



Escuela Técnica Superior de
Ingeniería Informática

TRABAJO FIN DE GRADO

Extrayendo conocimiento a partir de análisis clínico de datos CDM usando la herramienta Atlas

Realizado por
Da. Maria del Valle Alonso de Caso Ortiz

Para la obtención del título de
Grado en Ingeniería de la Salud

Dirigido por
Dr. Julián Alberto García García
Dra. María José Escalona Cuaresma

En el departamento de
Lenguajes y Sistemas Informáticos

Convocatoria de junio, curso 2023/24

A mi padre y a mi madre, por inculcarme la pasión por el estudio y acompañarme incondicionalmente en cada etapa del camino.

Agradecimientos

Nuevamente a mi familia, a mi padre Francisco José Alonso de Caso, a mi madre María del Valle Ortiz y a mis cuatro hermanos: Manuel, Ignacio, Quico y Juan Pablo; por haber sido apoyo incondicional e inspiración de los valores del trabajo, esfuerzo y sacrificio durante mis años de estudio y durante toda mi vida.

A todos mis compañeros de clase que en algún momento me han acompañado y ayudado durante el transcurso del grado de Ingeniería de la Salud y, especialmente, a aquellos que considero mis amigos y amigas, que no solo me han acompañado sino que también han amenizado este camino, llenándolo de diversión y pasión por nosotros y por estos estudios que hemos disfrutado juntos.

A todos los profesores con los que he coincidido, especialmente a Julián y María José, que además han tutelado y supervisado este Trabajo Fin de Grado.

Por último, a todos los profesionales del departamento de Innovación Tecnológica del Hospital Universitario Virgen del Rocío, que me han guiado durante el período de las prácticas curriculares, apostando por esta iniciativa y ayudándome a llevarla a cabo tutorizando y supervisando su desarrollo, especialmente a Silvia y a Carlos.

Resumen

Incluya aquí un resumen de los aspectos generales de su trabajo, en español.

Palabras clave: Palabra clave 1, palabra clave 2, ..., palabra clave N

Abstract

This section should contain an English version of the Spanish abstract.

Keywords: Keyword 1, keyword 2, ..., keyword N

Índice general

1	Introducción, conceptos previos y Motivación	1
1.1.	Introducción	1
1.2.	Contexto	2
1.3.	Estado del Arte	6
1.4.	Motivación	8
2	Objetivos del proyecto	9
2.1.	Objetivos del TFG	9
2.2.	Objetivos Personales	10
2.3.	Trazabilidad de Objetivos	10
3	Gestión del Proyecto	12
3.1.	Participantes del Proyecto	12
3.2.	Estructura de desglose de trabajo	13
3.3.	Estimación de recursos	13
3.4.	Planificación temporal	13
3.5.	Evaluación de costes	13
3.6.	Identificación de riesgos y planes de contingencia	14
4	Metodología	15
5	Marco Teórico	16
5.1.	¿Qué es OHDSI?	16
5.1.1.	Historia	19
5.1.2.	Actualidad	20
5.2.	¿Cómo generar evidencia?	21
5.3.	Estándares	25
5.3.1.	El Modelo de Datos Común	25
5.3.2.	El Vocabulario	27
5.3.3.	Investigación metodológica	28
5.4.	Herramientas	28
5.4.1.	ATLAS	28
5.4.2.	Otras herramientas	28
6	Documento de Requisitos	29
6.1.	Introducción	29
6.2.	Requisitos Funcionales	29
6.2.1.	Diagramas de casos de uso	29
6.2.2.	Descripción del requisito	29
6.3.	Requisitos no Funcionales	29
6.4.	Conclusiones	29

7 Documento de Análisis y Diseño	30
7.1. Introducción	30
7.2. Arquitectura del Sistema	30
7.3. Diagrama de Componentes	30
7.4. Conclusiones	30
8 Plan de pruebas	31
8.1. Introducción	31
8.2. Caso práctico	31
8.2.1. Comprobación calidad datos	31
8.2.2. Realización del estudio	31
8.3. Conclusiones	31
9 Resultados	32
9.1. Lecciones aprendidas	32
9.2. Trazabilidad de objetivos	32
10 Conclusiones	33
Bibliografía	34
A Manual de usuario	38
B Glosario	39

Índice de figuras

5.1. Banner de OHDSI. Extraído de web oficial [1]	16
5.2. Esquema de colaboración en OHDSI. Extraído de la web oficial [1] .	17
5.3. <i>The Journey from Data to Evidence</i> . Extraído del Libro de OHDSI [2] . .	19
5.4. Mapa de colaboradores de OHDSI. Extraído de la web oficial [1] . . .	20
5.5. Banner del Symposium Europeo 2024. Extraído de la web oficial [1] .	21
5.6. Dibujo simple del proceso de extracción de evidencia. Extraído de la web oficial [1]	22
5.7. Ejemplo de la plancha. Extraído de la web oficial [1]	23
5.8. Imagen de los <i>building blocks</i> . Extraída de la web oficial [1]	23
5.9. Alternativas para la implementación de un análisis observacional. Extraído del Libro de OHDSI [2]	24
5.10. Estructura del CDM v5.4. Extraída de la página de github [3]	26
5.11. Captura de pantalla del menú principal de ATHENA	27

Índice de tablas

2.1. Objetivos del Trabajo Fin de Grado	9
2.2. Objetivos personales del alumno	10
2.3. Trazabilidad de objetivos del Trabajo Fin de Grado	11
2.4. Trazabilidad de objetivos personales de la alumna	11
3.1. Descripción del primer participante del proyecto	12
3.2. Descripción del segundo participante del proyecto	12
3.3. Descripción del tercer participante del proyecto	13
3.4. Descripción del cuarto participante del proyecto	13
3.5. Descripción del quinto participante del proyecto	13

Índice de extractos de código

1. Introducción, conceptos previos y Motivación

Este primer capítulo del Trabajo Fin de Grado (TFG) se divide en cuatro secciones: se presenta en 1.1 una introducción descriptiva del mismo, en 1.2 el contexto teórico general, en 1.3 el estado del arte actual y en 1.4 la motivación que trasciende a la realización del trabajo.

1.1. Introducción

Este Trabajo Fin de Grado ha sido realizado por María del Valle Alonso de Caso Ortiz, alumna del grado de Ingeniería de la Salud por la Universidad de Sevilla (US), de la promoción 2020-2024 y bajo la tutela de D. Julián A. García García y Da. Maria J. Escalona Cuaresma, ambos pertenecientes al departamento de Lenguajes y Sistemas Informáticos de la Escuela Técnica Superior de Ingeniería Informática (ETSII) de la misma universidad. Además se realiza en conjunto con el Departamento de Innovación Tecnológica del Hospital Universitario Virgen del Rocío, mediante un convenio de prácticas curriculares de 337 horas, donde han ejercido la tutela D. Silvia Rodríguez Mejías y D. Carlos Luis Parra Calderón. Toda la información técnica relativa al desarrollo del TFG como los Objetivos del proyecto, la Gestión del proyecto y la Metodología empleada se presentan en los capítulos 2, 3 y 4, respectivamente.

Por otra parte, el trabajo abarca una perspectiva teórica muy amplia, de la que se presenta, en primer lugar, en la sección 1.2 de este mismo capítulo, de forma general el paradigma tecnológico y sanitario actual, con sus características y desafíos principales, con el fin de proporcionar al lector conocimientos generales amplios sobre la temática del trabajo. Una visión más orientada a las iniciativas reales que se están llevando a cabo actualmente en este ámbito se presenta en la siguiente sección, 1.3. Además, en esta línea, se presenta en el capítulo 5 una exposición teórica profunda y extensa sobre las herramientas y estándares de la organización Observational Health Data Sciences and Informatics (OHDSI), que es el foco central del trabajo, por lo que se pretende que el lector conozca en profundidad los aspectos importantes de la misma.

Una vez se presentan los conocimientos teóricos generales sobre la organización OHDSI y las herramientas más relevantes que ofrece, el trabajo se especializa en la implementación de la herramienta ATLAS y **la reproducción de un estudio de datos clínicos con la misma**. Concretamente, la herramienta se despliega e implementa a través de la iniciativa *Broadsea*, que origina la denominación *ATLAS Broadsea* de la herramienta a lo largo del TFG. El marco teórico en el que se presenta la herramienta abarca los capítulos 5, 6 y 7.

En tercer lugar, debido a la naturaleza práctica del TFG, por haber sido desarrollado en colaboración con el HUVR, se realiza **una reproducción de un estudio realizado por los investigadores del hospital el estudio es...** (véase 8). Además, se desarrolla el Anexo A como un documento de gran extensión y relevancia que describe en profundidad el proceso de instalación, despliegue y configuración de ATLAS Broadsea. Este Anexo es de gran interés porque no existe en internet ninguna guía completa de estas características sobre la herramienta, aportando un contenido a la comunidad científica y a mis compañeros del departamento de Innovación Tecnológica de gran valor.

Por último, los últimos dos capítulos 9 y 10 presentan una recopilación de resultados y conclusiones, respectivamente, obtenidos al término del desarrollo del TFG. También se adjunta el Anexo B, que consiste en un Glosario de Términos técnicos relevantes para la comprensión del trabajo.

Adicionalmente, por su naturaleza informática, este TFG se ha desarrollado paralelamente a un repositorio de github del proyecto [4], que ha servido como controlador de versiones y como administrador de archivos en la nube, permitiendo almacenar y compartir con el lector archivos relevantes del TFG, ya sean archivos necesarios para el despliegue de la herramienta, archivos producidos durante el análisis o los propios documentos en sí mismos.

1.2. Contexto

En esta sección de contextualización teórica, se presentan los aspectos y características fundamentales del panorama tecnológico-sanitario emergente, con los conceptos necesarios para comprender en profundidad el valor que provee este Trabajo Fin de Grado.

Con este propósito, se presenta el origen de la Industria 4.0 y su influencia en el sector sanitario, denominado Sanidad o Salud 4.0. Del panorama sanitario se describen tres características principales que son el motor del cambio de paradigma y, posteriormente, se destacan dos necesidades fundamentales: la interoperabilidad y la estandarización de las tecnologías sanitarias, especialmente en el tratamiento de los datos de salud. Finalmente se presentan algunos de los desafíos actuales de esta disciplina.

