



Escuela Técnica Superior de
Ingeniería Informática

TRABAJO FIN DE GRADO

Estudio y aplicación de la herramienta ATLAS de OHDSI para la estandarización de la investigación clínica

Realizado por
Da. María 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

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 y compañeras de clase, a las que me han acompañado y ayudado en algún momento durante el transcurso del grado de Ingeniería de la Salud y especialmente a aquellas que considero mis amigas, a Patricia, Angela, Marta, Gracia, Andrea y Paloma que no solo me han acompañado sino que también han amenizado el camino, llenándolo de diversión y pasión por nosotras y por estos estudios que hemos disfrutado juntas.

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.

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 Carlos.

Resumen

Este trabajo consiste en el estudio y aplicación de la herramienta de análisis de datos ATLAS perteneciente a la organización *Observational Health Data Sciences and Informatics* (OHDSI) con la finalidad de promover la estandarización de la investigación observacional con datos de salud. El proyecto se ha desarrollado en colaboración con el grupo de Informática de la Salud Computacional del Hospital Universitario Virgen del Rocío.

La necesidad de estandarización entre los sistemas informáticos y datos de salud es un aspecto de cada vez mayor relevancia a nivel mundial. En este aspecto, OHDSI se alza como una comunidad de investigadores con la finalidad común de unificar la forma de conducir estudios observacionales, a través del Modelo de Datos Común de OMOP y un complejo ecosistema de herramientas de procesamiento y análisis de datos, entre las que destaca ATLAS, una herramienta de análisis de datos *low-code* que permite ejecutar análisis siguiendo una metodología común.

Por la relevancia de la organización, el proyecto consta de una primera parte en la que se recopila información teórica sobre OHDSI, sus estándares y herramientas. Posteriormente, la investigación se complementa con un caso práctico en el que se reproduce con ATLAS un estudio realizado por el grupo de investigadores del hospital sobre posibles efectos adversos del tratamiento radioterápico en pacientes de cáncer de pulmón.

El proyecto en su totalidad confirma la relevancia de OHDSI en el sector de la informática clínica y los beneficios de la aplicación de la herramienta ATLAS y el Modelo de Datos Común de OMOP en la estandarización de la investigación observacional.

Palabras clave: Observational Health Data Sciences and Informatics, ATLAS, Modelo de Datos Común de OMOP, Broadsea.

Abstract

This work involves the study and application of the data analysis tool ATLAS, which belongs to the organization Observational Health Data Sciences and Informatics (OHDSI), with the aim of promoting the standardization of observational research with health data. The project has been developed in collaboration with the Computational Health Informatics group at the Virgen del Rocío University Hospital.

The need for standardization among health data and information systems is an increasingly important aspect worldwide. In this regard, OHDSI stands out as a community of researchers with the common goal of unifying the conduct of observational studies through the OMOP Common Data Model and a complex ecosystem of data processing and analysis tools, among which ATLAS stands out. ATLAS is a low-code data analysis tool that allows analyses to be performed following a common methodology.

Due to the organization's relevance, the project includes an initial part where theoretical information about OHDSI, its standards, and tools is collected. Subsequently, the research is complemented by a practical case in which a study conducted by the hospital's research group on possible adverse effects of radiotherapy treatment in lung cancer patients is reproduced using ATLAS.

The entire project confirms the relevance of OHDSI in the clinical informatics sector and the benefits of applying the ATLAS tool and the OMOP Common Data Model in the standardization of observational research.

Keywords: Observational Health Data Sciences and Informatics, ATLAS, Broadsea, OMOP Common Data Model

Índice general

1 Descripción del Proyecto	1
1.1. Introducción	1
1.2. Contexto	1
1.3. Estado del arte	6
1.4. Motivación	7
1.5. Estructura de la memoria	7
2 Objetivos del Proyecto	9
2.1. Objetivos del TFG	9
2.2. Objetivos personales	9
3 Gestión del Proyecto	11
3.1. Participantes del proyecto	11
3.2. Planificación temporal	12
3.3. Planificación financiera	15
3.4. Identificación de riesgos y planes de contingencia	18
4 Metodología	20
4.1. SofIA	20
4.2. Scrum	20
4.3. Control de versiones	22
5 Observational Health Data Sciences and Informatics (OHDSI)	24
5.1. Introducción	24
5.2. ¿Qué es OHDSI?	24
5.2.1. Características de la organización	26
5.3. ¿Qué es OMOP?	28
5.4. ¿Cómo generar evidencia?	29
5.4.1. Cohortes	30
5.4.2. Casos de uso para la investigación	32
5.4.3. Vías de implementación del análisis	34
5.5. Conclusiones	35
6 Documento de Requisitos	36
6.1. Introducción	36
6.2. Requisitos funcionales	36
6.2.1. Diagrama de casos de uso	36
6.2.2. Casos de uso del sistema	37
6.3. Requisitos no funcionales	49
6.4. Conclusiones	51

