundarios exige que éstos estén configurados de acuerdo a modelos comunes y estandarizados de datos.

La enorme disparidad de sistemas, proveedores y arquitecturas existentes en los sistemas asistenciales y administrativos de los dispositivos sanitarios, hace imposible el desarrollo de una herramienta universal para la extracción y reutilización de los datos. Sin embargo, podemos identificar unos pocos modelos arquitectónicos, y establecer recomendaciones y buenas prácticas para la extracción de información útil, y se transformación a modelos comunes de datos.

Existen en el mercado y en la comunidad de código abierto una gran cantidad de herramientas que nos permitirán cubrir el ciclo de vida completo de los datos, incluyendo la selección, extracción, transformación, análisis y visualización de los datos.

En todo este proceso, además de cuestiones de índole técnica y tecnológica, relacionadas con protocolos, estándares de normalización y arquitecturas de datos, es fundamental atender a los aspectos de interoperabilidad semántica. En este sentido, la definición y estandarización de ontologías y bibliotecas de arquetipos, facilita considerablemente el mapeo conceptual, desde los sistemas primarios a los repositorios finales, pasando por las herramientas de transformación y análisis de los datos. Avanzar en la implantación de modelos basados en arquetipos en todas las fases del ciclo de vida de uso secundario de los datos será esencial en los próximos años para mejorar y apalancar los procesos de obtención de conocimiento a partir de los datos.

Introducción

Audiencia

Este documento está destinado a todos los participantes del proyecto IMPaCT-Data, como referencia de tecnologías para la extracción, transformación y análisis de datos sanitarios, para ser utilizadas en el contexto del paquete de trabajo 4 de IMPaCT (gestión de datos e imagen clínica), así como su integración con datos genómicos (paquete 5), y su utilización en los casos de uso globales del proyecto. El documento puede ser también de utilidad para cualquier institución o grupo de investigación que quiera conocer las técnicas y herramientas más utilizadas actualmente para la extracción y reutilización de datos de salud.

Ámbito

El presente documento se empleará como referencia para la selección y uso de distintas tecnologías de interoperabilidad en los demostradores a desarrollar en el paquete 4 de IMPaCT-Data, así como en los procesos de integración global del paquete 5, y en los casos de uso globales propuestos por el paquete 6.

Relación con otros Entregables

Este entregable guarda relación con el entregable 4.1, donde se describen los principales estándares de interoperabilidad que se usan conjuntamente con las herramientas y técnicas descritas en el presente documento, y que son referenciadas frecuentemente en el mismo. Guarda relación también con el entregable 4.5, en el que se describen conceptos similares relacionados específicamente con imagen médica, y cuya información conjunta deberá ser utilizada para el desarrollo de los demostradores de integración del paquete 4.

Estructura Entregable

El presente documento se estructura en tres secciones diferenciadas, pero muy relacionadas.

En la primera sección, se analizan los distintos modelos de sistemas de información de dato primario (asistenciales), y la forma de acceder a ellos y estructurar su información de cara a un uso secundario de los datos.

En la segunda sección, se analizan los distintos modelos de procesos que, en función del tipo de arquitectura de los sistemas primarios, nos permiten extraer la información y transformarla a un modelo común de datos estandarizado.

La tercera sección es un compendio de las herramientas software para la realización de procesos de extracción, transformación, almacenamiento, análisis y presentación de resultados más utilizadas en la actualidad en el ámbito de la explotación de datos de salud, así como algunas arquitecturas de datos útiles para esta finalidad.

Por último, se exponen las conclusiones finales del análisis, así como recomendaciones para un buen diseño de arquitecturas de datos para uso secundario, y para las siguientes fases del proyecto.

1 Técnicas de gestión de la información de HCE. Acceso a fuentes primarias.

La utilización de información procedente de sistemas de Historia Clínica Electrónica (HCE) y otras fuentes de datos asistenciales y administrativas exige el acceso directo a los modelos de persistencia de datos que soportan estos sistemas de información, para poder realizar sobre ellos los procesos de extracción y transformación de los datos, para su posterior carga en los sistemas analíticos o en lagos de datos. Estos procesos de extracción y transformación, que pueden incluir tareas relacionadas con la calidad y normalización de datos, precisan de la definición de mapeos que relacionen cada dato extraído de una fuente en una posición concreta de la estructura de datos original con uno o más datos, replicados o transformados, en la estructura de datos del sistema de destino.

En todo caso, y esto es común e indiferente a la arquitectura o modelo de datos que soporte los datos de salud para su uso primario (la asistencia sanitaria), el principal problema del uso secundario de los datos (analítico, agregado, poblacional) es precisamente ése, que su uso es secundario. Eso no quiere decir que su uso sea menos importante, sino que el uso que se le da no es aquél para el que y por el que fue recogido. Cuando el uso analítico y agregado se realiza sobre datos recogidos explícitamente para ello (estudios de cohortes concretas, ensayos clínicos, etc.), la calidad y completitud de los datos suele estar contemplada, por lo general, en el diseño del estudio. Sin embargo, cuando realizamos un uso secundario de dato recogido para la práctica clínica, no debemos olvidar cómo y para qué se recogió ese dato. El dato suele estar con frecuencia recogido por un facultativo u otro personal sanitario para la atención específica de un paciente concreto (información de contexto), y para ser interpretado por humanos, con frecuencia el mismo humano que recogió y registró la información. Al hacer un uso secundario de la información, por un lado la información se va a procesar por máquinas, no por humanos, y por otro lado perdemos con frecuencia la información de contexto de los datos (¿en qué circunstancias se recogió un dato?, ¿en qué contexto patológico y asistencial?, ¿por qué razón no se recogió determinada información?, ¿por irrelevancia o porque se sobreentendía del contexto? ...).

Gran parte de estas incertidumbres son las que se intentan paliar con algunas técnicas avanzadas, como son la minería de datos y el procesamiento de lenguaje natural, y sobre todo mediante estándares de normalización de la información, extendiendo esta estandarización a las capas de codificación y clasificación de conceptos, modelos de persistencia de datos, y modelos de conceptualización del dominio de la información sanitaria.

Desgraciadamente, en los sistemas de información del mundo real, no hay un único

estándar en lo relativo a modelo de persistencia de datos, modelo de conocimiento, o esquemas terminológicos y de clasificación para registrar la información sanitaria. Esta realidad supone la imposibilidad de definir procesos o herramientas universales que nos permitan el acceso sencillo y homogéneo a las fuentes primarias de información sanitaria. El acceso a los datos primarios debe diseñarse e implementarse de forma específica y única en cada organización.

No obstante, sí es posible definir varios modelos genéricos de sistemas de información clínicos que, aunque no nos permitan en general el diseño de sistemas de extracción y carga directamente reutilizables, sí nos permitirán identificar algunas herramientas útiles, y sobre todo definir unas metodologías comunes que poder utilizar como guía a la hora de abordar un proyecto de acceso a datos primarios de salud, para su reutilización en un contexto de uso secundario.

1.1 Acceso a información sobre modelos propietarios

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:

- Los procesos de extracción de datos tendrán que diseñarse específicamente para cada una de las instancias, centros o servicios donde se implementen, haciendo que la reutilización de procesos ETL sea imposible en la práctica.

- El conocimiento de la estructura interna de los modelos de datos, las relaciones y dependencias entre entidades de información, es imprescindible para realizar una correcta interpretación de la información capturada. Es más, a la hora de interpretar correctamente el significado de un dato leído de determinado campo en determinada tabla del modelo de datos subyacente, es necesario con mucha frecuencia, conocer dónde y cómo se captura ese dato en el interfaz de usuario final del sistema de información. Es frecuente que, por distintas razones durante el ciclo de vida de las aplicaciones informáticas, la denominación de un determinado ítem de información no coincida entre lo explicitado en el interfaz de usuario y el nombre del campo y tabla donde ese dato se persiste, lo que puede llevar a cometer errores de interpretación de la información que estamos capturando y transformando, y que queremos reutilizar.

- Todo esto implica que sea necesario que los técnicos TIC responsables de la

explotación de los sistemas primarios (servicios de informática del hospital o servicio de salud), e incluso de la empresa u organismo que ha desarrollado o mantiene el aplicativo asistencial, participen activamente en la definición de las consultas que se vayan a utilizar para la extracción y transformación de la información entre los sistemas de uso primario y secundario de datos de salud.

En lo referente al contenido y normalización de la información, la situación real en los sistemas primarios no estandarizados (desarrollos propietarios y a medida) es algo mejor que en lo relativo a estructuras de datos, pero solo algo mejor. Estos sistemas han sido diseñados y creados en su mayoría en contextos en los que la interoperabilidad y la estandarización de la información no eran necesidades identificadas, lo que ha llevado a que, con demasiada frecuencia, las codificaciones y clasificaciones utilizadas en los mismos fueran también propietarias o particulares. En algunos casos, en los que un determinado servicio de salud (por ejemplo, un servicio autonómico) ha utilizado una misma solución tecnológica en todos los hospitales de la comunidad, no ha mantenido criterios comunes de codificación, dejando libertad a cada centro a elaborar y codificar múltiples conceptos, tales como la estructura organizativa del hospital, la cartera de servicios, o conceptos como los motivos de alta o ingreso.

Es cierto que, en la medida en la que la interoperabilidad de sistemas informáticos se ha ido extendiendo en el dominio de los sistemas de información sanitarios, muchas organizaciones han abordado proyectos de renormalización y recodificación de sus bases de datos, especialmente en dominios complejos y extensos como los diagnósticos, procedimientos o determinaciones analíticas. Pero todavía queda mucho camino por andar en esos sistemas propietarios hacia la interoperabilidad completa.

Cuando este trabajo no se ha llevado a cabo, o se ha hecho solo parcialmente, será preciso diseñar procesos de extracción y transformación de datos que realicen una recodificación de la información, de manera que los datos persistidos en el repositorio o lago de datos de uso secundario sí estén normalizados y homogeneizados.

Hay ocasiones en las que resulta muy costoso capturar y recodificar toda la información procedente de un sistema primario concreto, y se decide realizar la captura de forma incremental, recodificando inicialmente un conjunto mínimo de códigos que suponga un alto porcentaje de los registros, y recodificando paulatinamente el resto de códigos que, siendo mayor en número, se repiten menos veces en el conjunto de datos original. Pongamos como ejemplo un sistema de gestión de laboratorio (LIS) que no utilice códigos normalizados para identificar las determinaciones analíticas, y supongamos que queremos capturar la información de las analíticas sobre un repositorio codificado con LOINC. Si las distintas determinaciones recogidas en el LIS son muchas (típicamente varios miles si el sistema lleva algunos años funcionando), recodificar todas las determinaciones de inicio puede ser un trabajo arduo y costoso. Sin embargo, un rápido conteo de resultados puede hacernos ver fácilmente que apenas 100 o 200

determinaciones distintas pueden suponer entre un 60 y 70 por ciento de todos los resultados. Se puede empezar por codificar estas 200 determinaciones, e ir incrementando ese número en la medida en que van siendo necesarias analíticas menos frecuentes y más específicas, en un proceso iterativo de mejora continua.

Todas estas apreciaciones nos llevan a tener en cuenta una serie de implicaciones:

- Los procedimientos de extracción y transformación de los datos tendrán que diseñarse conjuntamente con los servicios técnicos de los hospitales y servicios de salud, y también y muy especialmente con los responsables funcionales de sistemas de información, que suelen ser los responsables de los procesos de codificación y normalización de los sistemas.

- Los procesos ETL tendrán que contemplar mecanismos de garantía de calidad de datos, de validación de códigos contra terminologías y clasificaciones estandarizadas, y de recodificación de códigos no estándar (muchas veces códigos específicos y únicos de una instalación concreta) hacia códigos normalizados.

- Los procesos ETL deberán diseñarse de forma que puedan ser ejecutados múltiples veces sobre los mismos datos, bien porque en un proceso de mejora continua de los sistemas primarios, se decide estandarizar algún concepto no normalizado inicialmente, bien por un planteamiento incremental en el proceso ETL.

1.2 Información sobre modelos basados en arquetipos

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 [1].

La arquitectura de modelo dual, propuesta en 2000 por Thomas Beale [2], propone una separación clara entre los modelos de información y de conocimiento. En los modelos de información clásicos, entidad-relación u orientados a objetos, el modelo de conocimiento -las entidades conceptuales con las que construimos el sistema: pacientes, centros sanitarios, visitas, episodios, diagnósticos, prescripciones...- están directamente implementadas sobre el modelo de información -objetos software, tablas y campos de las bases de datos...-, lo cual establece una dependencia enorme entre ambos niveles. En dominios tan complejos y variables como es el sanitario, esta dependencia lleva a una complejidad enorme del modelo de información, y a un costosísimo mantenimiento del sistema. Al estar el conocimiento codificado en el sistema -software, bases de datos-, cualquier modificación o ampliación del modelo de conocimiento -por ejemplo, empezar a

registrar información de un servicio hospitalario nuevo-, implica recodificar y alterar la base de datos y el software, y redesplegar una nueva versión del sistema.