Introducción: Industria 4.0

La Industria 4.0, o cuarta revolución industrial, fue un concepto concebido por el gobierno alemán en noviembre de 2011 como una estrategia tecnológica para abordar el crecimiento industrial proyectado para 2020. Su uso internacional se popularizó en abril de 2013 durante la feria industrial de Hannover (*Hannover Messe*). Este concepto representa la cuarta fase de la industrialización, sucediendo a la mecanización, electrificación e informatización, y destaca la integración digital de tecnologías avanzadas [5]. Se centra principalmente en la digitalización y la

necesaria convergencia entre los sistemas físicos y cibernéticos (*Cyber-Physical Systems, CPS*). Esta integración se busca a través de nuevas tecnologías de la información y telecomunicación (TICs), como el internet de las cosas (*Internet of Things, IoT*), la generación y análisis de datos masivos (*Big Data & Big Data Analytics*), la computación en la nube (*Cloud Computing*) y el auge de la Inteligencia Artificial (IA) [5][6][7]

Características de la Sanidad 4.0

La integración de los principios y tecnologías de la Industria 4.0 en el sector sanitario originó el concepto de Salud o Sanidad 4.0 (del inglés, *Healthcare 4.0*) [7][8]. En este contexto, este nuevo término se presenta como un complejo desafío destinado a abordar los nuevos escenarios generados por la creciente demanda de dispositivos y sistemas médicos más eficaces y alineados con las nuevas TICs y los avances ininterrumpidos en ciencias como la biotecnología y la ingeniería genética. [9].

La Sanidad 4.0 origina un nuevo ecosistema interseccional del que se destacan tres características principales: (1) la provisión continua de cuidado sanitario, (2) la orientación de la medicina hacia el paciente y (3) la prevención y predicción de enfermedades.

1. La provisión continua del cuidado sanitario se basa en el cuidado continuo (*continuum of care*) [10]. Gracias a las nuevas tecnologías de la Industria 4.0, mayoritariamente a las TICs y al IoT, la sociedad se encuentra estrechamente comunicada entre sí de forma prácticamente ininterrumpida. También a raíz de la pandemia del COVID-19 se han acelerado las telecomunicaciones, que en el ámbito sanitario han potenciado el desarrollo de la telemedicina y la salud digital (o *e-Health* [9] a través del desarrollo de programas informáticos para teleconsultas, monitoreo de actividad mediante pulseras o relojes, nuevos implantes inteligentes y un largo etcétera. Con la digitalización y el seguimiento continuo de la salud, los dispositivos médicos que monitorizan a los pacientes en su vida cotidiana generan enormes cantidades de datos médicos de distintas índoles que, además, cada organización recoge con distintos propósitos y estructura, lo que conlleva que los sistemas de salud digital frecuentemente almacenen grandísimas cantidades de datos inconsistentes, incoherentes o inaccesibles entre sí, produciéndose registros electrónicos de salud muy extensos y dispares. [10].
2. La orientación de la medicina hacia el paciente se refiere a la priorización del paciente como objeto central de la provisión de salud [7]. La atención sanitaria cada vez es más específica para cada individuo, gracias al seguimiento remoto de su actividad diaria y al auge de la medicina de precisión. Esta última constituye una nueva disciplina médica que aboga por un estudio clínico detallado que incluya aspectos como genoma, proteoma, condiciones medioambientales o rutina de vida del paciente [11]. La posición del foco de la salud en el paciente, fomentado por la Unión Europea, implica reestructurar el sistema sanitario alrededor del mismo, pues el paciente debe

ser el cliente final, juez y receptor de todos los servicios y aplicaciones de la salud digital [12] [13]. En términos informáticos esto implica la reconfiguración de los sistemas médicos de modo que se recoja de manera central para cada individuo su historial clínico electrónico (HCE) completo, que incluya tanto datos médicos, como farmacéuticos y otros datos de interés.

3. La última característica es que sea preventiva y predictiva en lugar de meramente reactiva. Esto quiere que decir, que a diferencia del enfoque tradicional en el que la medicina es curativa (posterior a la aparición de una enfermedad), se debe transicionar hacia la provisión de salud de manera previa a la aparición de una enfermedad, de modo que esta pueda ser (i) predecida a través del análisis del HCE del paciente y/o exhaustivos análisis de precisión, y (ii) prevenida a través de monitorearización y provisión de tratamientos preventivos en el cuidado continuo de la salud [11]. En esta línea el análisis del historial clínico de un paciente genera un desafío muy complejo por las características inherentes a los datos ya comentadas, es decir, por su complejidad, desorden y extensión, de modo que las técnicas de análisis de datos tradicionales habitualmente resultan insuficientes. La prevención y la predicción se alcanza gracias al constante desarrollo de técnicas y algoritmos cada vez más sofisticados de inteligencia artificial y aprendizaje automático y herramientas cada vez más poderosas de ciencia y análisis de datos masivos.

Principios fundamentales: Interoperabilidad y Estandarización

Estas tres características de la Sanidad 4.0 se edifican sobre dos principios fundamentales de creciente interés internacional: (a) la estandarización y (b) la interoperabilidad de los sistemas médicos. Ambos conceptos están relacionados entre sí mediante una relación causa-consecuencia, según el Institute of Electrical and Electronics Engineers (IEEE, 2013), "la interoperabilidad se hace posible mediante la implementación de estándares"[14].

- a. La implementación de estándares o estandarización consiste principalmente en establecer acuerdos entre las grandes organizaciones de la salud para definir marcos específicos a través de los que estructurar los registros clínicos electrónicos de manera única, reduciendo el desorden y la disparidad de los datos y permitiendo el intercambio de mensajes entre sistemas pertenecientes a distintas organizaciones. La estandarización es un requisito fundamental para alcanzar la interoperabilidad [13]. Actualmente existen muchos estándares reconocidos y utilizados internacionalmente, tales como HL7 (Health Level Seven), DICOM (Digital Imaging and Communications in Medicine), SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms) o IHE (Integrating the Healthcare Enterprise). Con los estándares nace también un concepto importante: el código abierto o *Open Source*. Sin ir más lejos, HL7, la mayor de las organizaciones entre las anteriores comenzó ofreciendo sus servicios de manera privada hasta 2012 cuando se decidió a promover el código abierto liberando la mayor parte de su propiedad intelectual para que pudiera ser accesible de forma gratuita, lo que potenció y

promovió la adopción de estándares y la consecuente interoperabilidad entre las organizaciones sanitarias [14].

- b. La interoperabilidad entre sistemas y datos es el objetivo final de la revolución industrial, tecnológica y sanitaria actual. La necesidad de interoperabilidad en la administración ya había sido identificada desde principios de siglo por la Comisión Europea [15] aunque no fue hasta 2010 que verdaderamente comenzaron las iniciativas para poner en práctica estrategias interoperables. Este mismo año se adoptó el primer Marco Europeo de Interoperabilidad (*European Interoperability Framework, EIF*) junto a los programas Soluciones de interoperabilidad para las administraciones públicas europeas (ISA y ISA²) y tres años más tarde, en 2013, el IEEE definió rigurosamente el concepto como "la habilidad de los sistemas de intercambiar información y utilizar dicha información intercambiada de forma efectiva"[14].

Recientemente, en 2017 la Unión Europea adoptó el nuevo Marco de Interoperabilidad Europea (*new EIF*) a través del cual ofrecer recomendaciones, modelos y guianza a fin de mejorar la calidad de los servicios públicos europeos alegando que "la falta de interoperabilidad es el mayor obstáculo para progresar"[10]. También, en la Comisión Europea del mismo año, se actualizó la definición de interoperabilidad como "la habilidad de las organizaciones de interactuar hacia objetivos mutuamente beneficiosos, involucrando el intercambio de información y conocimiento entre dichas organizaciones a través de los procesos empresariales que soportan." otorgando una importancia cada vez mayor al concepto [13][16][17]. En noviembre de 2018 se lanzó la Red Europea de Datos y Evidencia en Salud (*European Health Data & Evidence Network, EHDEN*) con el objetivo de "abordar los desafíos actuales en la generación de conocimientos y evidencia a partir de datos clínicos del mundo real a escala, para ayudar a los pacientes, médicos, pagadores, reguladores, gobiernos y la industria" [18].

Desafíos en el tratamiento de los datos

No obstante, aún con tantas iniciativas a nivel global y europeo, la transición hacia la interoperabilidad y la estandarización sigue siendo muy dificultosa, debido a la gran complejidad y sensibilidad de los sistemas de información en salud. El tratamiento de datos sanitarios requiere de gestiones muy precisas, con protocolos de ciberseguridad muy estrictos y leyes sobre privacidad y confidencialidad muy bien definidas, que dificultan la implementación coordinada en diferentes regiones. A continuación se presentan algunos de los desafíos en el tratamiento de los datos clínicos expuestos en el Foro de Seguridad y Protección de Datos organizado por la SEIS en 2024 [19] [20]:

- I. Por un lado, la ciberseguridad de los datos clínicos representa un desafío crítico. El creciente auge de amenazas cibernéticas constantes, requiere de actualizaciones y mejoras en las medidas de protección de la información médica. Las instituciones de salud deben estar a la vanguardia en la

implementación de tecnologías de seguridad robustas para salvaguardar la integridad y la confidencialidad de los datos.

- II. De esta forma, la confidencialidad y privacidad de los datos clínicos también conforman per se un desafío relevante. Garantizar que solo las partes autorizadas tengan acceso a la información médica de los pacientes requiere no solo de protocolos tecnológicos sólidos, sino también de una cultura organizacional comprometida con el cumplimiento de las regulaciones de protección de datos y la ética médica. Para ello, además se necesitan protocolos de anonimización y pseudoanonimización de las bases de datos, que garanticen la privacidad de la información.
- III. El consentimiento para el uso secundario de datos clínicos es otro aspecto crucial a considerar. A medida que se exploran nuevas formas de aprovechar los datos para la investigación y la mejora de la atención médica, es fundamental asegurar que los pacientes comprendan y otorguen su consentimiento informado para cualquier uso adicional de su información médica, respetando siempre su autonomía y derechos individuales.
- IV. Por último, la infraestructura tecnológica adecuada es un requisito fundamental para el manejo eficiente de los datos clínicos. La falta de interoperabilidad entre sistemas, la obsolescencia de la tecnología y las limitaciones presupuestarias pueden obstaculizar los esfuerzos para integrar y compartir datos de manera efectiva entre diferentes entidades de atención médica, dificultando así la coordinación y la prestación de servicios de salud centrados en el paciente.

Todos estos desafíos son los puntos débiles de la actual Sanidad 4.0 pero también son los puntos de mayor atención, pues incidiendo se forma especial en ellos se podrá alcanzar una solución global que facilite la provisión de salud y el aprovechamiento de la información clínica.

1.3. Estado del Arte

Frente al panorama sanitario descrito en [1.2](#), se presenta en esta sección las iniciativas que se están llevando a cabo más recientemente para afrontar los desafíos del sector a nivel global, europeo y nacional.