7 Entorno de Trabajo	52
7.1. Introducción	52
7.2. Estándares de OHDSI	52
7.2.1. Modelo de Datos Común de OMOP	52
7.2.2. Vocabulario	56
7.3. Herramientas de OHDSI	57
7.3.1. ATLAS	57
7.3.2. Otras herramientas	62
7.4. Programas informáticos empleados	63
7.5. Conclusiones	65
8 Arquitectura del Sistema	66
8.1. Introducción	66
8.2. Arquitectura teórica del sistema	67
8.3. Arquitectura de Broadsea	68
8.4. Arquitectura de ATLAS Broadsea	70
8.5. Conclusiones	73
9 Caso práctico	74
9.1. Introducción	74
9.2. Estudio realizado por el HUVR	74
9.3. Estandarización del estudio con ATLAS	77
9.3.1. Datos	78
9.3.2. Metodología	78
9.3.3. Resultados	87
9.4. Discusión de resultados	88
9.5. Conclusiones	89
10 Resultados	90
10.1. Trazabilidad de objetivos	91
10.2. Lecciones aprendidas	91
11 Conclusiones	93
Bibliografía	94
A Manual de ATLAS Broadsea	98
B Glosario	99

Índice de figuras

1.1. Esquema de contenidos de la sección 1.2 "Contexto"	2
4.1. Esquema de metodología <i>Scrum</i> . Extraída de [1]	21
5.1. Banner de OHDSI. Extraído de web oficial [2]	25
5.2. Mapa de colaboradores de OHDSI. Extraído de la web oficial [2] . . .	25
5.3. Ejemplo de la plancha. Extraído de la web oficial [2]	27
5.4. Dibujo del proceso de extracción de evidencia. Extraído de la web oficial [2]	28
5.5. <i>The Journey from Data to Evidence</i> . Extraído del Libro de OHDSI [3] . .	29
5.6. <i>The patient journey</i> . Extraído de la página web oficial [3]	31
5.7. "Anatomía de una cohorte". Extraída del Tutorial 2022 publicado en la web oficial [2]	31
5.8. Esquema simplificado de los casos de uso para la investigación en OHDSI. Extraído del Symposium 2023, publicado en la web oficial [2]	32
5.9. Esquema de los casos de uso encuadrado en la historia del paciente. Extraído del Symposium 2023 publicado en la web oficial [2]	33
5.10. Tres vías para la implementación de un análisis observacional. Extraído del Libro de OHDSI [3]	34
6.1. Diagrama de casos de uso	37
6.2. Diagrama de actividad de RF-01:Añadir base de datos	38
6.3. Diagrama de actividad de RF-02: Visualizar Reporte	39
6.4. Diagrama de actividad de RF-03: Definir una cohorte	41
6.5. Diagrama de actividad de RF-04: Definir un grupo de conceptos . . .	43
6.6. Diagrama de actividad de RF-05: Realizar Caracterización	44
6.7. Diagrama de actividad de RF-06: Realizar caracterización	46
6.8. Diagrama de actividad de RF-07: Realizar Predicción a nivel de Paciente	48
6.9. Diagrama de requisitos no funcionales	49
7.1. Estructura del CDM v5.4. Extraída de la página de github del CDM [4]	54
7.2. Modelo Entidad-Relación del CDM v5.4. Extraída de la página de github del CDM [4]	55
7.3. Captura de pantalla del menú principal de ATHENA	57
7.4. Logo de ATLAS. Extraída del repositorio de github [5]	58
7.5. Biblioteca de Métodos OHDSI. Extraída del Libro de OHDSI [3] . . .	59
7.6. Estructura de la WebAPI. Extraída de la wiki de github [6]	60
7.7. Captura de pantalla del menú principal de ATLAS demo	61
8.1. Esquema sencillo de Broadsea. Extraída de [7].	66
8.2. Esquema de arquitectura <i>three-tier</i> en Docker.	67
8.3. Vista general de todos los componentes de Broadsea. Extraída de [7].	69
8.4. Captura de pantalla del menú principal de Broadsea	69
8.5. Captura de pantalla del menú principal de ATLAS Broadsea	71

ÍNDICE DE FIGURAS

8.6. Captura de pantalla de pgAdmin de la estructura postgre del servidor de Broadsea	72
9.1. Seleccionador del reporte que se desea generar sobre sobre la base de datos S31/32 Registry HUVR	79
9.2. Reporte general de la información de la base de datos S31/32 Registry HUVR	79
9.3. Análisis de las rutas de la cohorte de cáncer de pulmón	85
9.4. Ajustes de la comparación para la estimación a nivel de población en ATLAS	86
9.5. Ajustes predeterminados para la estimación a nivel de población en ATLAS	86
9.6. Ajustes para la predicción a nivel de paciente en ATLAS	87