La arquitectura de modelo dual trata de resolver estos problemas separando por un lado el modelo de referencia (RM), que contiene las entidades básicas capaces de representar cualquier información contenida en la Historia Clínica Electrónica, y el conocimiento de dominio, que se implementa mediante un modelo de arquetipos (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. Ejemplos de arquetipos pueden ser un resultado analítico de glucosa en sangre, un antecedente familiar, un diagnóstico o una prescripción farmacéutica. En el sistema de Historia Clínica, los arquetipos se agrupan en plantillas, que representan actos o entidades clínicas complejas. Por ejemplo, un informe de alta hospitalaria puede ser una plantilla que integre una serie de arquetipos: paciente, diagnósticos, procedimientos realizados, estancia, medicación administrada, recomendaciones al alta, etc.

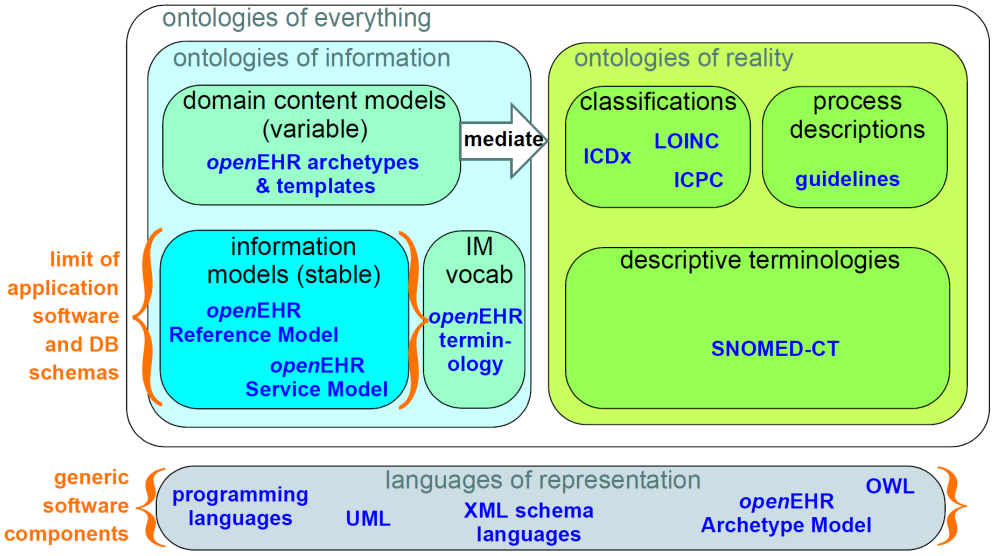


Figura 1. Arquitectura dual según OpenEHR

Es decir, los sistemas basados en arquitecturas duales permiten estructurar y modelar un conocimiento complejo y rápidamente cambiante, y gestionarlo en un modelo de información (software y esquemas de bases de datos) sencillo y estático, definido por el modelo de referencia.

La ventaja de los arquetipos, además de lo indicado, es que permiten que la información clínica y su significado (a través de la vinculación terminológica) sea fácilmente procesable de forma automática por sistemas informáticos, y por tanto permite la interoperabilidad semántica entre sistemas, al trabajar directamente sobre un dominio semántico, independientemente del modelo sintáctico o técnico utilizado en cada sistema. Además, dada la separación entre modelos de información y de conocimiento,

el primero es responsabilidad fundamental de los desarrolladores y responsables TIC, mientras que el modelo de conocimiento, es decir, el diseño de arquetipos y plantillas, puede ser perfectamente realizado por expertos del dominio sanitario, con la ayuda de sencillas herramientas que faciliten el diseño de los mismos.

Entre estas herramientas, cabe destacar el **Archetype Definition Language** (ADL), propuesto y desarrollado por OpenEHR [3] y utilizado también por otros modelos de arquitectura dual [4] es un lenguaje formal para expresar arquetipos, cuya sintaxis es una de las serializaciones posibles de los arquetipos. ADL está diseñado para ser legible tanto por personas como por sistemas informáticos. Puede ser usado para escribir arquetipos de cualquier dominio en el cual existan modelos formales que describan instancias de datos, aunque se ha desarrollado y utilizado fundamentalmente en el dominio sanitario. Los arquetipos son neutrales respecto del idioma utilizado, y pueden tener autoría y traducciones a cualquier idioma. Así, utilizando ADL se pueden modelar arquetipos utilizando estándares de información sanitaria específicos en cualquier ámbito y ubicación. A medida que el experto va creando arquetipos que modelan conceptos clínicos, éstos se van guardando en un repositorio o catálogo de arquetipos, que el experto puede utilizar, como si se tratara de piezas de LEGO®, para crear arquetipos más complejos o plantillas que desplegar en el sistema de Historia Clínica Electrónica.

En este tipo de sistemas, los arquetipos son los que modelan el conocimiento de dominio. Por tanto, en cualquier escenario de interoperabilidad, ya sea a nivel HCE-HCE, o a la hora de extraer información de sistemas asistenciales para llevarla a sistemas de uso secundario, lo fundamental es trabajar en ambos extremos con el mismo modelo de arquetipos, y no tanto que los modelos de datos o el software utilizado sean compatibles. La implementación de un proceso de interoperabilidad entre modelos de información sanitaria basados en arquetipos tendrá por tanto dos opciones. La primera consiste en la selección de arquetipos estándares ya creados y catalogados, y la alternativa es la elaboración de arquetipos propios a partir de acuerdos entre los miembros de las organizaciones que compartirán la información, cuando no existan arquetipos normalizados previamente.

En los casos en los que la información sanitaria no está modelada con arquetipos clínicos en origen (integración entre sistemas propietarios y sistemas basados en arquetipos), los arquetipos sirven como plantillas que definen qué datos deben ser extraídos de las distintas fuentes y, de requerirse, qué transformaciones deben llevarse a cabo para cumplir con las definiciones de los arquetipos.

Así, es fundamental entender que estas especificaciones, que pueden parecer complejas, son aplicadas al nivel de profundidad que la madurez tecnológica de la organización permita, y que el caso de uso demande. Los arquetipos clínicos pueden servir desde para acordar un modelo común a extraer de múltiples sistemas heterogéneos en sus formatos de origen (y posteriormente procesar con técnicas

convencionales) hasta, en escenarios complejos, construir una plataforma de datos para usos primarios y/o secundarios [5].

1.3 Lagos de datos

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.

Si un sistema de información contiene datos que están estructurados, normalizados y relacionados, entonces estamos hablando de un *data warehouse* o almacén de datos. En este caso, los datos sí se pueden consumir sin necesidad de un procesamiento adicional. Es habitual que, cuando se habla de lago de datos, en realidad nos estemos refiriendo a almacenes de datos masivos que recogen información procedente de fuentes diversas. Siguiendo la terminología utilizada clásicamente en los sistemas de inteligencia empresarial (BI), los lagos de datos se corresponden con los *Operational Data Store* (ODS) intermedios entre las fuentes primarias y los almacenes de datos. La principal diferencia entre un ODS y un lago de datos es que, mientras el ODS suele ser efímero y su contenido se descarta una vez estructurados y normalizados los datos en el almacén, el lago de datos tiene carácter permanente. La segunda diferencia es que el ODS no contempla en ningún caso la captura de datos desestructurados, mientras que el lago de datos admite cualquier tipo de información.

Últimamente se están desarrollando e implementando soluciones que tratan de superar las limitaciones que tienen los lagos de datos, derivadas de la compleja gobernanza de los datos, y sobre todo de la carencia de sistemas informáticos capaces de gestionar de una forma sencilla y homogénea los distintos tipos de datos. Esta carencia lleva en la práctica a que los lagos de datos sean infraestructuras formadas por múltiples productos interconectados, cada uno de los cuales es capaz de gestionar un tipo de dato o un proceso concreto dentro de la infraestructura. Estos nuevos sistemas que van ganando presencia en el mundo de la gestión de datos se denominan *lakehouses*. Un *lakehouse* es un sistema basado en una arquitectura que combina los mejores elementos de lagos y almacenes de datos. Son soluciones que se implementan sobre almacenamiento de bajo coste en la nube, y usando formatos abiertos. Probablemente los *lakehouses* sean el futuro más o menos inmediato de los almacenes masivos de datos para uso analítico,

pero no hay de momento instancias de *lakehouse* en el contexto de los sistemas sanitarios que podamos utilizar como referencia a día de hoy.

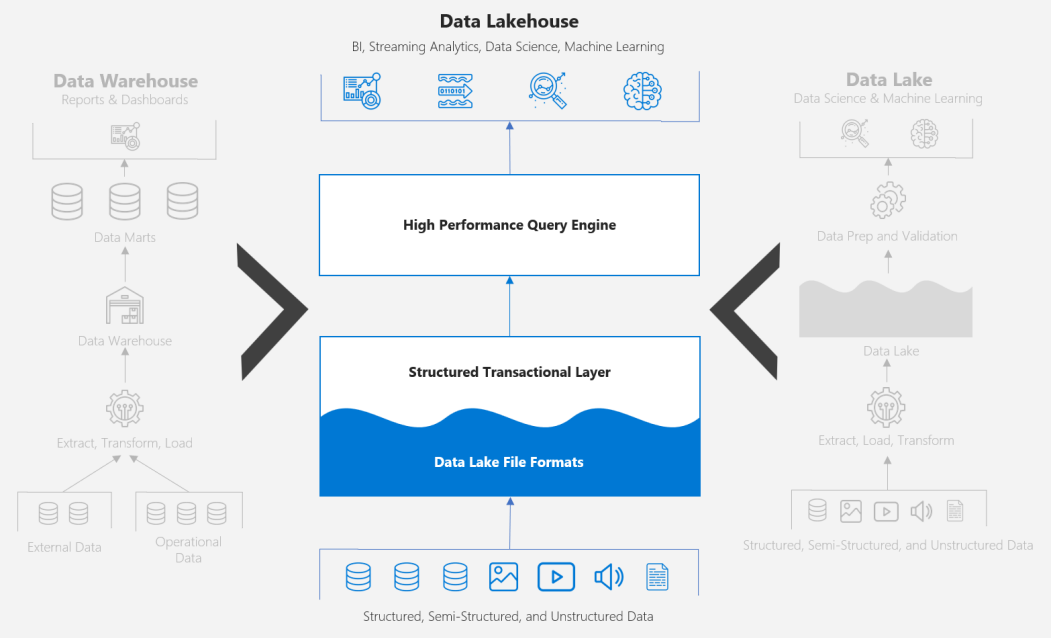


Figura 2. Arquitectura Data Lakehouse[1]

Las cantidades masivas de datos en el sector sanitario crecen a gran velocidad y de forma exponencial. Por este motivo, el almacenamiento y la calidad de los datos médicos suponen un gran reto para el personal sanitario, pacientes y futuros ensayos clínicos. Los *data lakes* o *data warehouses* son las infraestructuras más utilizadas para recoger, almacenar y analizar grandes cantidades de datos de distintos formatos, y desacoplarlos de los sistemas de información operacionales (asistenciales o administrativos). Esta agrupación de datos de distintas fuentes hace incrementar los errores en los datos, la duplicación de estos y otras inconsistencias que obligan a un proceso de limpieza de los datos (*data cleaning*) previo a su uso o análisis.

La implementación de estas infraestructuras intermedias masivas (ya sean almacenes de datos, lagos o *lakehouses*) implica el desarrollo de procesos ETL adicionales. Por un lado, se ejecutarán procesos ETL para capturar toda la información posible de los sistemas asistenciales y administrativos a fin de alimentar el almacén masivo, y por otro lado, necesitaremos procesos que, a partir del almacén masivo, alimenten modelos comunes de datos estructurados o herramientas de inteligencia de negocio sobre modelos de datos propios.

No obstante, las ventajas de disponer de este tipo de infraestructuras de datos son claras.

[1] CC https://www.taygan.co/blog/2022/05/15/data-lakehouse

- Se desacoplan totalmente los sistemas operacionales de los analíticos. Las infraestructuras de datos masivos se diseñan pensando en una ingesta de datos muy rápida, lo que minimiza el impacto sobre los sistemas primarios. Los procesos ETL diseñados para esta tarea se pueden ejecutar en ciclo rápido de refresco (diario), aunque se podrán adaptar a otras frecuencias en función de las características de la fuente de datos, y se ejecutan normalmente en ventanas horarias de mínimo impacto, independientemente de cuándo se necesitan los resultados analíticos.
- Se dispone en un mismo espacio de datos procedentes de fuentes diversas de datos. Si bien eso supone, como ya se ha comentado, riesgo de duplicación e inconsistencia de datos, también permite desarrollar procesos de limpieza de datos basados en la comparación de distintas fuentes, mejorando en conjunto la fiabilidad de los datos obtenidos. En estos procesos se realizarán tareas que mejoren y garanticen la calidad de los datos, tales como la correcta identificación de los pacientes entre distintas fuentes *(data linkage)*, la detección de duplicados, la fusión de historias y episodios clínicos, y la seudonimización de los datos personales.
- A partir de un mismo espacio de datos, es posible obtener de una forma relativamente homogénea distintos subconjuntos de datos en función de las necesidades, ya sean modelos clásicos de *data warehouse* y *data marts* (modelos en estrella o copo de nieve), o modelos comunes de datos para análisis (OMOP-CDM y similares).

Un datalake puede almacenarse en los servidores locales de la organización o en servidores en la nube. El almacenamiento en la nube destaca por su capacidad de escalar a medida que crece el volumen de datos que se pretenden almacenar sin que ello suponga una disminución en el rendimiento del sistema. Este hecho hace que cada vez más organizaciones opten por esta vía; sin embargo, dada la sensibilidad de los datos recogidos en las HCE hace que no siempre se delegue el almacenamiento de estos a terceros.