A nivel global, en Estados Unidos, el IEEE ha desempeñado un papel crucial en el desarrollo de estándares para la interoperabilidad de datos en salud, con iniciativas como el estándar IEEE 11073 para dispositivos médicos interoperables. Además, el National Institutes of Health (NIH) y la organización HL7 lideran esfuerzos para promover la colaboración y el intercambio de datos. Otra organización de gran popularidad y presencia en Estados Unidos es OHDSI, que además que se está convirtiendo en un referente a nivel global en el campo de la ciencia de datos en salud.

En China, otra de las grandes potencias, el gobierno ha lanzado múltiples programas y proyectos para mejorar la interoperabilidad de los datos de salud,

como el China Health Information Interoperability Project (CHIIP), que busca establecer estándares y protocolos para la integración de datos de salud en todo el país.

En Europa, desde marzo de 2020 la red europea de datos y evidencia EHDEN colabora con OHDSI para proporcionar un espacio de datos interoperables y estandarizados. La colaboración comenzó con el fin de realizar estudios sobre Covid-19 aunque su relación se mantiene en la actualidad, ejemplo de ello fue la participación de los socios de EHDEN en el Simposio Europeo de OHDSI en junio de 2022. Ese mismo año OHDSI mostraba también su interés en el proyecto, que dió comienzo en 2021, DARWIN EU (*Data Analysis and Real World Interrogation Network European Unión*) [21] para proporcionar evidencia del mundo real de toda Europa sobre enfermedades, poblaciones y los usos y rendimiento de medicamentos [22]. Otra iniciativa más reciente, comenzada en 2023, es el proyecto de EUCAIM (*Cancer Image Europe*) que pretende establecer una red federada interoperable de compartición de imágenes oncológicas. Para la selección del estandar que debe seguir la federación se está considerando la participación de HL7 FHIR o de OHDSI [23].

Por otro lado, la Infraestructura de Servicios Digitales de eSalud (eHDSI) [24] representa un hito crucial en el impulso de la interoperabilidad y la integración de los sistemas de información sanitaria en Europa. Este marco establece estándares y protocolos para facilitar el intercambio seguro y eficiente de datos de salud entre los Estados miembros de la Unión Europea, con el objetivo de mejorar la calidad de la atención médica y promover la movilidad de los pacientes en el espacio europeo de salud digital [25]. También el proyecto European Genomic Data Infrastructure (GDI) [26] busca establecer una infraestructura unificada para gestionar y compartir datos genómicos en Europa, abordando desafíos de interoperabilidad y ética. Su objetivo es promover la colaboración y la innovación en genómica, posicionando a Europa como líder en el uso responsable de datos genómicos para mejorar la salud.

A nivel estatal, España está colaborando en muchas de las iniciativas europeas como EUCAIM o eHDSI, y conforma uno de los nodos de colaboración con OHDSI más grandes de Europa. Muchas organizaciones a lo largo del territorio español ya están colaborando con el estándar de OHDSI como la Agencia Española de Medicamentos y Productos Sanitarios (AEMPS) o Quirónsalud entre otros [27]. En Sevilla, especialmente, la colaboración con OHDSI la llevan a cabo el IBIS (Instituto de Biomedicina de Sevilla), la fundación FISEVI (Fundación para la Gestión de la Investigación en Salud en Sevilla) y los hospitales universitarios Virgen Macarena y Virgen del Rocío, con participación muy importante en el proyecto de EUCAIM y la comunidad de OHDSI. El pasado octubre de 2023 el hospital Macarena celebró el 'Innodata 2023' [28], un congreso nacional sobre investigación de datos en salud, en la que se presentó una ponencia que trató las herramientas y experiencias de OHDSI.

Por otra parte, el Hospital Virgen del Rocio también está participando en estos proyectos innovadores a cargo del departamento de Innovación Tecnológica, siendo esta la sede del estudio práctico que ha acompañado al desarrollo del TFG,

que tratará de aquí en adelante la importancia de la organización OHDSI, su estándar y herramientas.

1.4. Motivación

Mi curiosidad e interés por el mundo de la ciencia de datos ha sido una constante a lo largo de mis cuatro años de estudio y la principal motivación para realizar este trabajo de fin de grado. El origen se sitúa en el primer año de carrera, allá en el 2020, cuando por primera vez el profesor de estadística nos habló a mi y a mis compañeros sobre el 'Big Data' como una disciplina emergente de gran interés a nivel laboral. Esta primera toma de contacto, fue la que me llevó a continuar investigando sobre dicha disciplina y todo lo relacionado con ella. En tercero de carrera tuve la oportunidad de realizar el programa de movilidad ERASMUS al Politecnico di Milano, una de las mejores universidades de ingeniería del mundo [29], por lo que opté a seleccionar el mayor número de asignaturas de Data Science que mi convenio de estudios me permitió. Este año de estudio en Milán confirmó que, lo que había nacido como una mera curiosidad, se había convertido en una pasión, por lo que a mi regreso del Erasmus me decidí a orientar mi carrera profesional y mi TFG hacia el mundo del análisis de datos clínicos, hasta el día de hoy en que este trabajo es escrito.

También ha sido de gran importancia la motivación de mis profesores y tutores de la Escuela Técnica Superior de Ingeniería Informática de la Universidad de Sevilla y la colaboración, mediante el convenio de prácticas, del grupo científico del Departamento de Innovación Tecnológica del Hospital Universitario Virgen del Rocío, quienes confiando en mi me han apoyado, motivado y dado las herramientas y conocimientos necesarios para completar mi formación sobre ATLAS y OHDSI y la informática clínica en general.

2. Objetivos del proyecto

En este capítulo se presentan los objetivos del Trabajo Fin de Grado, consensuados por el alumno, los tutores de la Universidad de Sevilla y los del Hospital Universitario Virgen del Rocío. Se presentan en 2.1 los objetivos generales para el desarrollo del TFG, en 2.2 los objetivos personales del alumno y en 2.3 la trazabilidad de los objetivos definidos.

2.1. Objetivos del TFG

Los objetivos relativos al desarrollo teórico y práctico del TFG son tres y se presentan a continuación en la siguiente tabla:

ID	Descripción
Obj-001	Instalación, configuración y despliegue de ATLAS Broadsea
Obj-002	Estudio teórico minucioso de funcionalidades y arquitectura de OHDSI y ATLAS Broadsea
Obj-003	Estudio de caso práctico de análisis de datos clínicos proporcionados por el hospital

Tabla 2.1: Objetivos del Trabajo Fin de Grado

El Obj-001 consiste en la "Instalación, configuración y despliegue de ATLAS Broadsea" y la redacción de toda la documentación relativa al proceso en el Anexo A del TFG. Este objetivo es de importancia trascendental incluso para el propio TFG, pues el anexo reúne en un único documento inédito información difícilmente accesible y desperdigada en la red, constituyendo un documento de gran relevancia para toda la comunidad científica, especialmente para el equipo del Hospital, que contará con mayor facilidad a la hora de realizar estas tareas sobre Broadsea.

El Obj-002 consiste en el "Estudio teórico minucioso de funcionalidades y arquitectura de OHDSI y ATLAS Broadsea". Este objetivo sí es puramente relativo al TFG, aunque no por ello menos importante, pues proporciona al alumno un marco de fundamentación y comprensión necesario para poder extraer verdadero valor del uso de ATLAS y de la comunidad científica de OHDSI.

El Obj-003 consiste en el "Estudio de caso práctico de análisis de datos clínicos proporcionados por el hospital". Este objetivo está ligado en igual medida al TFG y a las prácticas realizadas en el Hospital, pues consiste en **replicar un estudio ya realizado previamente sobre unos datos proporcionados por el HUVR** pero utilizando, en este caso, las herramientas OHDSI. La colaboración con el hospital en este caso es crucial para el alcance de este objetivo que de forma práctica complementa a la documentación teórica del TFG.

2.2. Objetivos Personales

Los objetivos personales, relativos a la ambición, interés y curiosidad de la alumna son tres y se presentan a continuación en la siguiente tabla:

ID	Descripción
Obj-Pers-001	Aumentar mi conocimiento del estándar OHDSI y sus herramientas
Obj-Pers-002	Aumentar mi conocimiento del mundo del análisis de datos
Obj-Pers-003	Aumentar mi experiencia en el mundo del análisis de dato

Tabla 2.2: Objetivos personales del alumno

El Obj-Pers-001 consiste en "Aumentar mi conocimiento del estándar OHDSI y sus herramientas", pues en origen mi conocimiento sobre la comunidad científica de OHDSI era nulo, y a medida que iba investigando descubría nuevas iniciativas, redes colaborativas y nuevas herramientas de gran utilidad que despertaron un creciente interés sobre la organización, además de la necesaria recopilación de información para estructurar un trabajo coherente y bien fundamentado.

El Obj-Pers-002 consiste en "Aumentar mi conocimiento del mundo del análisis de datos", pues si bien durante mis estudios de grado he aprendido y obtenido grandes conocimientos sobre este sector de las ciencias de datos, el conocimiento nunca sobra, por lo que de este trabajo también se espera aumentar en mayor profundidad los conocimientos teóricos, generales y específicos a ATLAS sobre análisis de datos.

El Obj-Pers-003 consiste en "Aumentar mi experiencia en el mundo del análisis de datos", pues si bien también durante mis estudios de grado he interactuado de forma experimental con este sector de las ciencias de grado, la experiencia nunca sobra, por lo que gracias a la realización de la parte práctica de este trabajo, en colaboración con el grupo de Innovación Tecnológica del Hospital, también se espera adquirir experiencia real en el tratamiento de datos clínicos, específicamente usando ATLAS.

2.3. Trazabilidad de Objetivos

Los objetivos estipulados se han cumplido al término del desarrollo del Trabajo de Fin de Grado, permitiendo elaborar las siguientes tablas o matrices de trazabilidad para cada objetivo. Cada tabla incluye los capítulos del TFG donde se alcanzan los objetivos (según la numeración del índice dle trabajo) y el tiempo en horas invertido a cada uno (extraído del estudio exhaustivo en el capítulo 3):

ID	Secciones del TFG	Tiempo total invertido
Obj-001	Anexo A	horas
Obj-002	Secciones 5, 6, 7 y Anexo A	horas
Obj-003	Sección 8	horas

Tabla 2.3: Trazabilidad de objetivos del Trabajo Fin de Grado

El Obj-001 se desarrolla y alcanza a través del Anexo A del TFG "Manual de instalación, despliegue y configuración de ATLAS Broadsea", para el que se invierte en total horas

El Obj-002 se desarrolla y alcanza a través de los capítulos 5 "Marco teórico específico", 6 "Documento de requisitos" y 7 "Documento de Análisis y Diseño" y el Anexo A del TFG "Manual de instalación, despliegue y configuración de ATLAS Broadsea", para los cuales se invierte horas, horas, horas, respectivamente, sumando un total de horas.