Índice de tablas

3.1. Descripción del primer participante del proyecto	11
3.2. Descripción del segundo participante del proyecto	11
3.3. Descripción del tercer participante del proyecto	11
3.4. Descripción del cuarto participante del proyecto	12
3.5. Descripción del quinto participante del proyecto	12
3.6. Planificación y dedicación real del primer sprint	13
3.7. Planificación y dedicación real del segundo sprint	13
3.8. Planificación y dedicación real del tercer sprint	14
3.9. Planificación y dedicación real del cuarto sprint	14
3.10. Coste estimado de personal del proyecto	16
3.11. Coste real de personal del proyecto	18
3.12. Posibles riesgos y planes de contingencia	19
3.13. Matriz de impacto	19
 4.1. Comparación de características entre metodologías tradicionales y ágiles en proyectos informáticos.	21
6.1. Caso de uso de RF-01:Añadir base de datos	39
6.2. Caso de uso de RF-02:Visualizar Reporte	40
6.3. Caso de uso de RF-03:Definir una cohorte	42
6.4. Caso de uso de RF-04:Definir un grupo de conceptos	43
6.5. Caso de uso de RF-05: Realizar caracterización	45
6.6. Caso de uso de RF-06: Realizar Estimación a nivel de Población	47
6.7. Caso de uso de RF-07: Realizar Predicción a nivel de Paciente	49
6.8. RNF-01: Rendimiento	50
6.9. RNF-02: Seguridad	50
6.10. RNF-03: Usabilidad	50
6.11. RNF-04: Portabilidad	50
6.12. RNF-05: Interoperabilidad	50
6.13. RNF-06: Mantenimiento	51
 7.1. Dominios del CDM v5.4. Extraída del Libro de OHDSI [3]	56
9.1. Recopilación de resultados del estudio del HUVR. Extraída de [8]	77
9.2. Reporte de las condiciones registradas en la base de datos S31/32 Registry HUVR	80
9.3. Reporte de las observaciones registradas en la base de datos S31/32 Registry HUVR	80
9.4. Reporte de los procedimientos registrados en la base de datos S31/32 Registry HUVR	81
9.5. Listado de los 12 grupos de conceptos definidos en ATLAS Broadsea	81
9.6. Listado de las 14 cohortes definidas en ATLAS	82
9.7. Definición del análisis estadístico de las cohortes principales	84

ÍNDICE DE TABLAS

9.8. Definición del análisis de la ruta de la cohorte de cáncer de pulmón	84
10.1. Trazabilidad de objetivos con resultados	91

1. Descripción del Proyecto

Este primer capítulo del Trabajo Fin de Grado (TFG) se divide en cinco secciones: [1.1 Introducción](#), [1.2 Contexto](#), [1.3 Estado del arte](#), [1.4 Motivación](#) y [1.5 Estructura de la memoria](#).

1.1. Introducción

El proyecto consiste en el *Estudio y aplicación de la herramienta ATLAS de OHDSI para la estandarización de la investigación clínica* a través de la realización de un estudio teórico exhaustivo de la organización Observational Health Data Sciences and Informatics (OHDSI) y la aplicación de un caso práctico utilizando la herramienta de análisis de datos ATLAS.

Los contenidos de este capítulo consisten principalmente en la descripción del panorama sanitario y tecnológico actual, de especial relevancia para conocer la importancia de la organización OHDSI en el contexto que envuelve a la informática clínica actual.

En la sección [1.2 "Contexto"](#) se presentan las características de la Sanidad 4.0 y los desafíos del sector en paralelo a las propuestas más relevantes que introduce OHDSI para paliar estas dificultades.

En la sección [1.3 "Estado del arte"](#) se presentan las alternativas a OHDSI más empleadas actualmente a nivel global en términos de estandarización y herramientas de análisis de datos clínicos.

Por último, en la sección [1.4 "Motivación"](#) se presenta la motivación personal de la alumna para realizar el proyecto y la colaboración con el Hospital Universitario Virgen del Rocío en esta labor y en la sección [1.5 "Estructura de la memoria"](#) se expone brevemente la estructura seguida a lo largo de la memoria y los contenidos que se tratan en la misma, incluyendo los anexos.