Una herramienta de código abierto que permite el despliegue y gestión de un *datalake* es *Apache Hadoop* [6], un *framework* para programar aplicaciones distribuidas que manejen cantidades masivas de datos (estructurados o no) siguiendo el paradigma map-reduce. Apache Hadoop ofrece soluciones para cubrir necesidades referentes al acceso, gestión, integración, gobernanza y seguridad de los datos.

1.4 Conclusión

Las ventajas descritas en los almacenes masivos de datos hacen de éstos una herramienta muy útil a la hora de diseñar infraestructuras de datos para uso secundario

de la información sanitaria. La realización de procesos analíticos sobre sistemas operacionales siempre está desaconsejada, por el impacto en consumo de tiempo y recursos de los sistemas operacionales (podemos ralentizar considerablemente la respuesta de estos sistemas a las consultas que un facultativo está realizando durante una atención clínica), y porque las estructuras de datos de los sistemas operacionales no están optimizadas para la realización de procesos analíticos.

La creación de modelos comunes de datos a partir de sistemas primarios, propios, propietarios o basados en arquetipos, puede ser una alternativa viable cuando los sistemas accedidos no sean muy grandes, y el impacto sobre los mismos esté muy acotado. Cuando el impacto en consumo de recursos (tiempo, CPU, memoria) afecte al normal funcionamiento de la actividad asistencial, será necesario pensar en sistemas intermedios de ingesta rápida, sobre los que realizar procesos de limpieza, normalización y estructuración de la información, ya sean lagos de datos, *Operational Data Stores* (ODS) o similares.

Finalmente, el disponer en los sistemas primarios de modelos basados en arquetipos facilita considerablemente la captura y mapeo de la información desde los sistemas primarios a cualquier sistema destino, ya que la correspondencia se realiza directamente a nivel conceptual *(interoperabilidad semántica)*, y no solo estructural. Cuando esto no sea posible, la disponibilidad de arquetipos normalizados de la información sanitaria facilitarán también la captura de información semánticamente ordenada desde cualquier sistema de información sanitario.

2 Modelos de extracción de conjuntos de datos y cohortes para uso secundario

En el proceso del uso secundario de datos sanitarios para la realización de análisis de dichos datos, independientemente de la finalidad última del análisis, suele ser necesario construir conjuntos de datos (*datasets*) en formatos tabulares o esquemas relacionales, conteniendo exclusivamente los datos necesarios para la realización del estudio o análisis. La mayoría de los algoritmos estadísticos o de aprendizaje automático que trabajan con dato estructurado, lo hacen a partir de series, vectores o matrices de datos, por lo que las estructuras tabulares suelen ser el punto de partida de casi cualquier flujo

de análisis.

En algunos casos, serán solo un conjunto de datos demográficos y clínicos de un grupo determinado de pacientes que cumplen cierto criterio de selección. En otros casos, serán combinaciones de datos más complejas, tales como la selección de una cohorte determinada de pacientes que cumplen un cierto criterio de inclusión durante un periodo determinado de tiempo, y de los que se recoge una colección amplia de datos sanitarios, todos ellos relacionados a través del seudónimo del paciente, y tal vez de otras variables de contexto.

Los estudios de cohortes (EC) consisten en conformar grupos de personas sobre las que se quiere realizar un análisis que permita un mejor entendimiento de un evento de interés (EI) así como los factores que influyen en éste. Estos estudios, además de tener un carácter observacional y analítico, son longitudinales, ya que recogen la información de cada uno de los sujetos de la cohorte a lo largo de un periodo más o menos prolongado de tiempo. Esta característica longitudinal nos permite hacer una distinción entre EC prospectivos y retrospectivos.

A la hora de definir una cohorte de pacientes para un estudio hay que tener en cuenta que lo importante en la definición de la misma son los criterios de inclusión / exclusión, y no los elementos concretos que forman parte de la misma en un determinado momento, aunque sean esos elementos concretos los que, en última instancia, se van a someter al análisis. Esto implica una serie de peculiaridades de las cohortes de pacientes:

- Una persona puede pertenecer a múltiples cohortes.
- Una cohorte puede tener 0 o más personas.
- Una persona puede entrar y salir de la cohorte múltiples veces.

Cuando en un estudio de cohortes van a participar varios nodos, cada uno con su subconjunto propio de datos (estudios multicéntricos, interautonómicos o internacionales), ya sea juntando los datos en un repositorio único o mediante algún procedimiento de análisis federado, es imprescindible que todos los nodos participantes dispongan sus datos respectivos utilizando un mismo modelo de datos, que será compartido por todos los participantes. Estos modelos comunes pueden ser esquemas de datos *ad hoc*, consensuados entre todos los participantes en el estudio, pero habitualmetne intentan ceñirse a modelos ya estandarizados, tales como i2b2 [7], Sentinel [8], o OMOP, desarrollado por la comunidad OHDSI [9]. Por esta necesidad de utilizar modelos comunes de datos, independientemente de si accedemos directamente a los sistemas primarios, o lo hacemos a través de repositorios intermedios (almacenes o lagos de datos), e independientemente del modelo de datos utilizado como origen (propietario o basado en arquetipos), será preciso diseñar y ejecutar procedimientos que extraigan y transformen los datos desde el repositorio origen hacia ese modelo común.

Este proceso variará, obviamente, dependiendo de si estamos capturando los datos desde sistemas basados en arquetipos o lo hacemos desde sistemas relacionales propietarios. Si nuestra fuente de datos es un repositorio que ya está estructurado de acuerdo con el modelo común de datos a utilizar, no serán necesarias transformaciones adicionales, y bastará con hacer una selección de los datos que se van a utilizar en el estudio.

Incluso cuando ya partimos de un repositorio estructurado según el modelo común de datos (CDM) a utilizar, y por supuesto en cualquier otro caso, la transformación de nuestros conjuntos de datos a un CDM nos resuelven la interoperabilidad sintáctica entre los nodos participantes en el estudio. Será preciso no obstante prestar una especial atención a la interoperabilidad semántica, es decir, asegurarse de que dentro del CDM compartido, los conceptos clínicos, significados, unidades y otros atributos asociados a cada uno de los datos y entidades del modelo, son también compartidos entre todos los participantes. Los repositorios basados en arquetipos simplifican considerablemente esta tarea, dado que los arquetipos conllevan una profunda carga semántica en sí mismos, pero eso no ocurre por lo general en los modelos relacionales.

2.1 Extracción a partir de sistemas basados en arquetipos

La mayor parte de las consultas a bases de datos dependen de esquemas de datos y representaciones físicas particulares. Los usuarios deben conocer el esquema de datos físicos de una base de datos en particular para escribir una consulta válida, a la vez que una consulta escrita para un esquema no suele ser útil en otros sistemas, ya que los esquemas de datos en general son distintos aunque almacenen los mismos datos.

Tal como explicamos en el apartado 1.2**¡Error! No se encuentra el origen de la r eferencia.** separan el modelo de conocimiento (arquetipos) del modelo de información (modelo de referencia), que es el utilizado para gestionar y persistir los datos físicamente. Por tanto, cuando utilizamos sistemas de información basados en arquetipos, lo ideal es poder realizar consultas directamente a nivel semántico (arquetipos de conceptos clínicos), y no a nivel físico (modelo de referencia), de manera que la consulta sea ejecutable en sistemas distintos que comparten el mismo modelo de conocimiento, aunque sus modelos de referencia sean distintos, e independientemente de la solución de persistencia física adoptada para ello (bases de datos SQL, NoSQL, ficheros, etc).

Archetype Query Language (AQL) [10] es un lenguaje de consulta declarativo desarrollado específicamente para expresar consultas utilizadas en la búsqueda y recuperación de datos almacenados en repositorios basados en arquetipos. La sintaxis es independiente de cualquier modelo de información, aplicación, lenguaje de

programación, ambiente de sistema y modelo de almacenamiento.

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. En términos prácticos, las consultas AQL en general se ejecutan a través de una API de librería o servicio, la cual probablemente provea una estructura de resultados "anotada" (es decir, con metadatos que faciliten un procesamiento eficiente de los resultados).

De esta forma, se pueden realizar consultas AQL sobre repositorios de arquetipos basados en modelos de referencia estándares que incluyan los criterios deseados para definir cohortes de pacientes. A partir de los resultados de estas consultas, se podrán realizar las transformaciones necesarias sobre los datos arquetipados para crear repositorios basados en modelos de datos comunes estandarizados. En algunos casos, debe realizarse una transformación manual ad hoc y una posterior carga de los datos, por ejemplo para alimentar bases de datos de investigación específicas (e.g. REDCap [11]) o incluso enriquecidas con metadatos para la aplicación de los principios FAIR usando herramientas de metadatación como CEDAR [12]. En otros casos, como ocurre con los modelos comunes de datos como OMOP CDM o i2b2, se ofrecen herramientas a los usuarios que facilitan la realización de dichas transformaciones (como Rabbit-in-a-hat [13], OHDSI ATLAS [14], etc.).

En definitiva, cuando partimos de modelos de información basados en arquetipos, la forma más natural de transformar los mismos consiste en realizar búsquedas semánticas sobre el repositorio mediante el uso de consultas desarrolladas en AQL, y obtener como resultado conjuntos de datos que sean conformes o fácilmente transformables a los modelos comunes de datos que necesitemos para el proyecto analítico concreto.

2.2 Extracción a partir de almacenes y lagos de datos propietarios

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 clasificaciones de enfermedades (CIE9, CIE10 OMS, CIE 10-ES, CIAP, SNOMED), o pautas de prescripción de fármacos, en múltiples sistemas de clasificación. Mientras en un modelo basado en arquetipos o en un modelo común de datos normalizados, las codificaciones están estandarizadas, y la búsqueda se realiza en el dominio semántico, en el caso de modelos propietarios corresponderá al grupo investigador mapear los criterios de selección de las cohortes a la clasificación disponible en el repositorio, que no necesariamente tiene por qué coincidir con la utilizada en la especificación formal de la cohorte o el estudio.

La organización propietaria, que dispone del conocimiento sobre el modelo de datos y el contenido de sus diversas instancias – que ha descrito en metadatos-, será la encargada de:

1. Generar los algoritmos y procedimientos de extracción, que pueden ser consultas de SQL u otros sistemas.

2. Documentar y versionar (se asume que es un proceso que requiere posiblemente más de una iteración) el algoritmo de búsqueda para tener trazabilidad y poder detectar cualquier error realizado

3. Generar la cohorte / tabla o tablas finales con los datos, garantizando que las mismas cumplen los requerimientos de la solicitud, y los requisitos legales derivados del hecho de estar trabajando con datos personales sensibles.

Dado que el objetivo de esta selección y extracción de conjuntos de datos es poder disponer de ellos en un modelo común de datos (OMOP-CDM u otro similar) consensuado entre los participantes en el estudio, será preciso mapear todos los datos extraídos desde el lago o almacén de datos hacia el modelo común de datos. Este proceso es particularmente complejo y costoso, y normalmente no basta con relacionar cada campo de cada tabla origen con un campo y tabla del modelo común de destino. Con frecuencia nos encontraremos con que los valores de una misma columna en una tabla tengan que ir a más de una columna de una tabla destino, o incluso a distintas tablas dependiendo del valor de otras columnas, o que un único valor de origen implique cumplimentar varias columnas de una tabla de destino con valores que dependan de otros parámetros de contexto.

El problema adicional, cuando partimos directamente de lagos o almacenes propietarios, es que este costoso procedimiento de emparejamiento entre modelos de datos debe

hacerse, como ya se ha indicado, de forma diferenciada para cada proyecto, siendo escasamente reutilizables los procedimientos. Este trabajo se simplifica considerablemente cuando partimos de modelos basados en arquetipos, como se ha explicado en la sección anterior, dado que el contexto semántico de estos deja menos margen a la ambigüedad. Además, al trabajar con modelos estandarizados, la reutilización de procesos y código es automática en el momento en que disponemos de un mapeo entre el modelo de arquetipos y el modelo común de datos de destino.

Pero incluso cuando partimos de tecnologías o modelos propietarios, el uso de los arquetipos clínicos hace más eficiente la gestión y la explotación de los datos. Así, acordando y formalizando previamente los modelos de información con recursos de arquitectura dual, es posible homogeneizar, incluso automatizar, los procesos ETL sobre el repositorio [15]. En estudios previos se ha demostrado que a partir de los arquetipos clínicos definidos se pueden desarrollar los scripts SQL sobre el modelo propietario, e implementarlos a través de vistas en sus respectivas bases de datos. Así mismo, si el almacén dispone de una capa que enlaza a un modelo de datos común, por ejemplo, OMOP-CDM, se pueden utilizar herramientas específicas para la generación de la cohorte como el ATLAS de OHDSI, aunque eso implica la transformación completa del lago de datos a un modelo común de datos, previo a la selección y extracción de la cohorte. Diferentes proyectos como EHDEN [16] están facilitando la disponibilidad de estas bases con estructuras comunes de datos. Este modelo lo trataremos con más detalle en las siguientes subsecciones.