El Obj-003 se desarrolla y alcanza a través de el capítulo 8 "Plan de pruebas", para el cuál se invierte en total horas.

ID	Secciones del TFG	Tiempo total invertido
Obj-Pers-001	Secciones 1, 5 y Anexo A	horas
Obj-Pers-002	Secciones 1, 5, 6, 9, 10	horas
Obj-Pers-003	Secciones 8, 9, 10	horas

Tabla 2.4: Trazabilidad de objetivos personales de la alumna

El Obj-Pers-001 se desarrolla y alcanza a través de los capítulos 1 "Introducción, Contexto y Motivación" y 5 "Estudio Previo" y el Anexo A del TFG "Manual de instalación, despliegue y configuración de ATLAS Broadsea", para los cuales se invierte horas, horas, horas, respectivamente, sumando un total de horas.

El Obj-Pers-002 se desarrolla y alcanza a través de los capítulos 1 "Introducción, Contexto y Motivación", 5 "Estudio Previo", 6 "Documento de requisitos", 9 "Resultados" y 10 "Conclusiones", para los cuales se invierte horas, horas, horas, respectivamente, sumando un total de horas.

El Obj-Pers-003 se desarrolla y alcanza a través de los capítulos 8 "Plan de pruebas", 9 "Resultados" y 10 "Conclusiones", para los cuales se invierte en total horas, horas, horas, respectivamente, sumando un total de horas.

3. Gestión del Proyecto

En este capítulo se presenta toda la información relacionada con la gestión del proyecto de la elaboración del TFG. El capítulo se divide en seis secciones: [3.1 Participantes del Proyecto](#), [3.2 Estructura de Desglose de Trabajo](#), [3.3 Estimación de recursos](#), [3.4 Planificación temporal](#), [3.5 Evaluación de costes](#) y [3.6 Identificación de riesgos y planes de contingencia](#).

3.1. Participantes del Proyecto

Los participantes del proyecto TFG se presentan a continuación mediante una tabla que recoge su nombre, institución a la que pertenece, rol asignado durante la elaboración del proyecto, tareas asignadas durante la elaboración del proyecto e información de contacto.

Es importante destacar que los tres primeros participantes corresponden a la propia alumna y tutores de la Escuela Técnica Superior de Ingeniería Informática de la Universidad de Sevilla y los dos últimos participantes, a los tutores de las prácticas realizadas en el Departamento de Innovación Tecnológica del Hospital Universitario Virgen del Rocío.

Participante	María del Valle Alonso de Caso Ortiz
Institución	Universidad de Sevilla
Rol	
Tareas asignadas	
Información de contacto	

Tabla 3.1: Descripción del primer participante del proyecto

Participante	Julián García García
Institución	Universidad de Sevilla
Rol	Tutor
Tareas asignadas	
Información de contacto	

Tabla 3.2: Descripción del segundo participante del proyecto

Participante	María José Escalona Cuaresma
Institución	Universidad de Sevilla
Rol	
Tareas asignadas	
Información de contacto	

Tabla 3.3: Descripción del tercer participante del proyecto

Participante	Silvia Rodríguez Mejías
Institución	Hospital Universitario Virgen del Rocío
Rol	
Tareas asignadas	
Información de contacto	

Tabla 3.4: Descripción del cuarto participante del proyecto

Participante	Carlos Luis Parra Calderón
Institución	Hospital Universitario Virgen del Rocío
Rol	
Tareas asignadas	
Información de contacto	

Tabla 3.5: Descripción del quinto participante del proyecto

3.2. Estructura de desglose de trabajo

3.3. Estimación de recursos

- PC, licencias windows, office; recursos open-source de OHDSI a través de youtube, github, docker...

3.4. Planificación temporal

- Scrum, planificación por sprints, estimación del tiempo, desviación...

3.5. Evaluación de costes

- PC, licencias windows, office, teams, ATLAS, OHDSI; gastos indirectos (luz)...

3.6. Identificación de riesgos y planes de contingencia

- Quedarme sin wifi para trabajar en latex
- Que se caiga el servidor de latex

4. Metodología

Metodología usada para la gestión del proyecto

- Scrum
- sofIA???

Metodología usada para el desarrollo del proyecto

- Docker, Github - Servidores, bases de datos del Hospital - Herramientas de OHDSI

5. Marco Teórico

En esta sección se muestra un estudio comprensivo del estandar OHDSI utilizado: qué es, su

5.1. ¿Qué es OHDSI?

OHDSI, pronunciado en inglés "Odysee", son las siglas de Observational Health Data Science and Informatics. OHDSI es una organización colaborativa de ciencia abierta cuyo propósito, de forma muy resumida, es mejorar la investigación científico-sanitaria a través de la ciencia de datos y la informática clínica. No obstante, no es solo una organización, sino una comunidad global abierta a todo el que esté interesado y alineado con su misión, visión y objetivos.

La comunidad se asigna por tanto la misión de "mejorar la salud empoderando a una comunidad para generar de manera colaborativa evidencia que promueva mejores decisiones de salud y una mejor atención", y comparte la visión de "un mundo en el que la investigación observacional produzca una comprensión integral de la salud y la enfermedad" [1][2].

Por otra parte, en *El Libro de OHDSI* la organización se define así misma como "una comunidad de ciencia abierta que tiene como objetivo mejorar la salud empoderando a la comunidad para generar de manera colaborativa evidencia que promueva mejores decisiones de salud y mejor atención" [2]. La web oficial presenta otra definición algo diferente, se presenta como "una colaboración de ciencia abierta, interdisciplinaria y de múltiples partes interesadas para resaltar el valor de los datos de salud a través de análisis a gran escala" [1].



Figura 5.1: Banner de OHDSI. Extraído de web oficial [1]

Por tanto, a la pregunta sobre *qué es OHDSI* se puede responder apoyándose en tres características fundamentales: (i) una comunidad o red colaborativa, (ii) de ciencia abierta y (iii) con la finalidad de promover la extracción de evidencia a partir de datos clínicos.

Una comunidad o red colaborativa

La organización es una comunidad, es decir, se presenta abierta a la incorporación de todo aquel que esté comprometido con su misión. Además se muestra siempre abierta e interesada en la incorporación de nuevos colaboradores, lo que muestran constantemente con el eslogan *"Join the Journey"*, en español, "únete a la aventura".

El *Libro de OHDSI* en el capítulo 2 presenta una guía completa de cómo unirse a la comunidad y participar en sus proyectos y eventos. Los proyectos y eventos de OHDSI se realizan a través de una red colaborativa distribuida por todo el mundo, con múltiples nodos en diferentes países y continentes.

Esta red de colaboradores busca conformarse de un gran equipo multidisciplinar, pues se entiende que el propósito que persigue la organización es tan extenso y complejo que es complicado que una única persona albergue todo el conocimiento técnico para desarrollar a la perfección cada etapa de un proyecto, por ello hace especial hincapié en recibir colaboradores expertos en diferentes materias pero que contribuyan al proyecto común de OHDSI.

En el Symposium de 2022 se presentó el esquema de la Figura 5.2 que muestra los cuatro tipos de colaboradores de OHDSI y las áreas de estudio en las que se requiere su participación. No obstante, se detallarán en profundidad más adelante.

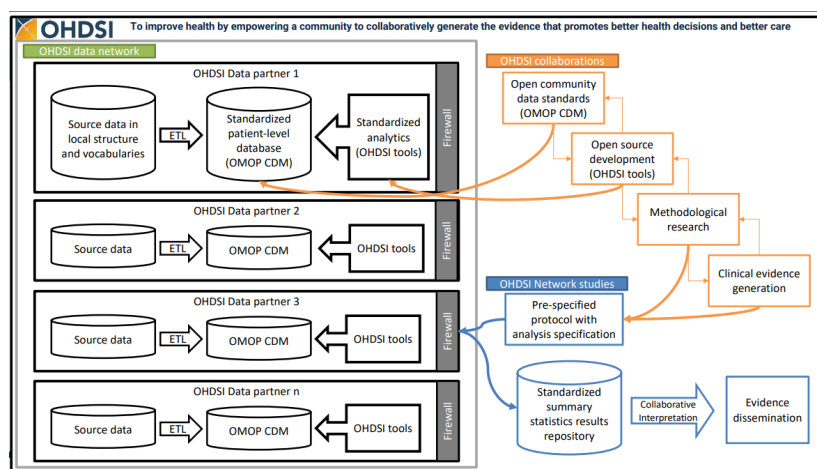


Figura 5.2: Esquema de colaboración en OHDSI. Extraído de la web oficial [1]

Una organización estandarizada y de ciencia abierta

La forma de trabajar de la organización es muy importante, puesto que promueve la estandarización de los estudios metodológicos y la ciencia abierta, evidentemente relacionado con la importancia de la colaboración.

OHDSI promueve la estandarización en dos planos, a través de un modelo común de datos clínicos y un modelo estándar de estudio para extraer evidencia de los datos. Ambos conceptos se presentarán más adelante con mayor detenimiento.

Sin embargo, la estandarización y la colaboración no tienen sentido sin la ciencia abierta. Todos los eventos, publicaciones, herramientas y documentación que elabora OHDSI están disponibles públicamente y de forma gratuita en internet, para que pueda unirse quien quiera (en el caso de los eventos) o consultarse y usarse en cualquier momento (en caso de las herramientas e información). Las dos vías de información por excelencia sobre OHDSI son su página web [1] y el *Libro de OHDSI* [2].

Además, OHDSI asegura la fiabilidad y reproducibilidad de sus estudios a través del cumplimiento de los principios FAIR, que desarrolla en gran extensión en la sección 3.7 de su libro [2].

Por último, como dato de interés, frente a la ciencia abierta, la organización se mantiene económicamente a través del Centro de Coordinación Central, situado en el Centro Médico Irving de la Universidad de Columbia, que es quien asume los costes asociados a la infraestructura central y la coordinación comunitaria por medio del apoyo de los miembros de la comunidad y del patrocinio [1].

El propósito de extracción de evidencia a partir de datos clínicos

Es importante destacar la finalidad de OHDSI de, no solo recopilar y almacenar la información clínica, sino también extraer información o evidencia de ella; lo que se denomina comunmente "el uso secundario de los datos".

La organización identifica la dificultad de extraer información trascendental de los datos clínicos debido a sus distintas morfologías y estructuras en las que son recogidos. Por ello elabora el slogan "*The journey from data to evidence*", en español, "el camino desde los datos hasta la evidencia", para acompañar y facilitar a los investigadores esta ardua tarea.