1.2. Contexto

El contexto en el que se desarrolla el proyecto se caracteriza por el impacto transformador de la Industria 4.0 y las tecnologías que la acompañan en el sector sanitario, que dan lugar a la Sanidad 4.0. De este nuevo paradigma tecnológico-sanitario emergen nuevas necesidades de interoperabilidad entre los sistemas informáticos y desafíos en el tratamiento de la información sanitaria.

Frente a ello, la organización OHDSI se levanta como una solución innovadora y potente para paliar las necesidades de la industria. A continuación, en la Figura [1.1](#)

se presenta un flujo sencillo de los contenidos que se desarrollan esta sección.



Figura 1.1: Esquema de contenidos de la sección 1.2 "Contexto"

1.2.1 La Industria 4.0 y las tecnologías emergentes

La Industria 4.0, o cuarta revolución industrial, fue un concepto concebido por el gobierno alemán en noviembre de 2011. Nace como una estrategia tecnológica para abordar el crecimiento industrial proyectado para 2020 y representa la cuarta fase de la industrialización, sucediendo a la mecanización, electrificación e informatización, y destaca la integración digital de tecnologías avanzadas [9].

Dicho concepto se centra principalmente en la digitalización y la necesaria convergencia entre los sistemas físicos y los sistemas ciberneticos (*Cyber-Physical Systems, CPS*). Esta integración se busca mediante el despliegue de nuevas tecnologías de la información y telecomunicación (TICs), como el tan sonado 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 tremendo auge de la Inteligencia Artificial (IA) [9, 10, 11]

1.2.2 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 (*Healthcare 4.0*) [11, 12]. Esto origina un nuevo paradigma del que se destacan a continuación tres características principales:

- **Cuidado sanitario continuo (*continuum of care*)**. Las tecnologías TIC y el IoT, han permitido a la sociedad estar altamente conectada, lo que ha impulsado el desarrollo de la telemedicina y la e-Salud, especialmente tras la pandemia del COVID-19 [13]. Se han desarrollado numerosos dispositivos portátiles, como pulseras y relojes inteligentes, para monitorear a los pacientes de forma continua tanto dentro como fuera del entorno hospitalario. Estos dispositivos generan de forma casi ininterrumpida grandes cantidades de datos médicos que se combinan con registros clínicos para formar los llamados 'Datos del mundo real' (*Real World Data, RWD*) [14].

La gestión de estas grandes y dispares cantidades de información es una tarea muy compleja. Usualmente los datos se recopilan de distintas formas según su finalidad. OHDSI tiene como objetivo poner fin a la disparidad estructural de la información sanitaria proveyendo un modelo de datos común que permita recopilar los datos con fines observacionales.

- **Centrada en el paciente**. Esta perspectiva enfatiza al paciente como el eje central de la atención sanitaria [11]. Con el avance de la medicina de precisión

y el seguimiento remoto de la actividad diaria, la atención médica se ha vuelto cada vez más personalizada [15]. La Unión Europea promueve esta orientación, exigiendo una reestructuración del sistema sanitario para que el paciente sea el principal beneficiario, evaluador y centro de los servicios de salud digital [16, 17]. Esto implica la implementación de sistemas informáticos que administren el historial clínico electrónico (HCE) completo de cada individuo, incluyendo observaciones de datos médicos, farmacéuticos así como cualquier otro relevante.

En este aspecto, OHDSI presenta un modelo de datos en el que el paciente es el núcleo central y alrededor de él se recoge información clínica interseccional muy diversa.

- **Preventiva y predictiva.** Esta característica implica un enfoque proactivo en la salud en lugar de uno reactivo. La medicina se orienta hacia la prevención de enfermedades, utilizando análisis detallados del historial clínico del paciente y técnicas de aprendizaje automático (*Machine Learning, ML*) para predecir y prevenir enfermedades antes de su aparición [15]. Se emplean algoritmos avanzados de inteligencia artificial y aprendizaje automático, así como herramientas sofisticadas de análisis de datos, para abordar este desafío complejo y evolucionar hacia una atención médica más preventiva y predictiva.

La organización OHDSI presenta técnicas de ML embebidas en su herramienta de análisis por excelencia, ATLAS, expuesta y utilizada en este trabajo.

1.2.3 Necesidad de interoperabilidad

La interoperabilidad entre sistemas y datos es el objetivo principal de la actual revolución industrial, tecnológica y sanitaria. Esta necesidad es fundamental en todos los sectores y sistemas de información de organizaciones públicas y privadas, y ha sido reconocida por la Comisión Europea desde principios de siglo [18]. En 2013, el IEEE definió la interoperabilidad como "la habilidad de los sistemas de intercambiar información y utilizarla de forma efectiva".