La correcta documentación de todos los pasos dados respecto a la extracción de la cohorte, la trasformación de los datos (trasformación de variables cualitativas en cuantitativas, tratamiento de datos ausentes, valoraciones de la calidad del dato), con un control de versiones que permita la trazabilidad de los cambios realizados en la definición de la misma en el caso de existir, es siempre muy recomendable. En el caso de procesos diseñados y ejecutados sobre lagos y almacenes de datos propietarios, esta recomendación se convierte en una obligación.

2.3 Definición de cohortes a partir de un modelo común de datos

Una 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 el ya mencionado OMOP-CDM, gestionado por la comunidad OHDSI. En este caso, por tanto, no crearemos un repositorio en formato CDM con los datos específicos para un proyecto de análisis

concreto, sino que tendremos todos los datos posibles estructurados de acuerdo con ese CDM. Esta configuración tiene, como todas, sus ventajas y sus inconvenientes. 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.

A la hora de trabajar con lagos de datos estructurados según un CDM, existen dos aproximaciones a la hora de definir una cohorte: Aquellas basadas en reglas y las que siguen un enfoque probabilístico.

Las **definiciones basadas en reglas** hacen uso de reglas explícitas que determinan cuándo un paciente pertenece a la cohorte. Inicialmente se deben establecer los criterios de inclusión que deben de cumplir los sujetos objeto de estudio. En el caso de OHDSI no existe distinción entre criterios de inclusión y exclusión ya que éstos se formulan como criterios de inclusión. Por ejemplo “excluir personas que hayan padecido de hipertensión previamente” se formularía como “incluye personas con 0 ocurrencias de hipertensión previa”.

A la hora de definir la cohorte se debe tener en cuenta:

- El evento inicial que determina el momento de entrada a la cohorte.
- Qué criterios de inclusión se aplican a los eventos iniciales.
- Qué define el momento de salida de la cohorte.

El evento inicial define el momento de entrada a la cohorte y viene definido por eventos registrados en el CDM, tales como visitas, exposición a fármacos, procedimientos, etc. El conjunto de personas para las que se tiene un evento de entrada es conocido como cohorte de evento inicial y el momento en el que estas pasan a formar parte de la cohorte se conoce como momento de inclusión.

El criterio de inclusión se aplica sobre la cohorte de evento inicial a modo de filtro, limitando así el número de sujetos que componen el conjunto. Los pilares estándares sobre los que se construye el criterio de inclusión son: el dominio (tipo de información clínica), el conjunto de conceptos estándares que abarca los datos de la entidad clínica de interés, los atributos específicos del dominio y la lógica temporal. Aquellos sujetos en la cohorte de evento inicial que satisfacen todos los criterios de inclusión conforman la cohorte calificada.

El evento de salida de la cohorte marca cuando un sujeto deja de pertenecer a la cohorte. Factores que determinan la salida de la cohorte pueden ser: el final del periodo de observación, un intervalo de tiempo fijo relativo al evento de entrada inicial o el último evento en una secuencia de observaciones relacionadas (como la exposición persistente a un fármaco).

El diseño probabilístico de cohortes es un método basado en modelos de aprendizaje automático que permiten acelerar la selección de los sujetos de la cohorte. Al contrario que las definiciones basadas en reglas que dependen en gran medida del conocimiento del experto que diseña la cohorte acerca del área terapéutica de interés el diseño probabilístico requiere un input mínimo por parte del experto [17].

Como paso previo a la extracción de las historias clínicas electrónicas (HCE) se debe contar con un listado de palabras clave o ancla relacionadas con la patología o EI, a través de las tablas de vocabulario predefinidas del CDM podemos obtener los sinónimos y términos relacionados con estas palabras ancla [17] completando así, el listado de palabras que se usará para encontrar sujetos cuya HCE incluye estos términos. Este proceso es semiautomático, ya que, se deben establecer las palabras clave manualmente, siendo automática la incorporación de términos relacionados. La negación de términos se tiene en cuenta para no incurrir en la inclusión de sujetos como casos cuando realmente no lo son [18].

Los modelos de ML de aprendizaje supervisado son entrenados con un conjunto de HCEs etiquetadas, y una vez entrenados, los modelos reportan la probabilidad de que un sujeto pueda ser candidato a pertenecer a una cohorte. La probabilidad reportada siempre toma un valor continuo entre 0 y 1 y puede convertirse en una clasificación binaria usando un punto de corte.

Un ejemplo de aplicación es el paquete de R APHRODITE (Automated PHenotype Routine for Observational Definition, Identification, Training and Evaluation) [19], herramienta de código abierto mantenida por la comunidad OHDSI [17], que hace uso de datos de la HCE disponible en OHDSI CDM para construir o reutilizar modelos de clasificación automática.

La definición de cohortes se ve simplificada cuando los datos se encuentran almacenados en un lago de datos en un modelo común de datos. Existen herramientas que permiten la extracción de cohortes a partir de lagos de datos con un modelo común de datos. También es posible realizar la definición de las cohortes mediante consultas SQL al lago de datos.

2.4 Integración de cuadernos clínicos

Los cuadernos clínicos son una herramienta ampliamente utilizada en el contexto de estudios observacionales, y su objetivo es la recogida multicénctrica de información clínica más o menos exhaustiva de un conjunto de pacientes, para un estudio analítico concreto. La recogida en los cuadernos suele hacerse de forma manual, aunque las herramientas informáticas de gestión de cuadernos clínicos suelen tener mecanismos de captura masiva de datos. Los cuadernos clínicos suelen utilizarse, normalmente, cuando se va a recoger un número de variables más amplio que el que normalmente se puede obtener directamente de los sistemas de HCE, y conlleva un cierto proceso de captura manual de toda o parte de la información. En ocasiones, los cuadernos clínicos suelen utilizarse simplemente como una herramienta para compartir modelos de datos entre múltiples participantes de un proyecto de investigación, sin necesidad de acudir a modelos de datos estandarizados.

Al igual que desde los sistemas de HCE, los datos recogidos en cuadernos clínicos de investigación también pueden ser combinados bajo un modelo común de datos. Para ello, es requisito que estos sistemas de recogida se diseñen en base a buenas prácticas de modelado y estandarización que haga que los conceptos que implementan puedan ser universalmente entendidos y procesados sin pérdida de significado. Aquí, de nuevo, juega un papel fundamental los arquetipos clínicos, permitiendo establecer un marco común de estructura y contenido terminológico que haga compatibles los metadatos de los sistemas propietarios de HCE, de las plataformas de cuadernos de investigación como REDCap [11], probablemente la plataforma de cuaderno clínico más extendida y conocida, y de los modelos normalizados de repositorios como OMOP-CDM o i2b2.

REDCap es una aplicación web segura para crear y gestionar encuestas y bases de datos en línea. Aunque REDCap puede utilizarse para recopilar prácticamente cualquier tipo de datos en cualquier entorno (incluido el cumplimiento de la norma 21 CFR Parte 11, FISMA, HIPAA y GDPR), está orientada específicamente a apoyar la captura de datos en línea y fuera de línea para estudios de investigación.

Entre las características que tiene RedCap destacan las siguientes:

- Diseño rápido y flexible de la base de datos/encuestas.
- Acceso multicéntrico a las bases de datos/encuestas.
- Creación de reportes personalizados.
- Módulo de programación de eventos y citas.
- Funcionalidades avanzadas: autovalidación, campos calculados, lógica de ramificación.
- Exportación de datos en Excel, PDF, SAS, Stata, R o SPSS.
- Conexión de aplicaciones externas para recuperar o modificar mediante

programación datos o configuraciones dentro de REDCap, como por ejemplo, realizar importaciones/exportaciones de datos automatizadas desde un proyecto REDCap específico.

- Interoperabilidad con la HCE a través de recursos FHIR y la autorización OAuth2

El registro de datos en REDCap puede realizarse de diferentes formas:

- Registro manual de los datos.
- Importación de datos en formato CSV.
- Importación de datos de HCE u otras fuentes a partir de la API de REDCap. Para este proceso se necesita crear un código de programación en R, PHP, Perl, Python, Ruby, Java o cURL, en el que se especifique los datos de origen y las transformaciones que necesitan para obtener el modelo de datos del proyecto en REDCap.
- Importación de datos de HCE a partir de recursos FHIR. Permite extraer datos e la HCE de las siguientes áreas: lista de problemas, datos demográficos, vacunas, alergias, laboratorio y medicación, y cargarlos en la base de datos REDCap mediante recursos FHIR de la HCE y la autorización OAuth2.

Dada las opciones de importación de datos que proporciona REDCap, es una herramienta que facilita la recogida de datos en estudios de investigación multicéntricos, ya que es capaz de adaptarse a los diferentes grados de madurez de cada organización. Así, un centro que cuente con una HCE con interfaces FHIR puede utilizar la opción 4 para realizar su carga en REDCap. Un centro que cuente con una HCE modelada de acuerdo a estándares, puede utilizar la opción de importación directa desde HCE y crear un código de programación en el que se especifiquen las transformaciones de datos necesarias para realizar la carga en REDCap. En este caso, puesto que el origen está basado en estándares, si las operaciones de datos definidas también se construyen de forma estructurada y normalizada, se crea un código reutilizable y fácilmente adaptable a los diferentes casos de uso. Por último, para aquellas organizaciones que cuentan con un menor desarrollo tecnológico REDCap les ofrece la posibilidad de introducir los datos de forma manual o mediante archivos CSV.

Uno de los usos que resulta muy interesante al trabajar con cuadernos clínicos es el de integrar sus datos con otros datos procedentes de fuentes automatizadas, enriqueciendo la información disponible, y poder combinarla luego con datos contenidos en el resto de los nodos de una red de investigación.

Este enriquecimiento de los datos contenidos en el cuaderno clínico y su posterior combinación con los de otros nodos, se facilita considerablemente con la utilización de modelos comunes de datos. Si la estructura del esquema utilizado en el cuaderno clínico se ha realizado de acuerdo con un modelo común de datos, o bien siguiendo un modelo semántico basado en arquetipos que ya esté mapeado a un CDM. La extracción de datos del cuaderno clínico a un CDM es directa o fácilmente automatizable, y ya en el

CDM se puede combinar con datos procedentes de otras fuentes.

El principal problema que nos encontramos a la hora de combinar datos procedentes de cuadernos clínicos y de almacenes de datos procedentes de HCE, está en el enlazado de datos entre ambos. Habitualmente, y de acuerdo con las exigencias del RGPD y la LOPDGDD, los datos recogidos a partir de la Historia Clínica se conservan de forma seudonimizada. Sin embargo, los datos recogidos en los cuadernos clínicos pueden estar anonimizados, seudonimizados utilizando un algoritmo de seudonimización distinto del usado en el almacén de datos, o incluso identificado, si se dispone de consentimiento de los pacientes y el estudio así lo requiere.

Existen distintas formas de integrar los cuadernos clínicos con los almacenes o lagos de datos, pero todas ellas pasan por conseguir el enlazado de los registros de un mismo paciente asociándolo a un mismo seudónimo tanto en los registros del cuaderno clínico como en el lago o almacén de datos. Si el almacén de datos está ya en formato seudonimizado, o bien revertimos la seudonimización para enlazar con los datos del cuaderno clínico, para nuevamente volver a seudonimizar el conjunto de datos completo, o asimilamos el cuaderno clínico al resto de fuentes primarias que alimentan nuestro lago de datos. De esta manera, seudonimizaremos los registros del cuaderno clínico con el mismo algoritmo o sistema que se utilice para seudonimizar los registros de HCE antes de extraerlos y cargarlos en el almacén de datos, y enlazaremos toda la información en el almacén antes de realizar la extracción y adaptación de los datos enriquecidos en un modelo común de datos.

2.5 Análisis comparativo de CDM

El modelo común de datos **OMOP** [9] permite el análisis sistemático de bases de datos observacionales de gran diversidad. El objetivo de este modelo es transformar los datos contenidos en diversos sistemas de información, tales como la Historia Clínica Electrónica o los sistemas de Receta Electrónica, en un formato común (modelo de datos), así como una representación común (terminologías, vocabularios, esquemas de codificación), que permitan además realizar análisis sistemáticos utilizando una biblioteca de rutinas analíticas estándar basadas en su formato común.

El problema que existe con OMOP frecuentemente es que, al estar basado en una ontología relativamente simple, y cuyo origen ha estado muy vinculado al desarrollo de proyectos relacionados con la industria farmacéutica y los estudios post-autorización, no siempre es capaz de cubrir el 100% de los conceptos que se necesitan compartir para un proyecto de investigación concreto. Esta carencia se puede suplir, no obstante, mediante guías de interoperabilidad que especifique la forma en la que se han mapeado los conceptos no estándar, mediante la creación de elementos (tablas y/o columnas no estándar), o a través de los mecanismos que ofrece OHDSI para extender el modelo

OMOP CDM de una forma ordenada. El problema de estas soluciones es que convierten al modelo común utilizado en ese proyecto en un modelo parcialmente desnormalizado, o llevan mucho tiempo de desarrollo, si estamos hablando de la definición de una extensión al modelo.