En el capítulo 1 del Libro de OHDSI se identifican las distintas formas en las que los datos son recogidos y se presentan tres tipos de estudio según el tipo de evidencia que se quiere extraer de ellos.

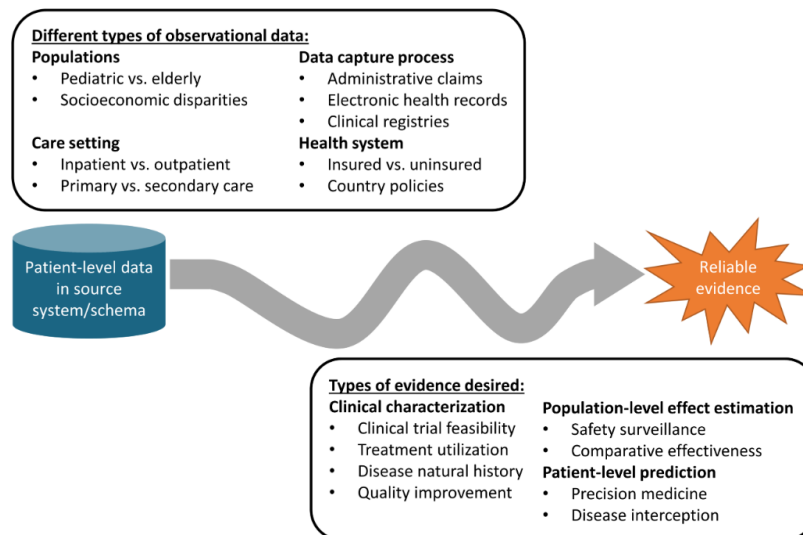


Figura 5.3: *The Journey from Data to Evidence*. Extraído del Libro de OHDSI [2]

Estos son los principios sobre los que se asienta la organización y las herramientas que esta utiliza y proporciona abiertamente a sus colaboradores. Por tanto, es una característica muy importante y que se desarrollará en mayor extensión más adelante.

5.1.1. Historia

Es común encontrar en internet los términos OHDSI y OMOP (*Observational Medical Outcomes Partnership*), utilizados de forma casi indistintiva. Si bien es verdad que OMOP se suele asociar mayoritariamente al CDM (*Common Data Model*) también OHDSI mantiene gran relación con este modelo común de datos. Entonces, ¿cuál es la relación entre estas dos entidades?

La iniciativa de OHDSI se origina en 2014, posterior al proyecto OMOP, que finalizó en 2013, pues la relación que guardan estas dos entidades es parental, OHDSI es la sucesora de OMOP.

OMOP nació en 2008 como una asociación público-privada presidida por la Administración de Alimentos y Medicamentos de EE. UU. con el objetivo de establecer buenas prácticas en estudios observacionales retrospectivos. El proyecto además fue administrado por la Fundación de los Institutos Nacionales de Salud y financiado por un consorcio de compañías farmacéuticas en colaboración con otros investigadores académicos y socios de datos de salud [30]. El propósito inicial de OMOP era impulsar la ciencia de la vigilancia activa de la seguridad de los productos médicos mediante el análisis de datos observacionales de atención médica [30]. Sin embargo, durante su desarrollo, se enfrentó a los desafíos técnicos de llevar a cabo investigaciones en bases de datos observacionales muy heterogéneas entre sí.

El resultado fue el desarrollo de un Modelo Común de Datos (CDM) como un

mecanismo para estandarizar la estructura, el contenido y la semántica de los datos observacionales y hacer posible escribir código de análisis estadístico que fuera reutilizable para estudios en distintas fuentes de datos [31]. Los experimentos de OMOP demostraron la viabilidad de establecer un CDM que además reuniese diferentes vocabularios estandarizados, reuniendo en un mismo estándar diversos tipos de datos de diferentes entornos de atención y representados por diferentes vocabularios de origen. Esta característica facilitó la colaboración y aumentó el interés entre diferentes instituciones lo que promovió o un enfoque de ciencia abierta [2]. OMOP puso todo su trabajo a disposición del público, incluidos diseños de estudio, estándares de datos, código de análisis y hallazgos empíricos, para mejorar la transparencia y fomentar la confianza en su investigación.

Al término del proyecto, el Modelo Común de Datos (CDM) de OMOP había evolucionado hasta respaldar un abanico amplísimo de aplicaciones analíticas, incluida la efectividad comparativa de intervenciones médicas y políticas de todo el sistema de salud, no solo de la industria farmacéutica, por tanto, el equipo de investigación acordó que el fin de dicho proyecto debería ser el origen de uno nuevo. a partir de esta idea nació OHDSI [2].

5.1.2. Actualidad

Por tanto, lo que nació en 2014 como la continuación del proyecto OMOP ha evolucionado hasta convertirse en una extensa red colaborativa global. En la actualidad, la comunidad de OHDSI cuenta con la participación de más de tres mil colaboradores distribuidos en 80 países.

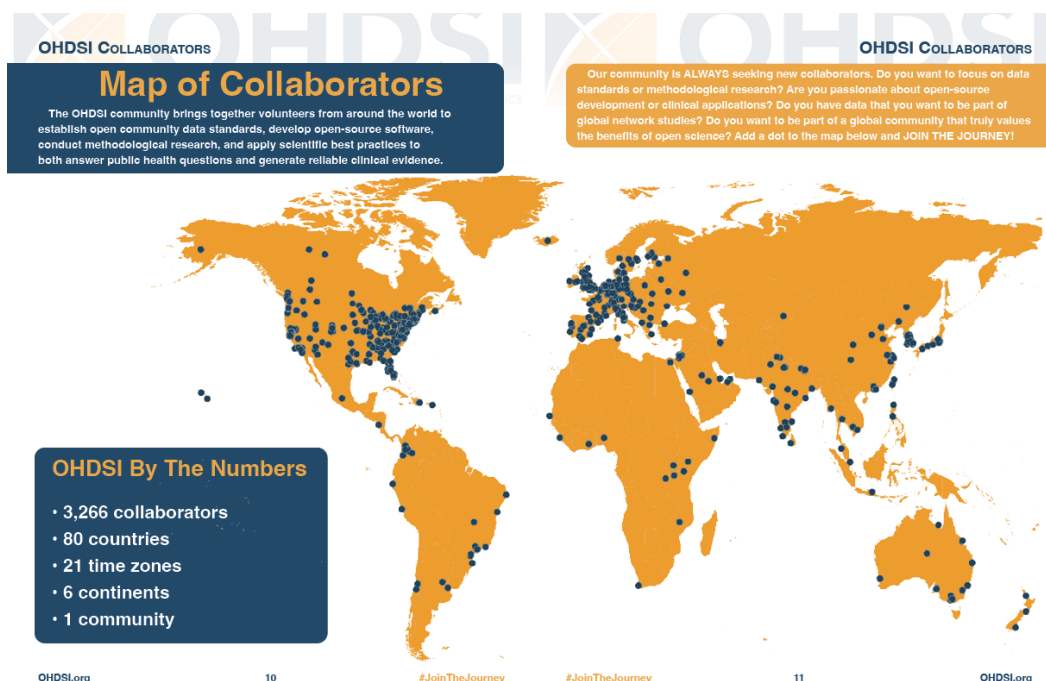


Figura 5.4: Mapa de colaboradores de OHDSI. Extraído de la web oficial [1]

La colaboración con OHDSI se realiza a través de las diferentes fuentes de información que aporta la organización. Por su característica de ciencia abierta, la información sobre OHDSI está espacida por toda la red de internet mediante publicaciones científicas [32], tutoriales para principiantes, grabaciones de las reuniones semanales de la comunidad o las conferencias anuales a través de su canal de youtube [33], canales de mensajería abierta como discord [34] o MS Teams [35], cientos de repositorios de github con información técnica de cada herramienta [36] y los foros de la comunidad para solventar dudas y preguntas [37], entre otros. No obstante, las fuentes de mayor rigor para acceder a la información sobre la organización son la web oficial [1] y el Libro de OHDSI [2].

Además, tal y como se presenta en 1.3, desde que se inició su colaboración con EHDEN (European Health Data Evidence) en 2020, OHDSI está adquiriendo cada vez mayor relevancia a nivel europeo. Ejemplo de ello es la celebración, este mes de junio, en Rotterdam del quinto Symposium Europeo de OHDSI (véase Figura 5.5), que tiene el fin de reunir a los expertos y miembros de la comunidad para presentar los grandes proyectos nacionales y europeos que se están realizando en toda Europa con las herramientas de la comunidad.



Figura 5.5: Banner del Symposium Europeo 2024. Extraído de la web oficial [1]

Por ejemplo, en el Symposium Europeo del pasado año 2023, se presentaron proyectos relativos al almacenamiento de los datos de UCI en Holanda [38], la integración del CDM de OMOP con el laboratorio de datos de salud alemán [39], la estandarización de la base de datos nacional francesa SNDS al modelo de OMOP [40], la armonización de los HCE hospitalarios en Ruanda al CDM [41] y a la estandarización de los datos del registro europeo de sarcomas a OMOP [42], entre otros.

5.2. ¿Cómo generar evidencia?

Una vez que se conoce qué es OHDSI su misión y sus características fundamentales (expuestos en la sección anterior, véase 5.1), se conoce la importancia de generar

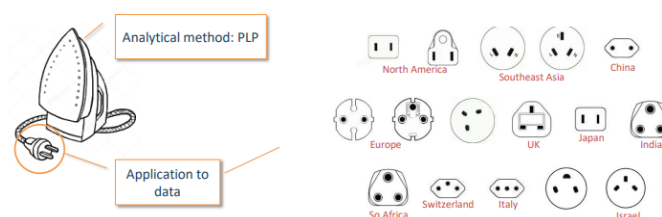
evidencia a partir del estudio de los datos clínicos. No obstante, también se identifican las numerosas dificultades que confronta este proceso, debido a la naturaleza heterogénea de los datos y de los sistemas de información relacionados con su tratamiento (véase 1.2). El complejo camino que se recorre desde el almacenamiento de los datos hasta la extracción de información es lo que la organización denomina *"The Journey from data to evidence"* y se muestra en numerosas ocasiones con el dibujo de la Figura 5.6.