Actualmente, el nuevo Marco de Interoperabilidad Europea (*new EIF, 2017*) se encarga de ofrecer recomendaciones para mejorar la calidad de los servicios públicos europeos en términos de interoperabilidad, ya que se considera que "la falta de interoperabilidad es el mayor obstáculo para progresar"[14]. Aunque la clasificación de los tipos de interoperabilidad aún es confusa y no existe una única clasificación concreta [19], la literatura coincide generalmente en tres tipos de interoperabilidad:

- **Interoperabilidad semántica.** 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 la información clínica de manera única. De este modo, se reduce el desorden y la disparidad de los datos, permitiendo el intercambio de mensajes entre sistemas pertenecientes a distintas organizaciones. Además

con los estándares nace también un concepto importante: el código abierto o *Open Source* que facilita el acceso libre a la información y permitir consensuar un estándar común. En este caso, OHDSI aboga por la interoperabilidad semántica aportando un modelo de datos *open-source* que combina su propio estándar con otros estándares utilizados hasta el momento, bajo la premisa "adoptá en vez de inventá" (*Adopt instead of build*).

- **Interoperabilidad técnica.** Este tipo de interoperabilidad pone el foco en la conectividad, comunicación y operación relacionadas con las entidades interactivas y los elementos de tecnológicos de los sistemas informáticos. [19]. La capa técnica abarca las aplicaciones e infraestructuras que vinculan sistemas y servicios, incluyendo especificaciones de interfaz, servicios de interconexión e integración de datos, presentación y intercambio de datos, y protocolos de comunicación segura [20].

Para la interoperabilidad técnica entre sus sistemas, la organización propone diversas formas de implementación de su ecosistema de herramientas, sin imponer una única tecnología con el objetivo de que el usuario configure el entorno que le sea más conveniente.

- **Interoperabilidad organizacional.** Este nivel se centra en la interoperabilidad inter e intra organizacional, en cuanto a la definición común de reglas de negocio, políticas y restricciones, alineación de procesos y las acciones necesarias para hacer que las organizaciones colaboren [21]. También se refiere a cómo los sistemas de los participantes alinean sus procesos, responsabilidades y expectativas para lograr objetivos acordados comúnmente.

OHDSI no solo es una organización científica sino una *red de colaboradores* en la que los integrantes comparten la misma misión, visión y valores.

1.2.4 Desafíos en el tratamiento de los datos

A pesar de las numerosas iniciativas a nivel global y europeo, la transición hacia la interoperabilidad y estandarización en salud sigue siendo muy desafiante debido a la complejidad y sensibilidad de los sistemas de información sanitarios. El manejo de datos médicos requiere gestiones precisas con protocolos de ciberseguridad estrictos y leyes de privacidad y confidencialidad bien definidas, lo que dificulta su implementación coordinada en diferentes regiones.

A continuación, se presentan algunos de los desafíos en el tratamiento de los datos clínicos, identificados en el Foro de Seguridad y Protección de Datos organizado por la SEIS en 2024 [22, 23].

- **Ciberseguridad del sistema.** La ciberseguridad de los datos clínicos representa un desafío crítico debido al creciente auge de amenazas cibernéticas constantes. 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 sus datos.

- **Confidencialidad y privacidad.** La confidencialidad y privacidad de los datos clínicos conforma sin duda un desafío cada vez más relevante. Se necesitan protocolos de anonimización y pseudoanonimización de las bases de datos, que garanticen la privacidad de la información personal de los pacientes además de organizaciones comprometidas con las regulaciones de protección de datos y la ética médica.
- **El uso secundario.** El uso secundario de los datos clínicos consiste en permitir el uso de los datos clínicos con una finalidad distinta de la que fueron recogidos. Cada vez se exploran más formas de aprovechar los grandes volúmenes de información sanitaria recopilada en bases de datos con el objetivo de favorecer la investigación y la mejora de la atención médica. Sin embargo, para ello es fundamental que los pacientes comprendan y otorguen su consentimiento informado para cualquier uso adicional de su información médica, presentándose esto muchas veces como un impedimento en el uso de la información sanitaria.
- **Infraestructura tecnológica.** Por último, la infraestructura tecnológica adecuada es un requisito fundamental para el manejo eficiente de los datos clínicos. La arquitectura de los datos cada vez es más compleja y requiere infraestructuras tecnológicas muy potentes y costosas. Además, la falta de interoperabilidad entre sistemas, la obsolescencia de la tecnología y las limitaciones presupuestarias pueden obstaculizar los esfuerzos para la prestación de servicios TIC de salud.

1.2.5 Propuesta de solución: Observational Health Data Science & Informatics

Ante las necesidades y desafíos del complejo panorama sanitario actual, se propone a la organización **Observational Health Data Science & Informatics (OHDSI)** como la solución óptima a la interoperabilidad en estudios observacionales con datos de salud, a través del Modelo de Datos Común de OMOP y la herramienta de análisis de datos ATLAS.