El componente principal de **i2b2** [7] es el repositorio de datos clínicos, cuya base de datos se basa en un diseño de esquema en estrella, similar al modelo habitual en sistemas de *Data Warehouse*, formado por la tabla central Observation_Fact, que almacena todas las observaciones de salud entendiendo observación como cualquier evento en la salud del paciente), y un conjunto de tablas adicionales, enlazadas a esta, que le aportan información adicional. Así, i2b2 normaliza el modelo de datos, pero no establece un marco común de conceptos del dominio clínico para el contenido del mismo, como sí hacen otras propuestas de repositorios normalizados como OMOP CDM. Esto hace a i2b2 flexible en la incorporación de estándares terminológicos y de clasificación de acuerdo con las necesidades específicas de cada caso de uso y los orígenes de datos. Por contra, esto implica que cada proyecto debe acordar previamente este marco conceptual entre las organizaciones que carguen sus datos al repositorio, dependiendo nuevamente de guías de implementación específicas para cada proyecto.

ICGC Argo [20] es una iniciativa del International Cancer Genome Consortium cuyo modelo de datos es mucho más sencillo que el de OMOP y el de i2b2. Básicamente se limita a proveer un diccionario de datos que describe un esquema fijo para la creación de 15 tablas clínicas orientadas a la investigación en Oncología genómica. Si bien en algunos casos se requiere el uso de expresiones regulares concordantes con los códigos de alguna terminología o clasificación específica, no tiene diccionarios terminológicos propios, y en la mayoría de los casos los valores permisibles a reportar corresponden a definiciones propias no normalizadas. Al ser un modelo de datos orientado a la Oncología, explicita las variables que deben ser representadas (por ejemplo, “menopause_status”) en lugar de permitir que se reporte un rango más amplio de variables (como ocurre con los conceptos aceptados de condition_concept_id de OMOP). Esto aumenta la completitud de datos específicos que deben ser representados, a la vez que le resta escalabilidad y flexibilidad al modelo.

MIDS (Medical Imaging Data Structure) [21] es la propuesta de FISABIO para la estandarización, organización y gestión de datos de imagen médica. MIDS es una extensión de BIDS (Brain Imaging Data Structure) que soporta la inclusión de datasets de múltiples partes del cuerpo obtenidas mediante varias metodologías. Los beneficios de MIDS respecto a otros sistemas de organización de los datos de imagen médica incluyen la organización de cada imagen por tipo de principio físico para facilitar las interacciones con el usuario y la capacidad de categorizar todos los métodos de recepción de una imagen, lo que permite la clasificación de imágenes de nuevos dispositivos con facilidad. Este tipo de estándar proporciona una estructura que facilita su

uso en proyectos de IA.

En resumen, se puede concluir que existen varios modelos comunes de datos, cada uno de los cuales puede ser suficiente o incluso adecuado en determinados contextos de dominio específico, pero que no son capaces de resolver las necesidades de organización de la información en otros dominios distintos. A día de hoy, el CDM más generalista y extendido que existe es OMOP-CDM. Este modelo, aunque no responde al 100% de las necesidades de representación de la información que nos podemos encontrar en distintos proyectos de investigación o análisis de datos biomédicos, dispone de mecanismos que habilitan la ruptura del estándar de una forma más o menos ordenada para dichos casos, y además es un modelo en constante ampliación, con una comunidad de colaboradores muy extensa, la que hace de OMOP-CDM, a día de hoy, la mejor opción de modelo común de datos a nivel general.

3 Herramientas de procesado de datos

3.1 Herramientas para ETL de dato clínico

Los procesos de extracción, transformación y carga (ETL) tienen lugar cuando se requiere trasferir e integrar los datos desde los sistemas de origen al lago de datos, y de éste a un modelo común de datos (CDM). Los procesos ETL para datos clínicos pueden realizarse usando herramientas de uso general o aquellas adaptadas especialmente a datos de salud. Tanto en un caso como en otro, existen herramientas de software propietario y de código abierto, algunas de las cuales ofrecen como opción soporte empresarial.

Dentro del conjunto de herramientas disponibles, cabe destacar el pipeline propuesto por la comunidad OHDSI, que hace uso de tres herramientas que facilitan el proceso ETL de los datos hacia el estándar OMOP. El pipeline empieza con un escaneado de los datos de origen usando **White Rabbit** [22] (herramienta desarrollada por OHDSI) que permite un mejor entendimiento de los datos a nivel de tabla, campos y contenido. Una vez realizado el escaneado, se puede proceder con la herramienta **Rabbit in a Hat** [13] (herramienta desarrollada por OHDSI) la cual permite al usuario definir la lógica del proceso conectando los datos de origen con las tablas y campos del CDM mediante una interfaz gráfica. Para mapear los conceptos al formato del CDM se usa otra herramienta de OHDSI llamada **Usagi** [23].

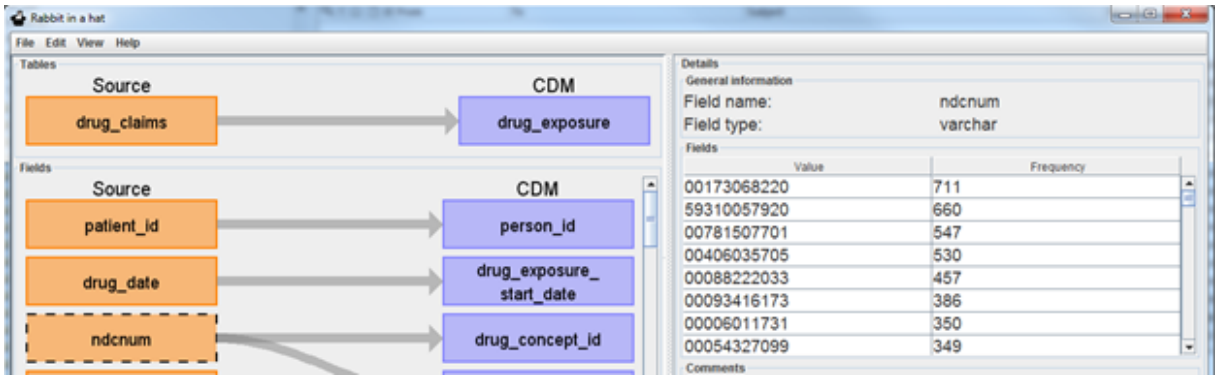


Figura 3. Ejemplo de mapeo entre modelo origen y OMOP en Rabbit-in-a-hat [2]

Como ventajas encontramos las interfaces gráficas proporcionadas por las herramientas OHDSI que da como resultado un instrumento muy intuitivo que además tiene soporte de la comunidad OHDSI a través de foros de discusión. Como desventajas, pueden destacarse la carencia de un paquete que integre las herramientas ODHSI mencionadas en el párrafo anterior junto con el formato del output generado que es gráfico, ya que, se requiere una posterior implementación de la ETL diseñada.

Además de las herramientas proporcionadas por OHDSI, que están muy orientadas a OMOP-CDM, la implementación de las ETL puede realizarse con herramientas de código abierto, como **Scriptella**. Scriptella [24] es una herramienta implementada en Java que permite la ejecución de scripts ETL ya sea en SQL u otros lenguajes que se adecuen a la fuente de datos para realizar las transformaciones necesarias. Para llevar a cabo el proceso ELT Scriptella provee tres elementos básicos como son la conexión a las fuentes de datos, scripts que contienen el código a ejecutar ya sea en SQL, javascript o incluso en un lenguaje específico del dominio (DSL) y consultas escritas en el DSL para ser lanzadas a la fuente de datos [Scriptella ETL Reference Documentation].

Otras herramientas de código abierto muy interesantes y útiles son **Talend Open Studio** [25] y **Pentaho DI** [26] (antes conocida como Ketlle). Ambas herramientas permiten el diseño de flujos ETL mediante una interfaz gráfica muy simple e intuitiva, incluyendo conexión a múltiples fuentes de datos de origen y destino, así como herramientas de filtrado, mapeo, recodificación por búsqueda (lookup) y ejecución condicional. La principal diferencia entre ambas es que mientras Pentaho guarda los procesos en un formato XML propietario, que debe ser ejecutado desde su propio motor de ejecución, Talend genera automáticamente, a partir del diseño de procesos, código Java altamente optimizado, que puede ser ejecutado en cualquier JVM, e incluso editado y modificado. Ambos productos disponen, además de la versión de código abierto, de soporte y versiones empresariales. En el caso de Pentaho, que nació como producto de código abierto, fue en su momento adquirido y es mantenido por la compañía Hitachi Vantara [27].

[2] http://ohdsi.github.io/WhiteRabbit/RabbitInAHat.html

Dentro de las opciones de código abierto, existen también paquetes de software para los principales lenguajes utilizados en ciencia de datos, como son **sqlRender** [28] en R, o **Luigi** [29] en Python. Estos paquetes simplifican enormemente el diseño de código para la ejecución de procesos ETL, que pueden ejecutarse de forma autónoma, o embeberse en cualquier pipeline de análisis. Estos paquetes permiten la ejecución de sentencias SQL parametrizadas, implementar estructuras de programación típicas como if-else o bucles, y la capacidad de adaptar consultas SQL estandarizadas a distintos dialectos, de manera que un mismo código pueda ser ejecutado contra distintos gestores y orígenes de datos, tales como SQL Server, Oracle, Postgres, Apache Spark o SQLite de forma transparente. [Rendering Parameterized SQL and Translation to Dialects • SqlRender (ohdsi.github.io)]

Algunas instituciones han optado por desarrollar sus propias herramientas, adaptadas de forma óptima a la arquitectura de sistemas de información. Tal es el caso de **TransformEHRs**, diseñada e implementada por el Hospital Universitario 12 de Octubre [30][31]. Esta plataforma permite generar el código del proceso ETL sin necesidad de realizar un desarrollo específico para cada caso de uso. Su diseño parte del análisis de los modelos de uso secundario relevantes (repositorios RWD como i2b2 y OMOP, y modelos de reportes de casos para investigación), así como del modelo de referencia de la Norma UNE-EN ISO 13606 (ISO, 2020b). Así, se ha definido un catálogo formal de operaciones de datos, implementado en SQL y R, agrupadas en “extracción”, “selección” y “transformación” [30][31]. Así mismo, analizando el modelo de uso secundario a obtener desde la HCE, se parametriza un fichero de configuración de ETL implementado con XML (W3C, 2022c), según las restricciones de la especificación de datos objetivo. Finalmente, a partir de ambos componentes, catálogo de operaciones de datos y fichero de configuración, se obtiene el código necesario para ejecutar el proceso ETL sobre los sistemas de HCE. Esta herramienta ha sido adoptada por la plataforma integral de datos del Hospital Universitario 12 de Octubre denominada “INFOBANCO” [32].

Al margen de estas soluciones descritas, desarrolladas ad hoc, o basadas en código abierto, prácticamente todos los fabricantes de soluciones de gestión de datos, tanto gestores de bases de datos como herramientas de procesamiento y análisis, tienen soluciones ETL propietarias, sujetas a licencia comercial. Tal sería el caso, entre otras muchas, de Oracle, Microsoft, IBM, Tibco, Amazon, o Informatica.com, que posee una fantástica suite de productos para la gestión, y transformación y análisis de datos, aunque la lista sería casi interminable. Estas soluciones pueden ser una alternativa interesante, especialmente cuando se configuran ecosistemas completos basados en las soluciones técnicas de estas compañías. El sobrecoste de las licencias de estos productos vendría en gran medida compensado por el soporte empresarial que estas compañías ofrecen, y que en general es mucho más difícil de conseguir en las soluciones de código abierto.

En resumen, existe un catálogo enorme de herramientas y paquetes de software que

facilitan el diseño de procesos de extracción, transformación y carga de datos, tanto para la captura de dato primario como para la generación de modelos comunes de datos. La selección de una herramienta concreta dependerá de múltiples factores: el tipo de normalización de información de la que se parte, la configuración o no de ecosistemas determinados para la gestión de datos, ya sean comerciales o de código abierto, la flexibilidad deseada en los desarrollos, y el nivel de capacitación del equipo técnico que va a hacerse cargo de esta parte del procesamiento de datos.

Lo cierto es que los procesos ETL, junto con los de gestión de la calidad de datos, son los procesos que más coste, tiempo y esfuerzo suponen en todo el *pipeline* de análisis de datos sanitarios (y probablemente, también en otros dominios), por lo que la selección de herramientas que faciliten el diseño de procesos ETL, que los hagan intuitivos, transparentes y mantenibles, es esencial a la hora de configurar un buen ecosistema de gestión y análisis de datos.

3.2 OHDSI ATLAS y otras herramientas de selección de cohortes

Si el punto de partida para la obtención de cohortes de pacientes en formato OMOP es un repositorio completo OMOP-CDM, el trabajo se simplifica considerablemente, ya que la única tarea que se requiere es la de identificación y selección de la cohorte, ya que no es preciso ningún tipo de transformación de los datos.

ATLAS [14] es una herramienta de software de código abierto cuya finalidad es que los investigadores puedan realizar análisis sobre datos de observaciones convertidos al CDM. Los investigadores pueden crear cohortes definiendo grupos de personas en función de la exposición a un fármaco o el diagnóstico de una afección, por lo que se trata de una definición de cohortes basada en reglas.