Figura 5.6: Dibujo simple del proceso de extracción de evidencia. Extraído de la web oficial [1]

El camino hacia la generación de evidencia se realiza a través de estudios observacionales o fenotípicos, es decir, que pretenden "simular" lo que sería un estudio clínico experimental pero sobre los datos ya almacenados de pacientes, en vez de realizar un seguimiento en vivo. Además se promueve que estos estudios sigan una misma estructura de modo que sean reciclables y fácilmente reproducibles. De esta forma, OHDSI promueve una vía para generar evidencia interoperable entre las distintas organizaciones que interactúan a través de su red mundial, dicho de otra forma, pretende dar soporte para que miles de estudios diferentes sigan una misma metodología que facilite su comprensión de forma global.

Esta idea se presenta en el Symposium de 2023 con un ejemplo muy intuitivo: la conexión a la corriente eléctrica a través de una plancha. La conexión de la plancha sería la realización de un estudio sobre unos datos, que serían el enchufe a la corriente eléctrica, siendo el objetivo establecer un enchufe estándar que permita la conexión de la plancha a la corriente eléctrica en cualquier lugar del mundo, es decir, la realización de un estudio siguiendo una misma estructura en cualquier lugar del mundo.



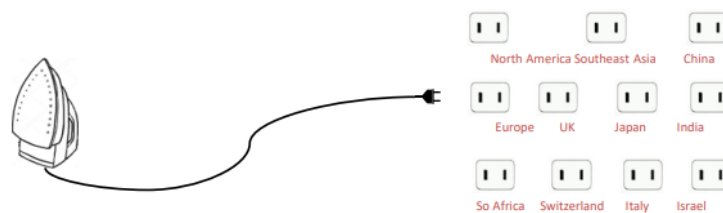


Figura 5.7: Ejemplo de la plancha. Extraído de la web oficial [1]

Para alcanzar este propósito se realiza una estandarización en dos planos: (a) estandarización de los datos clínicos al Modelo de Datos Común de OMOP (véase ??) y (b) estandarización del estudio en sí (véase 5.3.3).

Building blocks

Para comprender de manera general los aspectos fundamentales de cualquier estudio observacional implementado según las recomendaciones de OHDSI, es interesante comprender los siguientes *building blocks* o "bloques de construcción" que utilizados conjunta y correctamente facilitan la generación de evidencia.



Figura 5.8: Imagen de los *building blocks*. Extraída de la web oficial [1]

El primer bloque *databases* corresponde a las bases de datos. El camino hacia la evidencia comienza accediendo a una (o varias) bases de datos estandarizadas al Modelo de Datos Común de OMOP. De este modo se reduce la heterogeneidad en las diferentes fuentes de datos, aumentando la interoperabilidad entre los estudios.

El segundo bloque *phenotypes* corresponde a los fenotipos. Como se ha explicado anteriormente, los estudios que promueve la organización son estudios observacionales sobre características fenotípicas de los individuos. Por tanto, a la

hora de realizar un estudio es importante conocer cuál es el fenotipo que se quiere estudiar y los resultados o *outcomes* que se quieren evaluar. Este bloque presenta como actividad central la **definición de un cohorte**. un cohorte encapsula al conjunto de personas que presentan el/los fenotipo/s que se quiere estudiar. Este término será explicado con mayor detenimiento más adelante. También en el capítulo 10 del Libro de OHDSI [2] se presentan instrucciones e información sobre la definición de cohortes.

Los dos siguientes bloques *study design* y *methods* corresponden al diseño del estudio y la metodología, respectivamente. Los diferentes estudios se realizan a través de las especificaciones de los cohortes, como el período de observación sobre el que se va a realizar el estudio o la designación del comparador del *outcome*, que podrá ser otro cohorte, él mismo o ninguno. El diseño del estudio y la metodología se corresponderá con alguno de los siguientes tres casos de uso: (a) estudios de caracterización de cohortes, (b) estudios de estimación a nivel de población o (c) estudios de predicción a nivel de paciente. Cada uno de estos casos de uso se describen con mayor profundidad en 5.3.3 y también les corresponde un capítulo específico del Libro de OHDSI a cada uno, concretamente los capítulos 11, 12 y 13 [2].

Por último, el bloque *standardized tools* corresponde a las herramientas que ofrece la organización. Como también se ha mencionado anteriormente, OHDSI provee un robusto conjunto de herramientas para cubrir todos los pasos necesarios en el camino hacia la evidencia. Estas herramientas se describen y numeran con mayor detenimiento en 5.4. El hecho de que todas las organizaciones utilicen las mismas herramientas para la realización de estudios e investigaciones también contribuye notoriamente a la interoperabilidad.

Implementación del análisis

Para realizar el análisis *per se* OHDSI distingue tres vías alternativas para generar la evidencia a partir de la base de datos estandarizada al OMOP CDM. Estas tres alternativas se muestran a continuación en la Figura 5.9, extraída del capítulo 8 del Libro de OHDSI.

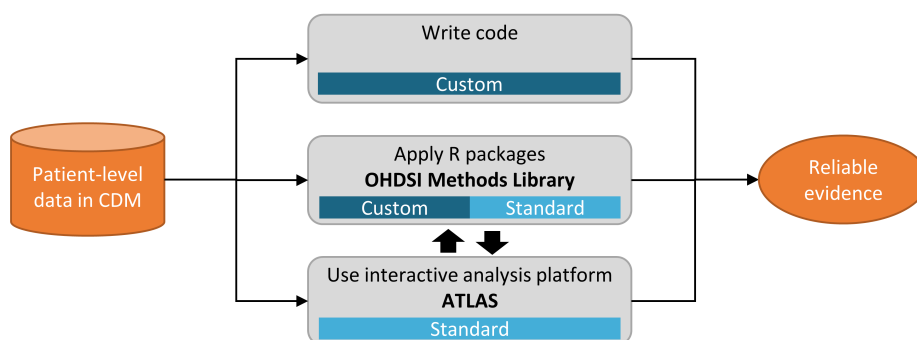


Figura 5.9: Alternativas para la implementación de un análisis observacional. Extraído del Libro de OHDSI [2]

La primera vía *Write code* consiste en extraer la información de la base de datos realizando consultas personalizadas sobre la misma. No hay ningún tipo de estandarización, los analistas escriben código de su propia cosecha utilizando el programa y/o lenguajes de programación que consideren conveniente. Esta vía es muy propensa a errores humanos.

La segunda vía *Apply R packages* consiste en aplicar las librerías estándares que OHDSI ofrece para análisis de datos en R (*OHDSI Methods Library*). De esta forma se hace un balance entre lo personalizado (el código en R) y lo estándar (las librerías), los analistas escriben código personalizado pero utilizan el mismo lenguaje de programación y métodos, aunque quizás distintos programas.

La tercera vía *Use interactive analysis platform* consiste en usar la herramienta interactiva *low-code* de análisis de datos que ofrece OHDSI, denominada **ATLAS**. Esta tercera vía es la vía de implementación que selecciona el TFG puesto que es la que presenta mayor porcentaje de estandarización, todos los analistas utilizan el mismo programa, que utiliza el mismo lenguaje de programación y los mismos métodos. Además, al ser *low-code* el analista no necesita programar nada específicamente, aunque ATLAS sí permite exportar el código que internamente genera (siguiendo siempre los mismos patrones), lo que exponencializa la interoperabilidad entre los estudios.

A partir de este momento se conoce que la implementación del estudio objeto del TFG se realizará a través de una implementación mediante ATLAS, luego toda información a continuación está estrechamente ligada con su utilidad y uso en los análisis de ATLAS. Esta herramienta, junto a otras que también contribuyen a la estandarización del análisis se presentan en mayor detalle en [5.4](#).

5.3. Estándares

En la generación de evidencia es crucial para la interoperabilidad de los estudios que los datos presenten una misma estructuración. Para ello OMOP diseñó dos herramientas fundamentales: el Modelo de Datos Común y el Vocabulario.

5.3.1. El Modelo de Datos Común

El Modelo de Datos Común o *Common Data Model* de OMOP es “un estándar de datos comunitario abierto, diseñado para estandarizar la estructura y el contenido de los datos de observación y permitir análisis eficientes que puedan producir evidencia confiable” [\[3\]](#), en definitiva, es un modelo estándar de estructuración de los datos de salud. La información más relevante y actualizada sobre el CDM se encuentra en su página de github [\[3\]](#) y en el capítulo 4 del Libro de OHDSI [\[2\]](#).

La estructura del CDM está diseñada de forma óptima para servir a la investigación y presenta características muy importantes en este aspecto. En la sección 4.1 del Libro de OHDSI se presentan todas las características del modelo, aunque ahora se destacan las más relevantes:

- Es un modelo centrado en el paciente (alineado con la característica de la Sanidad 4.0 comentada en 1.2), lo que conlleva que todos los eventos y tablas están relacionados con la tabla central del paciente, denominado "Person".
- Limita el acceso a la información personal de los pacientes, evitando en la medida de lo posible el acceso a información sensible como nombres o fechas de nacimiento, para fomentar la protección y privacidad de los datos (que es una dificultad que se identifica generalmente en el tratamiento de datos de salud, véase 1.2). Mayor información al respecto se encuentra en el apartado *Privacidad del paciente y OMOP* de la página de github [3].
- Para fomentar la estandarización e interoperabilidad (véase 1.2) no impone su propia terminología o vocabulario, sino que permite utilizar terminología de vocabularios ya existentes (ej. SNOMED, LOINC...) referenciándolos en su modelo. El conjunto de todos los vocabularios existentes adheridos al modelo de OMOP conforma el Vocabulario.
- El modelo no requiere una tecnología específica sino que puede estructurarse en cualquier base de datos relacional (ej. Oracle, SQL Server...), ajustándose a los requisitos tecnológicos necesarios de cada organización (identificado también como una dificultad en 1.2).

Actualmente el CDM va por la sexta versión, sin embargo, esta aún no está soportada por todas las herramientas de la comunidad, por lo que se sigue sugiriendo el uso del CDM v5.4, que es la última versión completamente funcional. A continuación, en la Figura 5.10 se presenta la estructura lógica de este modelo.