De esta forma el proyecto pretende demostrar la utilidad y los beneficios de extraer evidencia utilizando las herramientas estandarizadas de OHDSI a través de la estandarización utilizando ATLAS de un estudio clínico sobre los efectos adversos de la radioterapia en pacientes oncológicos, llevado a cabo por el Hospital Universitario Virgen del Rocío.

La relevancia de OHDSI a nivel europeo es innegable, en marzo de 2020, la red de datos y evidencia de la Unión Europea, EHEDEN (European Health Evidence & Data Network) comenzó a colaborar con OHDSI para poner fin a la disparidad de estándares presente en los distintos nodos de la Unión Europea y proporcionar un Modelo de Datos Común y un espacio de datos interoperable y estandarizado para todos. A partir de entonces OHDSI ha comenzado a ganar gran relevancia a través de su participación en proyectos europeos como DARWIN EU (*Data Analysis and Real World Interrogation Network European Unión*, 2022) [?] o EUCAIM (*EUropean Cancer Image*, 2023).

Además a nivel estatal, España conforma uno de los nodos de colaboración con OHDSI más grandes de Europa. Concretamente en Sevilla, 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..

1.3. Estado del arte

En base a lo expuesto anteriormente, aún no existe un consenso entre las grandes potencias mundiales que establezca una solución conjunta. En el ámbito del tratamiento de la información sanitaria, existen numerosas alternativas a OHDSI y organizaciones proveedoras de estándares y herramientas para paliar las necesidades y dificultades del sector. Sin embargo, paradójicamente la presencia de tantas alternativas diferentes es precisamente la principal dificultad para la interoperabilidad.

En el ámbito de la **interoperabilidad semántica**, algunos de los estándares más reconocidos y usados mundialmente son HL7 FHIR (*Health Level Seven - Fast Health Interoperability Resources*), HL7 CDA (*Health Level Seven Clinical Document Architecture*), DICOM (*Digital Imaging and Communications in Medicine*), SNOMED CT (*Systematized Nomenclature of Medicine - Clinical Terms*), IHE (Integrating the Healthcare Enterprise), openEHR (*Open Electronic Health Record*), LOINC (*Logical Observation Identifiers Names and Codes*), RxNorm (Prescription Norm) entre otros.

Solo en España cada comunidad autónoma utiliza un sistema informático sanitario distinto, cuyos datos están estructurados de formas distintas. En Andalucía el sistema de información es DIRAYA. Otros ejemplos son: en Madrid, Historia Clínica Digital de Atención Primaria (HCDSAP); en Cataluña, Sistema de Información de Atención Primaria (SIAP); en la Comunidad Valenciana, Sistema de Información Poblacional de Atención Primaria (SIPAP), en País Vasco, Osabide; en Galicia, SERGAS; entre otros.

Por otro lado, en el ámbito de la interoperabilidad técnica las alternativas son muy diferentes, desde aquellos que realizan análisis totalmente personalizados mediante scripts de código hasta el gran catálogo de software de procesamiento de datos actualmente disponible en el mercado. Los lenguajes de programación que más utilizan los analistas de datos son Python, R y SQL, y se implementan en diferentes entornos de desarrollo como JupyterLab o Jupyter Notebook para Python, Rstudio para R o multitud de plataformas de bases de datos (Oracle, Postgre, BigQuery...). Por otra parte, los software de análisis más extendidos son Tableau, Microsoft PowerBI, SAS, MatLab, Apache Spark, entre otras.

La falta de un estandar común es objeto de investigación en todo el mundo, lo que da lugar a alianzas entre organizaciones y competiciones en proyectos que pretenden dar solución a este aspecto, como por ejemplo la Infraestructura de Servicios Digitales de eSalud (eHDSI) [24] o el proyecto European Genomic Data Infrastructure (GDI) [25] que busca establecer una infraestructura unificada para gestionar y compartir datos genómicos en Europa. OHDSI también es una apuesta

muy interesante en este aspecto aunque todavía le queda un largo trecho hasta posicionarse como el único estándar común.

1.4. Motivación

La principal motivación para realizar este proyecto ha sido mi curiosidad e interés por el mundo de la ciencia de datos a lo largo de mis años de formación universitaria. El origen se sitúa en el primer año de carrera, en 2020, cuando por primera vez me hablaron del análisis de datos clínicos como una disciplina emergente de gran interés a nivel laboral. A partir de este momento continué investigando sobre esta disciplina hasta que en tercero de carrera tuve la oportunidad de realizar el programa de movilidad ERASMUS al Politecnico di Milano y aproveché para seleccionar el mayor número de asignaturas de *Data Science* que mi convenio de estudios me permitió.