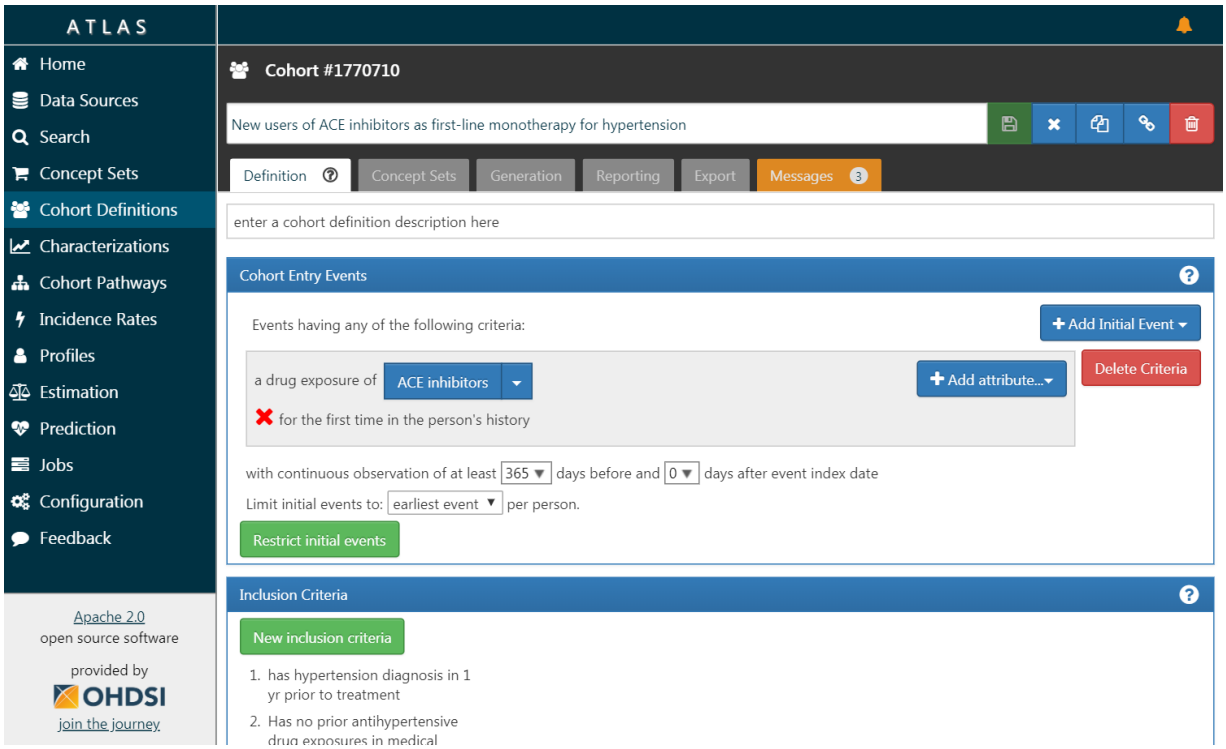


Figura 4. Interfaz de Atlas [3]

El primer paso a la hora de definir una cohorte con ATLAS es definir el evento inicial para el cual un sujeto pueda ser incluido en la cohorte, para ello se debe seleccionar el dominio sobre el cual se basa el criterio. Por ejemplo, si se buscasen pacientes que toman un determinado medicamento, el criterio para definir el evento inicial se basaría en el dominio DRUG_EXPOSURE.

Una vez seleccionado el dominio se debe definir un conjunto de conceptos necesario para aplicar el criterio del evento inicial, en el caso del dominio DRUG_EXPOSURE podemos indicar tanto el principio activo del fármaco así como todos los fármacos que contienen el principio activo (conceptos descendientes).

Además del evento inicial, es posible añadir criterios de inclusión para conocer la idoneidad del sujeto para el estudio, para ello ATLAS proporciona la opción de añadir atributos en la misma sección donde se define el criterio inicial, un ejemplo de criterio inclusión es que sea la primera vez que el paciente toma un fármaco.

Al igual que criterios de inclusión también es posible definir los criterios de expulsión, los cuales son todos aquellos eventos definidos por el usuario que hacen que un sujeto deje de ser elegible para la cohorte de estudio. Para incluir los criterios de expulsión se procede de la misma forma que cuando se añaden otros atributos a la definición de la cohorte.

[3] https://www.ohdsi.org/atlas-a-unified-interface-for-the-ohdsi-tools/

Una vez se ha realizado la definición de la cohorte es posible exportar la definición en formato JSON o en código SQL.

ATLAS es una herramienta gratuita, disponible públicamente y basada en web. Está desarrollada por la comunidad OHDSI, y facilita el diseño y la ejecución de análisis sobre datos observacionales estandarizados, a nivel de paciente, en el formato OMOP CDM. ATLAS se despliega como una aplicación web en combinación con la OHDSI WebAPI.

Esta herramienta provee múltiples funcionalidades, entre las que se encuentran:

- Fuentes de datos: visualización de informes descriptivos y estandarizados para cada una de las fuentes de datos que se han configurado dentro de la plataforma.
- Búsqueda de vocabulario: búsqueda y exploración del vocabulario estandarizado de OMOP para entender qué conceptos existen y cómo aplicar esos conceptos dentro de un análisis de los datos.
- Conjunto de conceptos: creación de colecciones de expresiones lógicas que pueden utilizarse para identificar un conjunto de conceptos que se utilizarán en los análisis normalizados. Un conjunto de conceptos se compone de múltiples conceptos del vocabulario normalizado en combinación con indicadores lógicos que permiten al usuario especificar que le interesa incluir o excluir conceptos relacionados en la jerarquía del vocabulario. Estos conjuntos de conceptos pueden ser guardados dentro de ATLAS y luego utilizados a lo largo de los análisis como parte de las definiciones de cohortes o especificaciones de análisis.
- Definición de cohortes: construcción de un conjunto de personas que satisfacen uno o más criterios durante un periodo de tiempos. Estas cohortes pueden servir de base para posteriores análisis.
- Caracterización: capacidad analítica que permite observar una o más cohortes y resumir las características de esa población de pacientes.
- Vías clínicas de cohortes: herramienta analítica que permite observar la secuencia de eventos clínicos que se producen en una o más poblaciones.
- Tasa de incidencia: herramienta que permite estimar la incidencia de los resultados en las poblaciones de interés.
- Perfiles: herramienta que permite explorar los datos de observación longitudinal de un paciente individual para resumir lo que ocurre en un individuo determinado.
- Estimación a nivel de población: definición de un estudio de estimación del efecto a nivel de la población, utilizando un diseño de cohorte comparativo mediante el cual se pueden explorar las comparaciones entre una o más cohortes objetivo y de comparación para una serie de resultados.
- Predicción a nivel de paciente: aplicación de algoritmos de aprendizaje automático para llevar a cabo análisis de predicción a nivel de paciente mediante

los cuales puede predecir un resultado dentro de cualquier exposición objetivo.

Así, ATLAS es una herramienta que proporciona tanto un análisis descriptivo de la información disponible en la base de datos OMOP, como la posibilidad de generar cohortes y diferentes tipos de estudios. Una de las principales ventajas que posee es que tanto las definiciones de cohortes como de análisis son exportables y, por tanto, pueden ser ejecutadas de la misma forma en cualquier organización que cuente con un repositorio OMOP.

Otra herramienta interesante, fuera del ecosistema OHDSI, es la plataforma La plataforma TriNetX [33], la cual permite el diseño estudios clínicos relativos a los centros hospitalarios contenidos en la red de TriNetX, de carácter internacional, incluyendo zona EMEA y Estados Unidos entre otras.

La herramienta permite la creación de consultas específicas para cada estudio, para lo cual el usuario es capaz de diseñar las cohortes necesarias para cada caso de uso determinando los rasgos demográficos de la cohorte, las pruebas clínicas o procedimientos que se precise añadir e incluso agregar filtros respecto a dichas características para realizar una búsqueda tan granular como se especifique. El usuario también debe señalar la red de centros hospitalarios que deban ser objeto de estudio, algunas de las cuales están incluidas por defecto en el sistema.

Una vez diseñada y ejecutada una consulta, la herramienta devolverá el número exacto de pacientes correspondientes a la misma, rasgos estadísticos elementales sobre la cohorte (incluyendo distribución geográfica) y su distribución según los rasgos demográficos o diferentes pruebas que se hayan incluido en el diseño de la consulta.

3.3 Virtualización de bases de datos

La virtualización de datos consiste en la creación, mediante herramientas software adecuadas, de una capa lógica de abstracción de una infraestructura de datos diversa y dispersa.

La virtualización de datos permite a sus usuarios acceder, combinar, transformar y obtener conjuntos de datos de una forma rápida y trasparente, como si se estuviera accediendo a un único sistema centralizado y homogéneo, cuando realmente se está accediendo a diversos sistemas de base de datos, tanto sistemas relacionales tradicionales, sistemas NoSQL, Big Data, ficheros o sistemas en la nube.

Los sistemas de virtualización de datos, por tanto, simplifican considerablemente la gestión, gobernanza e integración de sistemas heterogéneos y distribuidos de bases de datos.

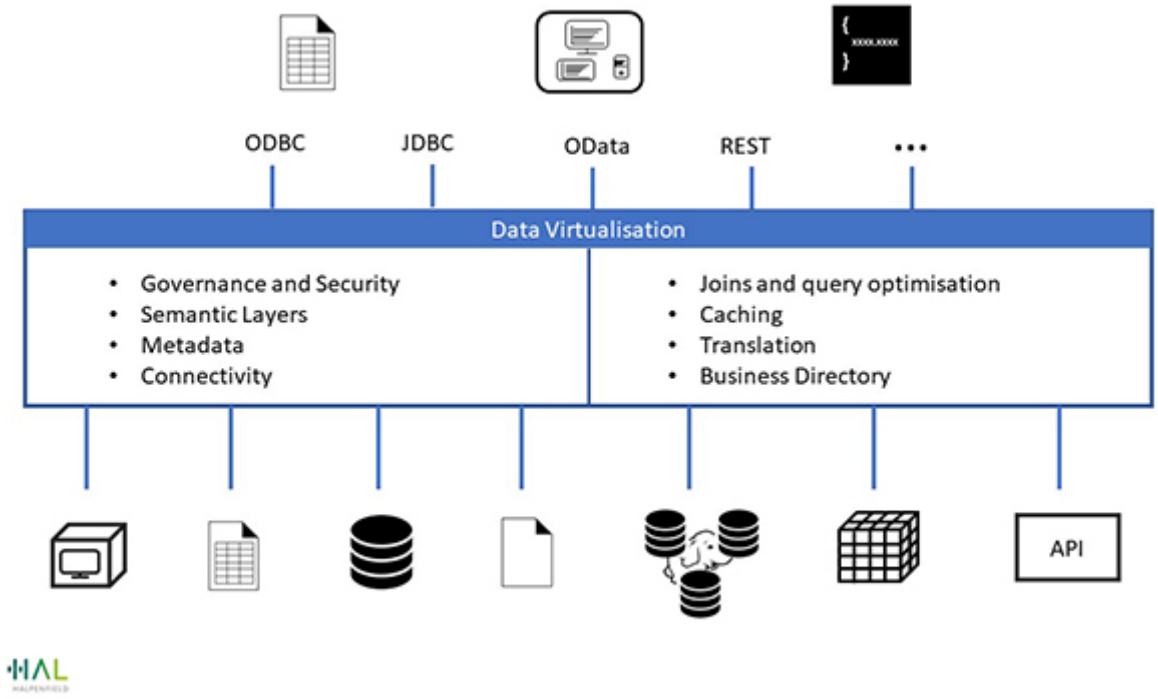


Figura 5. Esquema conceptual de la virtualización de bases de datos [4]

En el contexto del uso secundario de datos de salud, la virtualización de datos tiene dos ámbitos de utilidad:

- Intranodo, como forma de facilitar la integración y los procesos ETL en infraestructuras asistenciales con múltiples sistemas de información heterogéneos.
- Internodo, como forma de ejecutar pipelines de análisis de datos sobre infraestructuras federadas, de forma centralizada. La virtualización de datos puede mostrar un conjunto de nodos de bases de datos multicéntricas como una única instancia monolítica, sobre la que ejecutar consultas y procesos de análisis. Es posible con algunas de esas herramientas garantizar la no persistencia de ningún dato temporal, de manera que los datos no se moverían nunca de sus bases de datos origen, salvo su persistencia en la memoria del sistema de virtualización durante la ejecución de las consultas y análisis.

Al igual que ocurre en el caso de herramientas ETL, las soluciones existentes, tanto comerciales como de código abierto, sin interminables. Entre las soluciones comerciales, asociadas a ecosistemas más o menos monolíticos, podemos destacar las herramientas de Oracle o SAP entre otras, muy orientadas a trabajar dentro de sus propios ecosistemas. En este grupo, merecería una mención especial el IBM Cloud Data Pak, que además de la solución de virtualización de bases de datos, ofrece una completísima suite para diseñar pipelines completos de análisis de datos, desde la captura hasta el entrenamiento de modelos de inteligencia artificial y deep learning. Algo similar ocurre

[4] https://www.sdggroup.com/en-GB/insights-room/case-data-virtualization

con la solución de informatica.com. Otras soluciones comerciales más específicas y menos monolíticas, dignas de mencionar, son Denodo o Data Virtuality.

Entre las soluciones de virtualización de código abierto, podemos destacar JBoss Data Virtulization, un producto actualmente bajo propiedad de IBM, pero que mantiene su condición de código abierto, y que además ofrece soporte empresarial.