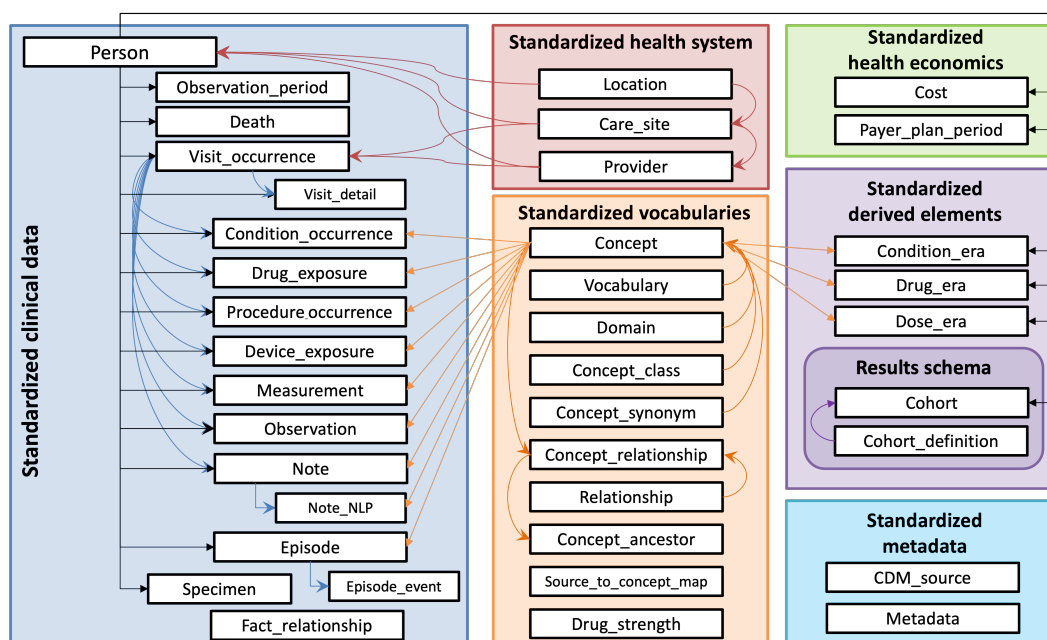


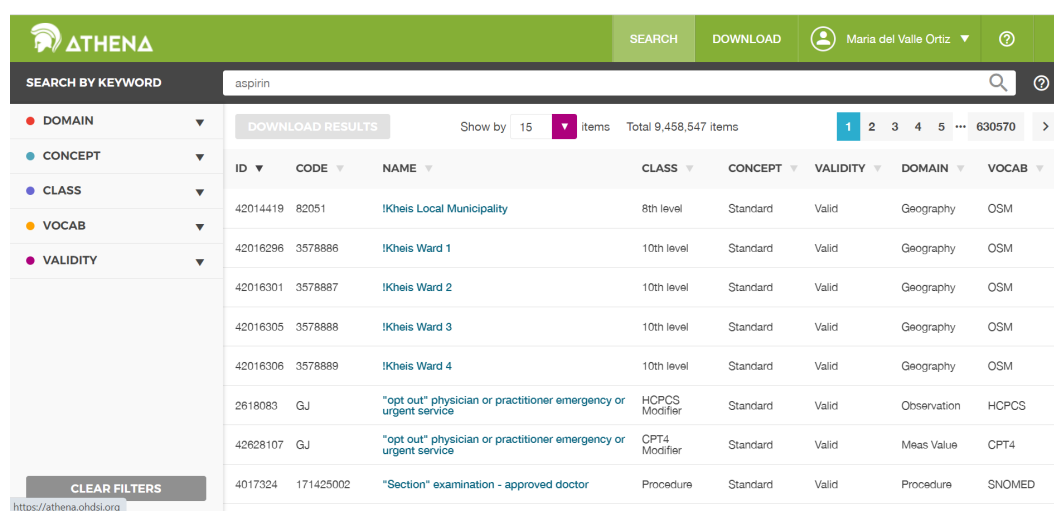
Figura 5.10: Estructura del CDM v5.4. Extraída de la página de github [3]

Explicación más detallada de las tablas y las relaciones entre sí cuando empiece a manejar ATLAS

5.3.2. El Vocabulario

El Vocabulario es uno de los elementos centrales del Modelo de Datos Común de OMOP y una gran herramienta de estandarización e interoperabilidad entre sistemas. Como se comentaba en varias ocasiones, actualmente hay muchos estándares distintos en funcionamiento que establecen las terminologías de los eventos clínicos, como son FHIR, SNOMED CT, RxNorm u otros. El beneficio del Vocabulario de OMOP es que integra todos los vocabularios ya existentes en su estructura a través de referenciación y claves primarias y secundarias.

El Vocabulario de OHDSI, por tanto, es un conjunto de vocabularios y, como todas las herramientas de la comunidad, está disponible online de forma pública. La información sobre el vocabulario se encuentra en el capítulo 5 del Libro de OHDSI [2] y en la página de github del CDM [3]. TamPor otra parte, existe un buscador online de términos en el Vocabulario denominado ATHENA [43].



The screenshot shows the ATHENA web application interface. At the top, there is a green header with the ATHENA logo, a search bar, and buttons for SEARCH and DOWNLOAD. Below the header, a search bar contains the keyword 'aspirin'. To the left of the search results, there is a sidebar with filters for DOMAIN, CONCEPT, CLASS, VOCAB, and VALIDITY. The main area displays a table of search results. The table has columns for ID, CODE, NAME, CLASS, CONCEPT, VALIDITY, DOMAIN, and VOCAB. The results show several entries related to 'aspirin', including 'IKheis Local Municipality', 'IKheis Ward 1', 'IKheis Ward 2', 'IKheis Ward 3', and 'IKheis Ward 4'. There are also entries for 'opt out' physician or practitioner emergency or urgent service and 'Section' examination - approved doctor.

ID	CODE	NAME	CLASS	CONCEPT	VALIDITY	DOMAIN	VOCAB
42014419	82051	IKheis Local Municipality	8th level	Standard	Valid	Geography	OSM
42016296	3578886	IKheis Ward 1	10th level	Standard	Valid	Geography	OSM
42016301	3578887	IKheis Ward 2	10th level	Standard	Valid	Geography	OSM
42016305	3578888	IKheis Ward 3	10th level	Standard	Valid	Geography	OSM
42016306	3578889	IKheis Ward 4	10th level	Standard	Valid	Geography	OSM
2618083	GJ	"opt out" physician or practitioner emergency or urgent service	HCPCS Modifier	Standard	Valid	Observation	HCPCS
42828107	GJ	"opt out" physician or practitioner emergency or urgent service	CPT4 Modifier	Standard	Valid	Meas Value	CPT4
4017324	171425002	"Section" examination - approved doctor	Procedure	Standard	Valid	Procedure	SNOMED

Figura 5.11: Captura de pantalla del menú principal de ATHENA

Actualmente hay más de nueve millones de términos registrados en el Vocabulario de OMOP, como se muestra en la Figura 5.11, y 155 vocabularios distintos coexisten juntos en el estándar.

Cada término regitrado en el vocabulario corresponde a un concepto o *CONCEPT*, según el modelo lógico de la Figura 5.10. Los términos se asocian al vocabulario al que corresponden mediante la tabla *VOCABULARY*

5.3.3. Investigación metodológica

Caracterización

Estimación a nivel de población

Predicción a nivel de paciente

5.4. Herramientas

5.4.1. ATLAS

- Extensa descripción de ATLAS (versión actual, anteriores, uso, aspecto...)
- Diferentes tipos de ATLAS (demo, Broadsea, AWS..)

ATLAS ADEMÁS IMPLEMENTA INTRÍNSICAMENTE DOS HERRAMIENTAS

-ATHENA (herramienta de búsqueda en el vocabulario del CDM) actualmente está implementada dentro de ATLAS/Search

- ACHILLES (data quality dashboard) también esta implementada actualmente dentro de ATLAS/Data source

5.4.2. Otras herramientas

breve descripción de cada una:

-HADES (herramientas de análisis pero en librerías R)

-WHITE-RABBIT y RABBIT-IN-A-HAND (para preparar las ETL)

USAGI (también para la ETL) ...

6. Documento de Requisitos

6.1. Introducción

6.2. Requisitos Funcionales

Hay que readaptar los requisitos para realizar el estudio???????

- RF00: Cargar datasets // De hecho este aún no sabemos cómo hacerlo. Estamos trabajando con datasets que ya vienen cargados
- RF01: Obtener un reporte del data set (Data source)
- RF02 : Definir un conjunto de conceptos del Vocabulario (Concept set)
- RF03: Configurar la muestra de trabajo (cohort definition)
- RF04: Caracterizar el cohort (characterization - primer gran bloque de metodología de OHDSI)
- RF05: Definir una estimación a nivel de población (estimation - segundo gran bloque de metodología de OHDSI)
- RF06: Hacer una predicción a nivel de paciente (prediction - tercer gran bloque de metodología de OHDSI)

Más cosas que se pueden hacer y no definimos la otra vez:

- Obtener un reporte de la ruta del cohorte (cohort pathway)
- Analizar los ratios de incidencia de un outcome (Incidence rate)
- Obtener un reporte de los datos para un paciente concreto (profile)

6.2.1. Diagramas de casos de uso

6.2.2. Descripción del requisito

6.3. Requisitos no Funcionales

6.4. Conclusiones

En este capítulo concluimos que...

7. Documento de Análisis y Diseño

7.1. Introducción

En este capítulo explicaremos...

7.2. Arquitectura del Sistema

Todo el sistema de OHDSI con todas sus herramientas se organiza...

Ampliamente - Plataforma open-source en github: toda la info y código de todo se encuentra aquí

- El modelo de vocabulario que se utiliza

Concretamente para este trabajo, se ha usado el "subsistema" de - OHDSI BROADSEA APPLICATIONS con la base de datos de EUNOMIA

*Se podría extender a más bases de datos pero aún no sabemos cómo hacerlo

- Implementado en el ordenador personal usando DOCKER

7.3. Diagrama de Componentes

7.4. Conclusiones

En este capítulo concluimos que...

8. Plan de pruebas

Este capítulo podría ser más bien Casos prácticos"

8.1. Introducción

8.2. Caso práctico

Reproducción del estudio oncológico realizado por los investigadores del HUVR pero utilizando herramienta ATLAS

8.2.1. Comprobación calidad datos

- Datos omopizados por TFG Paco - Calidad previa y post comprobada tb en TFG Paco

8.2.2. Realización del estudio

Check de los casos de uso/requisitos en el estudio real.

ej de la obtención dle reporte de la BD, ej de la creación de un cohorte concreto para un estudio concreto, ej de la predicción a nivel de paciente para un estudio concreto....

todos los ejemplos anteriores seguirían un mismo hilo conductor en cuanto al ESTUDIO CONCRETO

8.3. Conclusiones

En este capítulo concluimos que...

9. Resultados

9.1. Lecciones aprendidas

- Comprensión de la importancia de la estandarización (estandar OHDSI) en la interoperabilidad de los sistemas clínicos.
- Implementación de un entorno virtual en el PC (Entorno y webAPI de OHDSI en MV DOCKER)
- Aprendizaje de uso de la herramienta ATLAS
- ...