Aquel 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 en esta disciplina, hasta el día de hoy en que este Trabajo Fin de Grado es escrito.

El proyecto ha sido realizado por mi, 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 ha realizado en conjunto con el Departamento de Innovación Tecnológica del Hospital Universitario Virgen del Rocío, mediante un convenio de prácticas curriculares a través de la asignatura "Prácticas en Empresa", donde han ejercido la tutela Da. Silvia Rodríguez Mejías y D. Carlos Luis Parra Calderón.

De esta forma, también ha sido de gran importancia la motivación de mis profesores y tutores de la ETSII y compañeros del grupo científico del Departamento de Innovación Tecnológica del hospital, quienes confiando en mi me han apoyado, motivado y guiado durante mi formación sobre ATLAS, OHDSI y la informática clínica en general.

1.5. Estructura de la memoria

La memoria se estructura en diez capítulos y dos anexos que contienen toda la información relevante.

La información propiamente sobre el proyecto se encuentra en los capítulos: 1 "Descripción del Proyecto", 2 "Objetivos del Proyecto", 3 "Gestión del Proyecto" y 4 "Metodología".

A continuación, en el capítulo 5 "Marco Teórico", se presenta la información relevante sobre la organización Observational Health Data Science and Informatics (OHDSI) y su relación con la organización Observational Medical Outcomes Partnership (OMOP).

El capítulo 6 "Documento de Requisitos" presenta un catálogo de requisitos y casos de uso del sistema que utiliza el proyecto para su desarrollo. Este capítulo en conjunto con los capítulos 7 "Entorno de Trabajo" y 8 "Arquitectura del Sistema" proveen un conocimiento completo de las herramientas a tratar durante el proyecto.

Por otra parte, en el capítulo 9 "Caso práctico" se describe el contenido práctico del proyecto, que consiste en la estandarización de un estudio clínico realizado en el HUVR utilizando la herramienta ATLAS.

Por último, los siguientes dos capítulos 10 "Resultados" y 11 "Conclusiones", presentan una recopilación de resultados y conclusiones respectivamente obtenidos al término del desarrollo del TFG.

Adicionalmente, se adjuntan dos anexos. El anexo A "Manual de instalación, despliegue y configuración de ATLAS Broadsea" consiste en una guía de usuario completa sobre la herramienta empleada en el caso práctico, ATLAS Broadsea, y el Anexo B "Glosario de Términos", recopila los conceptos técnicos relevantes para la comprensión del trabajo.

Por su naturaleza informática, este TFG se ha desarrollado paralelamente a un **repositorio de github del proyecto** [26], que ha servido como controlador de versiones y como administrador de archivos en la nube, permitiendo almacenar y compartir con el lector final archivos relevantes al proyecto, ya sean archivos necesarios para el despliegue de la herramienta, archivos producidos durante el análisis o los propios documentos de la memoria y anexos en sí mismos.

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. El capítulo se divide en dos secciones: [2.1 Objetivos del TFG](#) y [2.2 Objetivos Personales](#).

2.1. Objetivos del TFG

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

- 1. Obj-001: Estudio teórico de organización OHDSI y herramienta ATLAS.** Este objetivo proporciona a la alumna un marco de fundamentación y comprensión necesario para poder extraer verdadero valor del uso de ATLAS y de todo el ecosistema de la comunidad científica de OHDSI.
- 2. Obj-002: Instalación, despliegue y configuración de ATLAS mediante Broadsea.** Este objetivo, acompañado de la redacción del Anexo [A](#) "Manual de instalación, despliegue y configuración de ATLAS Broadsea", es de gran importancia, puesto que el Anexo reúne en un único documento información de difícil acceso, desperdigada en la red. Así, constituye un documento de interés para toda la comunidad científica, especialmente para el equipo del Hospital Universitario Virgen del Rocío, que contará con una mayor facilidad a la hora de realizar estas tareas sobre Broadsea.
- 3. Obj-003: Estandarización de caso práctico de análisis de datos clínicos proporcionados por el HUVR.** Este objetivo está ligado en igual medida al TFG y a las prácticas curriculares realizadas en el HUVR, debido a que consiste en estandarizar un estudio realizado anteriormente sobre unos datos oncológicos proporcionados por el hospital pero utilizando, en este caso la herramienta ATLAS. 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:

- 1. Obj-Pers-001: Aumentar mi conocimiento sobre la comunidad OHDSI y sus herramientas.** Este objetivo se debe a que inicialmente mi desconocimiento sobre OHDSI era absoluto. Por tanto, aumentar mi

conocimiento sobre la organización es importante para comprender la utilidad de la misma y de las herramientas que proporciona y poder realizar un trabajo coherente y bien fundamentado.