Otra solución de código abierto que también resulta muy interesante es Apache Trino. Esta solución de virtualización, desarrollada por Facebook bajo en nombre “Presto”, se ha ramificado bajo el paraguas Apache bajo la denominación Trino. Trino es un motor SQL que soporta la práctica totalidad del estandar SQL’96, además de funcionalidades añadidas, y que en lugar de tener persistencia de datos propias, puede utilizar casi cualquier fuente de datos, estructurada y NoSQL, como base de persistencia. Eso permite crear un esquema de base de datos en Trino, cada una de cuyas tablas virtuales tiene su persistencia en cualquier sistema gestor, local o remoto. Las consultas SQL se realizan transparentemente sobre el esquema virtualizado, incluso combinando datos de distintas fuentes dentro de la consulta.

3.4 Herramientas para análisis y visualización de datos

En el desarrollo de los procesos integrales de análisis de datos, las últimas fases de los mismos, una vez recolectados, curados, depurados, normalizados y seleccionados los datos requeridos, y estructurados en un modelo de datos adecuado, son el análisis en sí y la visualización de los resultados. El análisis puede consistir en cualquier operación de tratamiento de los datos que permita obtener o extraer nuevo conocimiento de los mismos, ya sea mediante análisis matemático-estadísticos, procesos de aprendizaje automático o utilización de redes neuronales para su procesamiento.

Históricamente, el tratamiento de datos se ha realizado utilizando herramientas autocontenidas, con capacidad de aplicar distintos modelos de análisis sobre los datos. Las más conocidas, sobre todo dentro del contexto estadístico, son **SPSS [34]** y **Stata** [35], aunque otras herramientas como **Matlab** [36] de Mathworks o GNU **Octave** [37], de código abierto, se han utilizado también con este fin. En el caso de algoritmos de minería de datos y aprendizaje automático, herramientas como **Weka** [38] o **Rapidminer** [39] son también opciones utilizadas. En contextos empresariales, son también dignas de destacar las ya mencionadas soluciones de **Informatica.com**, el **IBM Cloud Data Pak** [40] o las soluciones analíticas de **Amazon**, todas ellas muy orientadas al trabajo en Cloud. IBM, actual propietario de SPSS, ha integrado los más que probados modelos estadísticos de SPSS en los data flows de su Cloud Data Pak, lo que hace de esta una herramienta muy potente, aunque sujeta a licenciamiento. Por otro lado, existe un clon de código abierto de SPSS denominado PSPP [41].

No obstante, el abanico de soluciones comerciales y de código abierto para el análisis de

datos, cada vez son más los analistas que desarrollan o adaptan sus propios códigos de análisis en el lenguaje de programación de elección, utilizando librerías y herramientas de código abierto y algoritmos de ciencia abierta. En este contexto, los lenguajes de programación más extendidos son, sin duda, **Python** [42] y **R** [43], aunque también hay quien desarrolla procesos analíticos en **Java** [44], **Scala** [45] o **Julia** [46]. Todos estos lenguajes disponen de una ingente cantidad de librerías ya desarrolladas y probadas, para la ejecución de multitud de algoritmos de análisis estadístico, de aprendizaje automático o de aprendizaje profundo. Librerías como **Pandas** [47] para el manejo de dataframes, **NumPy** [48] para manejar matrices, o **Scikit-learn** [49] para algoritmos de aprendizaje automático y minería de datos, hasta las librerías de redes neuronales de **Tensorflow** [50] o **Keras** [51], están disponibles en Python, y muchas de ellas portadas a otros lenguajes. En el caso de R, la biblioteca de librerías disponibles en **CRAN** [52] o R**-bioconductor** [53] (entre otros repositorios) es casi interminable. Además, librerías de análisis como Spark, que incluye módulos de gestión de dataframes, aprendizaje de máquina, SQL y grafos, es accesible y utilizable desde cualquiera de estos lenguajes.

Project **Jupyter [54]** ofrece herramientas para el análisis interactivo de los datos y la computación científica en múltiples lenguajes de programación, tales como Julia, Python o R. Jupyter Notebook es la opción que podemos considerar más simple. Esta permite crear y compartir documentos computacionales de forma optimizada. Los Notebook de Jupyter son populares para el análisis de datos por su componente visual a la hora de mostrar los resultados y la capacidad de ejecutar bloques de código independientes en un mismo fichero (ver Figura 6). Además, permite la inclusión de bloques texto en formato Markdown, que resultan útiles para añadir contexto entre el código. Por otro lado está JupyterLab, una versión más avanzada de Jupyter Notebook, ya que ofrece una interfaz web que permite trabajar con múltiples Jupyter Notebook además de con otros tipos de documentos (HTML, texto, Markdown, etc.) y consola. Finalmente, con JupyterHub posibilita el uso de las herramientas Jupyter entre grupos de usuarios de una forma customizable, flexible, escalable y portable.

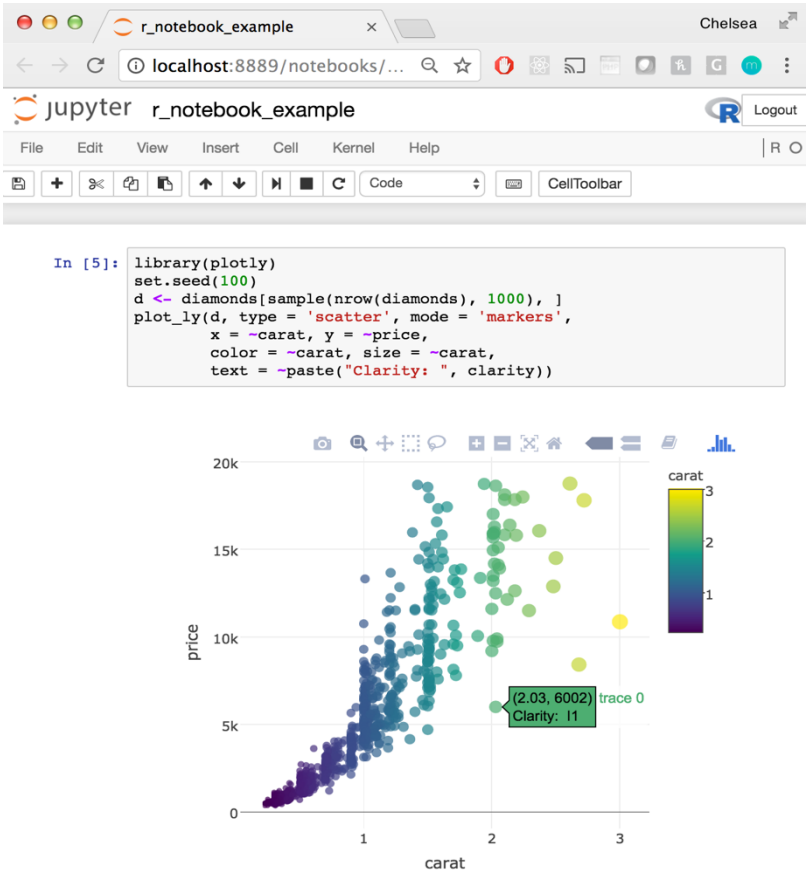


Figura 6. Ejemplo de Jupyter Notebook

Google Colab [55] ofrece el acceso a entornos interactivos llamados Colab notebook que permiten al usuario escribir y ejecutar código en Python. Entre sus ventajas se encuentra el acceso a GPUs sin coste económico, no ser necesario ningún tipo de configuracion local y la facilidad de compartición de los archivos, de forma que varios usuarios puedan editar el mismo archivo de forma remota. Su uso es similar a las Jupyter Notebook.

Otras herramientas muy similares a estas son **Apache Zeppelin** [56], centrada inicialmente en Scala y **Spark** [57], aunque dispone en la actualidad de intérpretes para decenas de lenguajes, o **Polynote** [58], un desarrollo reciente que integra la capacidad de ejecución de bloques de código con la visualización de los resultados a través de una librería de generación de gráficos muy conocida, **Matplotlib** [59].

R Studio Server [60] y **R Shiny** [61] son herramientas basadas en el lenguaje de programación R, que permiten, al igual que los notebooks mencionados más arriba, la creación de informes dinámicos combinando bloques de cálculo y análisis con texto estático y gráficos, y la creación de aplicaciones web interactivas que se pueden publicar y compartir.

En el caso de estudios en los que se deben analizar también imágenes médicas, se

precisan herramientas más específicas, que permitan, además de los posibles análisis utilizando librerías que combinen la gestión de imagen con los algoritmos de análisis o aprendizaje profundo, la visualización, selección de áreas, etiquetado y otras operaciones específicas sobre las imágenes. En este ámbito, cabe destacar soluciones como OHIF (Open Health Imaging Foundation) [62]. OHIF Medical Imaging Viewer es una plataforma web de código abierto para la creación y gestión de aplicaciones de imagen médica, que permite realizar todas estas operaciones a partir de imagenes capturadas en formato DICOM, así como representarlas en 2D y 3D, y definir flujos de análisis sobre conjuntos más o menos masivos de imágenes. No obstante, no profundizaremos mucho más en este tipo de herramientas, que serán amplia y debidamente comentadas en el entregable 4.5 del presente proyecto.

Desde el punto de vista de la publicación y visualización de datos y resultados, es preciso diferenciar los distintos objetivos que tiene un sistema de uso secundario de dato sanitarios, y los distintos perfiles profesionales que harán uso de ellos. No serán iguales las necesidades de un investigador biomédico, que las de un gestor que necesita responder a determinadas preguntas sobre el funcionamiento del sistema sanitario o la salud pública, que la de un facultativo médico que necesita comparar los datos de un paciente con los de una cohorte determinada, a la hora de tomar una decisión clínica.

De forma general, además de la utilización de librerías gráficas para la generación de gráficos y tablas, que se pueden utilizar desde distintos lenguajes de programación, e integrar en herramientas como las ya mencionadas anteriormente Jupiter, Zeppelin o R Shiny, existen herramientas de propósito general para la automatización de la visualización y publicación de indicadores y gráficos. **Tableau** [63], como solución comercial con versión básica gratuita, o **Redash** [64] como solución de código abierto, son dos ejemplos de soluciones que permiten el diseño de cuadros de mando a partir de componentes prediseñados (tablas y gráficos), sin necesidad de conocimientos técnicos avanzados. Es decir, son herramientas pensadas para técnicos funcionales y no personal informático o experto en gestión de datos. Los cuadros así diseñados se pueden publicar o convertir en portales web, de manera que cualquier decisor pueda consultar indicadores y gráficos que se van actualizando automáticamente con la frecuencia establecida.

Desde los ámbitos más clínicos, una de las utilidades más frecuentes y demandadas en el desarrollo de cuadros de mando clínicos de procesos asistenciales y condiciones de salud para identificar, priorizar y monitorizar grupos de pacientes por sus condiciones clínicas y sus determinantes de salud. Estas herramientas tienen como finalidad última simplificar la labor de los clínicos en cuanto a facilitar la consulta de información agregada, que actualmente reside en múltiples fuentes de información de diferente diseño y propósito.

Para el desarrollo de un cuadro de mandos de una condición de salud particular, se ha

de comenzar por identificar la cohorte, para posteriormente filtrar los modelos de información previamente definidos. Un mecanismo adecuado para identificar la cohorte en la HCE es a través de un 'Proceso asistencial' codificado de acuerdo al catálogo de evaluaciones SNOMED-CT [65]. Un proceso (o problema) que en la norma ISO-EN 13940 podría asociarse al concepto "Health issue" o "Asunto de salud" se comporta como un agrupador de episodios o visitas. A su vez, un proceso puede estar asociado a uno o varios diagnósticos codificados. Un diagnóstico es el concepto de Historia Clínica para registrar la observación en un momento temporal concreto de una enfermedad o evento asociado a la salud de un paciente. Estos diagnósticos nos permiten mayor granularidad en el filtrado de los datos del cuadro de mandos, siendo posible el filtrado por uno o más diagnósticos específicos dentro de la condición de salud. Para escalar un cuadro de mandos a otra condición, bastaría con definir la nueva cohorte en función del proceso que se emplee en esa condición clínica y filtrar el modelo de datos en base a esa cohorte.

Así mismo, desde la plataforma TriNetX [33] (mencionada anteriormente), es posible el estudio de las características de una o varias cohortes, incluyendo la comparación entre diferentes cohortes. De esta forma la herramienta ofrece al usuario la posibilidad de centrarse en uno o varios rasgos clínicos de la cohorte (demográfico, prueba clínica, procedimiento, diagnóstico, medicación…) y analizar sus datos estadísticos como la evolución el tiempo, análisis de riesgos, porcentaje que representa cada rasgo clínico en la cohorte, etc.; en el caso de varias cohortes se ofrece la comparación entre ambas para estudiar su similitud o rendimiento, por ejemplo en el caso de estudiar la mortalidad de una cohorte al agregar un medicamento u otro. La herramienta a su vez permite la exportación de los resultados en diferentes formatos, incluyendo Word, al terminal del usuario. De esta forma, el usuario puede elegir las estadísticas o comparativas que necesite en el estudio para directamente incluirlos en un documento o compartirlo *online* desde la plataforma de TriNetX a otros usuarios de la plataforma, así como compartir estudios para diseñarlos de manera colaborativa.