9.2. Trazabilidad de objetivos

10. Conclusiones

Bibliografía

- [1] Observational Health Data Sciences and Informatics. Ohdsi.org. <https://www.ohdsi.org/>, .
- [2] Observational Health Data Sciences and Informatics. The book of ohdsi. <https://ohdsi.github.io/TheBookOfOhdsi.html>, January 11 2021.
- [3] Observational Health Data Sciences and Informatics (OHDSI). Common data model, 2023. URL <https://ohdsi.github.io/CommonDataModel/index.html>.
- [4] vallealonsodc. Thesis-ATLAS-OHDSI. <https://github.com/vallealonsodc/Thesis-ATLAS-OHDSI>, 2024.
- [5] Heiner Lasi, Peter Fettke, Hans-Georg Kemper, Thomas Feld, and Michael Hoffmann. Industry 4.0: Towards future industrial opportunities and challenges. *Business & information systems engineering*, 6:239–242, 2014.
- [6] Chieh-feng Chen, El-Wui Loh, Ken N Kuo, and Ka-Wai Tam. The times they are a-changin’—healthcare 4.0 is coming! *Journal of medical systems*, 44:1–4, 2020.
- [7] Guilherme Luz Tortorella, Flávio Sanson Fogliatto, Alejandro Mac Cawley Vergara, Roberto Vassolo, and Rapinder Sawhney. Healthcare 4.0: trends, challenges and research directions. *Production Planning & Control*, 31(15):1245–1260, 2020.
- [8] Guilherme Luz Tortorella, Tarcísio Abreu Saurin, Flavio S Fogliatto, Valentina M Rosa, Leandro M Tonetto, and Farah Magrabi. Impacts of healthcare 4.0 digital technologies on the resilience of hospitals. *Technological Forecasting and Social Change*, 166:120666, 2021.
- [9] Susana Rubio Martín and Sonia Rubio Martín. ehealth y el impacto de la cuarta revolución industrial en salud, el valor del cuidado. *Enfermería en cardiología: revista científica e informativa de la Asociación Española de Enfermería en Cardiología*, (82):5–9, 2021.
- [10] Angelina Kouroubali and Dimitrios G Katehakis. The new european interoperability framework as a facilitator of digital transformation for citizen empowerment. *Journal of biomedical informatics*, 94:103166, 2019.
- [11] Rocío B Ruiz and Juan D Velásquez. Inteligencia artificial al servicio de la salud del futuro. *Revista Médica Clínica Las Condes*, 34(1):84–91, 2023.
- [12] Christina Ntafi, Stergiani Spyrou, Panagiotis Bamidis, and Mamas Theodorou. The legal aspect of interoperability of cross border electronic health services: A study of the european and national legal framework. *Health Informatics Journal*, 28(3):14604582221128722, 2022.

- [13] Dimitrios G Katehakis and Angelina Kouroubali. A framework for ehealth interoperability management. *Journal of Strategic Innovation and Sustainability*, 14(5):51–61, 2019.
- [14] Reid Berryman, Nathan Yost, Nicholas Dunn, and Christopher Edwards. Data interoperability and information security in healthcare. 2013.
- [15] Comisión Europea. Decisión no 1719/1999/ce del parlamento europeo y del consejo de 12 de julio de 1999 sobre un conjunto de orientaciones, entre las que figura la identificación de los proyectos de interés común, relativo a redes transeuropeas destinadas al intercambio electrónico de datos entre administraciones (ida). Technical report, Comisión Europea, 1999. URL <https://www.boe.es/doue/1999/203/L00001-00008.pdf>.
- [16] Comisión Europea. Marco europeo de interoperabilidad – estrategia de aplicación. Technical report, Comisión Europea, 2017. URL <https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:52017DC0134&from=LT>.
- [17] Cesar Casiano Flores, A Paula Rodriguez Müller, Shefali Virkar, Lucy Temple, Trui Steen, and Joep Cromptvoets. Towards a co-creation approach in the european interoperability framework. *Transforming Government: People, Process and Policy*, 16(4):519–539, 2022.
- [18] EHDEN Consortium. Ehden consortium, 2024. URL <https://www.ehden.eu/>.
- [19] ACTIVIDADES SEIS. Xxi foro de seguridad y protección de datos 2024-14/02/24- tercera sesión debate, feb 2024. URL <https://www.youtube.com/watch?v=x79UKXCh1V8>.
- [20] ACTIVIDADES SEIS. Xxi foro de seguridad y protección de datos 2024-15/02/24- octava sesión, feb 2024. URL <https://www.youtube.com/watch?v=6vbbgR7MUqA>.
- [21] OHDSI. Darwin eu initiative presentation, 2023. URL <https://www.ohdsi.org/darwin-eu-initiative-presentation/>.
- [22] Darwin eu, 2023. URL <https://www.darwin-eu.org/index.php>.
- [23] V. et al Kalokyri. Early release of the data federation framework. 2023. URL https://cancerimage.eu/wp-content/uploads/2023/10/D5.1_Early-release-of-the-Data-Federation-Framework_vf.pdf.
- [24] DigitalHealthEurope. eHDSI - European Health Data Space, 2023. URL <https://digitalhealtheurope.eu/glossary/ehdsi/>.
- [25] Comisión Europea. Servicios electrónicos sanitarios transfronterizos, 2023. URL https://health.ec.europa.eu/ehealth-digital-health-and-care/electronic-cross-border-health-services_es.
- [26] European Genomic Data Infrastructure (GDI) project. European genomic data infrastructure (gdi) project, 2022. URL <https://gdi.onemilliongenomes.eu/>.

- [27] OHDSI. Ohdsi spain, 2024. URL <https://www.ohdsi-europe.org/index.php/national-nodes/spain>.
- [28] Hospital Universitario Virgen Macarena. Innodata 2023 - mejorando la gestión y evaluación sanitaria a través de la innovación en el uso de datos, oct 2023. URL <https://www.hospitalmacarena.es/entrada-blog/innodata2023/>.
- [29] QS International. Politecnico di milano, n.d. URL <https://www.topuniversities.com/universities/politecnico-di-milano>.
- [30] P. E. et al Stang. Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership, 2010.
- [31] J. M. et al Overhage. Validation of a common data model for active safety surveillance research., 2012.
- [32] Observational Health Data Sciences and Informatics. Publications - ohdsi.org. <https://www.ohdsi.org/publications/>, .
- [33] Observational Health Data Sciences and Informatics. OHDSI youtube channel. <https://www.youtube.com/@OHDSI/playlists>, .
- [34] OHDSI discord server invitation. <https://discord.com/invite/xABFWShJYx>, .
- [35] Formulario de Microsoft Office. https://forms.office.com/Pages/ResponsePage.aspx?id=LAAPoyCRq0q6TOVQkC0y1ZyG6Ud_r2tKuS0HcGnqiQZUQ05MOU9BSzEw0ThZVjNQVVFgTDNZREN0iQ1QCN0PWcu, .
- [36] Observational Health Data Sciences and Informatics (OHDSI). Ohdsi github repository, 2023. URL <https://github.com/OHDSI/>.
- [37] OHDSI. Ohdsi forums, 2024. URL <https://forums.ohdsi.org/>.
- [38] A. et al Jagesar. The dutch icu data warehouse: towards a standardized multicenter electronic health record database, 2023. URL https://www.ohdsi.org/wp-content/uploads/2023/07/3-ICUdata-poster_portrait-3-A-Jagesar.png.
- [39] A. et al Finster. Integrating the omop cdm into the ai sandbox of the german health data lab, 2023. URL https://www.ohdsi.org/wp-content/uploads/2023/07/7-Finster_OMOP-at-the-HDL_Poster_2023Symposium-Me-Li.png.
- [40] G. et al Collumeau. Standardization of the french national database snnds in omop-cdm, 2023. URL https://www.ohdsi.org/wp-content/uploads/2023/07/12-Standardization_of_SNDS_Health_Data_Hub-Gaelle-Collumeau.png.
- [41] L. et al Halvorsen. The laisdar project – hospital ehr harmonization in rwanda through mapping to omop cdm; outcome, challenges and lessons learned, 2023. URL https://www.ohdsi.org/wp-content/uploads/2023/07/13-halvorsen_laisdarstatusposter_2023symposium-Lars-Halvorsen.png.

- [42] M. et al van Swieten. Standardizing european sarcoma registry data to the omop common data model, 2023. URL https://www.ohdsi.org/wp-content/uploads/2023/07/15-vanSwieten.Blueberry-OMOP-mapping_2023symposium-Maaike-van-Swieten.png.
- [43] OHDSI. Athena. URL <https://athena.ohdsi.org/search-terms/terms>.
- [44] Wikipedia. Macrodatos - wikipedia, la enciclopedia libre, 2024. URL <https://es.wikipedia.org/wiki/Macrodatos>.
- [45] Varios. Computación en la nube - wikipedia, la enciclopedia libre, 2024. URL https://es.wikipedia.org/wiki/Computaci%C3%B3n_en_la_nube.
- [46] Oracle. ¿qué es el internet de las cosas (iot)?, 2024. URL <https://www.oracle.com/es/internet-of-things/what-is-iot/>. Consultado el 11 de enero de 2024.
- [47] Wikipedia. Internet de las cosas - wikipedia, la enciclopedia libre, 2024. URL https://es.wikipedia.org/wiki/Internet_de_las_cosas.
- [48] Real Academia Española. Real academia española, 2024. URL <https://www.rae.es/>.

A. Manual de usuario

B. Glosario

Datos masivos (*Big Data*): Término que hace referencia a conjuntos de datos tan grandes, variados o complejos que requieren de aplicaciones informáticas y técnicas no tradicionales para ser procesados adecuadamente. También se conoce como Macrodatos [44].

Computación en la Nube (*Cloud Computing*): Red de servidores remotos virtualmente integrados y conectados a internet para almacenar, administrar y procesar bases de datos, otros servidores y software. También conocida como Servicios en la Nube, Nube de cómputo o simplemente "La Nube". [45]

Industria 4.0 (*Industry 4.0*): Concepto acuñado por el gobierno alemán en 2011 para referirse a la emergente cuarta revolución industrial basada fundamentalmente en la integración de los sistemas físicos con Internet a través de herramientas como Internet de las cosas, Big Data, Cloud Computing o Inteligencia Artificial.

Internet de las cosas (*Internet of Things, IoT*): Red de dispositivos, sistemas y servicios que incorporan sensores, software y otras tecnologías que permiten la conectividad avanzada y el intercambio de datos entre sí a través de Internet u otras redes de comunicación [46], [47].

Inteligencia Artificial (*Artificial Intelligence, AI*): Disciplina científica que se ocupa de crear programas informáticos que ejecutan operaciones comparables a las que realiza la mente humana, como el aprendizaje o el razonamiento lógico. [48]

Sanidad 4.0 (*Healthcare 4.0*:

También conocido como Salud 4.0.