Un requisito clave para la utilización de herramientas de visualización y cuadros de mandos es el uso de modelos de información comunes, que sean explotables para múltiples propósitos de uso secundario independientemente de la condición de salud a estudio. Aquí juegan un papel fundamental los arquetipos clínicos, permitiendo formalizar y compartir estas especificaciones comunes.

Como se ha explicado en secciones anteriores, las peculiaridades de cada institución, la adopción o no de soluciones comerciales homogéneas, y la capacitación técnica del personal de las instituciones determinará la solución idónea para el tratamiento y análisis de los datos en cada proyecto de investigación o pregunta analítica planteada.

4 Conclusiones

En el presente documento, se ha expuesto la problemática derivada de la enorme disparidad de sistemas de información primaria, asistencial y administrativa, en el contexto de los servicios y dispositivos sanitarios. Esta disparidad exige, en principio, el desarrollo de soluciones específicas para cada instalación o sistema de información de origen. Es posible, no obstante, identificar una serie de patrones o modelos de sistemas de información que nos permiten establecer igualmente unos patrones de soluciones técnicas para desarrollar los procesos de extracción y transformación de los datos.

Existen a disposición de los desarrolladores una gran cantidad de herramientas tecnológicas, tanto comerciales como de código abierto o mixtas (herramientas de código abierto con versiones extendidas y soporte comercial), para crear soluciones en todas las fases del ciclo de vida de los datos, desde su extracción inicial hasta la visualización final de los resultados, pasando por la transformación, aseguramiento de la calidad, persistencia, selección y análisis de los datos.

Sin embargo, más allá de las soluciones tecnológicas que resuelven, esencialmente, los aspectos relacionados con la interoperabilidad técnica y sintáctica, el mapeo conceptual entre los datos clínicos recogidos en una Historia Clínica Electrónica y los datos que son analizados en procesos de extracción de conocimiento es esencial. Estar seguros de que la interpretación que se hace de un resultado analítico, masivo y agregado, es la misma que la que hicieron los facultativos a la hora de registrar la información clínica en el sistema de Historia Clínica Electrónica es imprescindible si queremos que el conocimiento y las conclusiones obtenidas sean realmente útiles. Los procesos de aseguramiento de la calidad, igualmente, son imposibles sin una correcta interpretación semántica y una contextualización de los datos. Por esta razón, además de las herramientas y los modelos comunes de datos, es imprescindible dotar a todo el sistema de ontologías y bibliotecas de arquetipos de información estandarizadas y compartidas a lo largo de todo el ciclo de vida de los datos.

A partir de todo lo explicado y recopilado en este documento, podemos destacar una serie de comentarios y recomendaciones:

- Los sistemas de información primarios son muy diversos, tanto en estructuras de datos como en modelos organizativos. La extracción de nuevo conocimiento a partir de esos datos es imposible si no se comprende claramente la forma en la que estos datos son almacenados, y el correcto significado de cada uno de ellos. Por esta razón, a la hora de desarrollar proyectos o plataformas de uso secundario del dato de salud, es imprescindible que en él, junto con los técnicos y científicos de datos asignados a las tareas analíticas, se impliquen de forma clara el personal TIC de los servicios y dispositivos sanitarios, que son quienes mejor

conocen las estructuras de datos de los sistemas primarios, y también los técnicos funcionales responsables de la Historia Clínica Electrónica, que son quienes conocen la estructura semántica de la información de la Historia Clínica, y la forma en la que los facultativos introducen y utilizan esa información, así como los protocolos y estrategias de salud, imprescindibles para interpretar correctamente la información.

- Conocer, discriminar, instalar y utilizar con soltura la totalidad de herramientas informáticas que pueden ayudarnos en un proyecto de uso secundario de datos es una tarea absolutamente inabarcable para cualquier equipo de técnicos y científicos de datos. En este caso, como en otros, lo perfecto es enemigo de lo bueno, y las curvas de aprendizaje de manejo de las nuevas herramientas y tecnologías es mucho mayor que la velocidad a la que salen a la luz nuevas herramientas, con seguridad interesantísimas y muy útiles. Es fundamental, en cada proyecto, analizar y optar por un ecosistema de soluciones que cubra adecuadamente todo el ciclo de vida de los datos, y capacitar al personal para el uso de esas herramientas, sacándoles todo el rendimiento posible, sin saltar cada semana a una tecnología o herramienta nueva. En este caso también, la homogeneización de algunas de estas herramientas entre distintos proyectos, facilitará la reutilización de código y la compartición de experiencias y buenas prácticas, aunque no es un requisito de obligado cumplimiento, y cierto grado de diversidad puede ayudar a ir identificando aquellas herramientas que aporten valores diferenciales a este tipo de proyectos.
- En un contexto de desarrollo de proyectos multicéntricos, o en el escenario aparentemente próximo del desarrollo de espacios de datos geográficamente extensos (nacionales y europeos), tan importante o más que una completa y correcta estandarización de las capas tecnológica y sintáctica, es el poder disponer de estándares semánticos que vayan más allá de la común utilización de terminologías y clasificaciones clínicas. Es preciso en los próximos años trabajar en el desarrollo de bibliotecas compartidas de arquetipos y ontologías clínicas que cubran todo el conocimiento asociado a la investigación biomédica, incluyendo los distintos objetivos (investigación de base, monitorización de salud pública, desarrollo de políticas sanitarias) y los distintos tipos de información disponibles (dato clínico, imágen médica, dato genómico y proteómico, muestras biológicas y otros condicionantes de salud).
- La calidad de los datos en los sistemas de origen es fundamental para que los resultados obtenidos sean correctos, y las decisiones que se tomen a partir de los mismos, útiles. Sin embargo, el contexto en el que se capturan los datos clínicos en el ámbito de la consulta médica es totalmente distinto al de un estudio de laboratorio o un ensayo clínico. Es necesario que los sistemas analíticos ofrezcan un retorno a los clínicos y a los gestores, de manera que el procesamiento secundario de los datos suponga un beneficio directo para los usuarios finales, y éstos comprueben las ventajas de registrar la información de forma adecuada y

completa en la Historia Clínica. Este mecanismo de feedback, a modo de "tercera columna"[5] de la Historia Clínica, habilita un diálogo continuo entre los sistemas de uso primario y secundario de datos clínicos, que debe redundar en una mejora de la calidad de la información, y una mejora también de los procesos asistenciales, al disponer de herramientas de ayuda a la decisión integradas en el portal clínico.

[5] La "primera columna" es lo que el médico lee –historia, anamnesis-, y la "segunda columna", lo que el médico escribe –informes, escalas...-. La "tercera columna" serían los sistemas de retroalimentación, documentación contextual y sistemas de ayuda a la toma de decisiones.

Referencias

1. Moner Cano, D. (2021). Archetype development and governance methodologies for the electronic health record [Tesis doctoral]. Universitat Politècnica de València. https://doi.org/10.4995/Thesis/10251/164916
2. Beale, Thomas. (2002). Archetypes: Constraint-based Domain Models for Future-proof Information Systems. Eleventh OOPSLA Workshop on Behavioral Semantics. Serving the Customer.
3. https://www.openehr.org/
4. http://www.en13606.org/
5. Lozano-Rubí R, Muñoz Carrero A, Serrano Balazote P, Pastor X. OntoCR: A CEN/ISO-13606 clinical repository based on ontologies. J Biomed Inform. 2016 Apr;60:224-33. doi: 10.1016/j.jbi.2016.02.007. Epub 2016 Feb 18. PMID: 26911524.
6. https://hadoop.apache.org/
7. https://www.i2b2.org/
8. https://www.sentinelinitiative.org/methods-data-tools/sentinel-common-data-model
9. https://www.ohdsi.org/
10. https://specifications.openehr.org/releases/QUERY/latest/AQL.html
11. https://www.project-redcap.org/
12. https://metadatacenter.org/
13. http://ohdsi.github.io/WhiteRabbit/RabbitInAHat.html
14. https://www.ohdsi.org/atlas-a-unified-interface-for-the-ohdsi-tools/
15. Serrano-Balazote, Pablo & Moner, David & Sebastián Viana, Tomás & Maldonado, Jose & Navalón, Rafael & Gómez, Ángel. (2009). Utilidad de los arquetipos ISO 13606 para representar modelos clínicos detallados. RevistaeSalud.com.
16. https://www.ehden.eu/
17. Banda, J. M., Halpern, Y., Sontag, D., & Shah, N. H. (2017). Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. AMIA Summits on Translational Science Proceedings, 2017, 48.
18. Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., & Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of biomedical informatics, 34(5), 301-310.
19. https://github.com/OHDSI/Aphrodite
20. https://www.icgc-argo.org/
21. https://github.com/BIMCV-CSUSP/MIDS
22. https://www.ohdsi.org/analytic-tools/whiterabbit-for-etl-design/
23. https://www.ohdsi.org/analytic-tools/usagi/
24. https://scriptella.org/
25. https://www.talend.com/products/talend-open-studio/
26. https://github.com/pentaho/pentaho-kettle

27. https://www.hitachivantara.com/en-us/products/lumada-dataops/data-integration-analytics.html
28. https://cran.r-project.org/web/packages/SqlRender/index.html
29. https://luigi.readthedocs.io/en/stable/
30. Pedrera M, Garcia N, Rubio P, Cruz JL, Bernal JL, Serrano P. Making EHRs Reusable: A Common Framework of Data Operations. *Stud Health Technol Inform*. 2021;287:129-133. doi:10.3233/SHTI210831
31. Pedrera-Jiménez M, García-Barrio N, Cruz-Rojo J, et al. Obtaining EHR-derived datasets for COVID-19 research within a short time: a flexible methodology based on Detailed Clinical Models. *J Biomed Inform*. 2021;115:103697. doi:10.1016/j.jbi.2021.103697
32. https://cpisanidadcm.org/infobanco/
33. https://trinetx.com/
34. https://www.ibm.com/es-es/analytics/spss-statistics-software
35. https://www.stata.com/
36. https://www.mathworks.com/products/matlab.html
37. https://octave.org/
38. https://www.cs.waikato.ac.nz/ml/weka/
39. https://rapidminer.com/
40. https://www.ibm.com/es-es/products/cloud-pak-for-data
41. https://www.gnu.org/software/pspp/
42. https://www.python.org/
43. https://www.r-project.org/
44. https://www.java.com/es/
45. https://www.scala-lang.org/
46. https://julialang.org/
47. https://pandas.pydata.org/
48. https://numpy.org/
49. https://scikit-learn.org/stable/
50. https://www.tensorflow.org/
51. https://keras.io/
52. https://cran.r-project.org/
53. https://www.bioconductor.org/
54. https://jupyter.org/
55. https://colab.research.google.com
56. https://zeppelin.apache.org/
57. https://spark.apache.org/
58. https://polynote.org/latest/
59. https://matplotlib.org/
60. https://www.rstudio.com/products/rstudio/download-server/
61. https://www.rstudio.com/products/shiny/shiny-server/

62. https://ohif.org/
63. https://www.tableau.com/es-es
64. https://redash.io/
65. https://www.snomed.org/

Acrónimos y Abreviaturas

CDM	Common Data Model
CIAP	Codificación Internacional de Atención Primaria (ICPC)
CIE	Codificación Internacional de Enfermedades (ICD)
CMBD	Conjunto Mínimo Básico de Datos
CSV	Comma Separated Values – Formato de fichero de datos
DICOM	Digital Image Communication
EHDEN	European Health Data Evidence Network
ETL	Extracción, Transformación y Carga (Loading) de datos de un sistema de gestión de datos a otro.
FAIR	Findable, Accesible, Interoperable, Reusable
FHIR	Fast Health Interoperability Resources
HCE	Historia Clínica Electrónica
HIPAA	Health Insurance Portability and Accountability Act
HL7	Health Level 7
IMPaCT	Infraestructura de Medicina de Precisión asociada a la Ciencia y la Tecnología
IMPaCT-Data	Programa de ciencia de datos de IMPaCT
JSON	Javascript Object Notation
JSON-LD	Javascript Object Notation - Linked Data
LIS	Laboratory Information System
LOINC	Logical Observation Identifiers Names and Codes
LOPDGDD	Ley Orgánica de Protección de Datos y Garantía de Derechos Digitales
NLP	Natural Language Processing
NoSQL	Not only SQL – Define a los sistemas de bases de datos que trascienden el modelo relacional, incluyendo bases de datos clave-valor, documentales o basadas en grafos.
OAuth2	Open Authorization V.2. Estándar industrial de autorización de acceso a recursos informáticos o de información
ODS	Operational Data Store – Almacén provisional de datos en bruto, en las arquitecturas Data Warehouse
OHDSI	Observational Health Data Sciences and Informatics, organización responsable del modelo OMOP-CDM
OMOP-CDM	Observational Medical Outcomes Partnership – Common Data Model
openEHR	Open Electronic Health Record
RDF	Resource Description Framework
RGPD	Reglamento General de Protección de Datos
RIM	Reference Information Model

SNOMED-CT	Systematic Nomenclature on Medicina – Clinical Terms
SQL	Structured Query Language – Lenguaje estandarizado de consulta a bases de datos relacionales. Por extensión, denomina también a los sistemas de gestión de bases de datos relacionales que soportan dicho lenguaje.
TIC	Tecnologías de la Información y las Comunicaciones
UMLS	Unified Medical Language System