2. **Obj-Pers-002: Aumentar mi conocimiento del mundo del análisis de datos.** Este objetivo se debe a que, aunque es cierto que durante mis estudios de grado he aprendido y obtenido grandes conocimientos sobre las ciencias de datos, de este trabajo final también se espera aumentar en mayor profundidad los conocimientos teóricos, generales y específicos a una herramienta de gran interés europeo como es ATLAS para el análisis de datos.
3. **Obj-Pers-003: Aumentar mi experiencia laboral analizando datos clínicos.** Este objetivo busca aumentar la experiencia adquirida analizar datos clínicos fuera del marco meramente académico, sino en un entorno de trabajo real, con datos clínicos reales, gracias a la colaboración con el Grupo de Innovación Tecnológica del HUVR.

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 cuatro secciones: [3.1 Participantes del proyecto](#), [3.2 Planificación temporal](#), [3.3 Evaluación de costes](#) y [3.4 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 e información de contacto.

Es importante destacar que los tres primeros participantes corresponden a 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	Jefe de Proyecto & Desarrollador & Analista
Información de contacto	maraloort@alum.us.es

Tabla 3.1: Descripción del primer participante del proyecto

Participante	Julián García García
Institución	Universidad de Sevilla
Rol	Tutor del TFG & Supervisor
Información de contacto	juliangg@us.es

Tabla 3.2: Descripción del segundo participante del proyecto

Participante	María José Escalona Cuaresma
Institución	Universidad de Sevilla
Rol	Tutor del TFG & Supervisor
Información de contacto	mjescalona@us.es

Tabla 3.3: Descripción del tercer participante del proyecto

Participante	Silvia Rodríguez Mejías
Institución	Hospital Universitario Virgen del Rocío
Rol	Tutor de prácticas en empresa
Información de contacto	silvia.rodriguez.mejias@juntadeandalucia.es

Tabla 3.4: Descripción del cuarto participante del proyecto

Participante	Carlos Luis Parra Calderón
Institución	Hospital Universitario Virgen del Rocío
Rol	Supervisor de prácticas en empresa
Información de contacto	carlos.parra.sspa@juntadeandalucia.es

Tabla 3.5: Descripción del quinto participante del proyecto

3.2. Planificación temporal

La planificación temporal se realiza dentro del marco de la asignatura Trabajo de Fin de Grado que consta de 12 créditos ECTS y una duración aproximada de 300 horas. Además, el trabajo se ha realizado de forma lineal y combinada con las prácticas curriculares, de 13.5 créditos ECTS y 337 horas, por lo que la planificación contempla ambas tareas de forma conjunta.

Para ello se realiza una planificación basada en cuatro sprints, de cuatro semanas de duración cada uno (salvo el sprint final que dura seis semanas). De esta forma se pretende tener cada mes un nuevo incremento del proyecto y con ello, un feedback por parte del product owner (véase [4.2 "Scrum"](#)).

El comienzo del proyecto se estima al fin de las vacaciones de navidad, el 10 de enero de 2024. De esta forma, se calcula que la duración del proyecto será de 4 meses, con intención de ser entregado en la primera convocatoria de Trabajo Fin de Grado en mayo de 2024.

Es importante comentar que no todos los sprint estiman un esfuerzo igual, puesto que el producto y sus requisitos van incrementando en cada sprint, de modo que en los primeros sprints se estima una carga de trabajo menor (más ligada a la investigación) mientras que en los últimos sprints, más próximos a la fecha de entrega y con un incremento de producto mayor, se estima mayor esfuerzo.

A continuación se presenta una tabla descriptiva para cada sprint, con la planificación temporal y la dedicación real para cada tarea identificada. Posteriormente se justifica en mayor detalle la desviación en cuanto a dedicación real.

Sprint 1			
Inicio:	Fin:	Esfuerzo estimado:	Esfuerzo real:
Resumen	Este primer sprint se dedicará a la investigación teórica sobre la organización, el modelo de datos, la herramienta y otros aspectos contextuales necesarios para el facilitar la puesta en marcha al inicio de las prácticas curriculares		
Tarea	Categoría	Estimación	Dedicación real:
Lectura del Libro de OHDSI	Investigación	15 horas	15 horas
Visualización de tutoriales de Symposium	Investigación	5 horas	5 horas
Lectura de artículos científicos	Investigación	10 horas	10 horas

Tabla 3.6: Planificación y dedicación real del primer sprint

Sprint 2			
Inicio:	Fin:	Esfuerzo estimado:	Esfuerzo real:
Resumen	Este sprint se dedicará a la primera toma de contacto con la empresa. Se acordarán los objetivos y alcance del proyecto entre tutores de la universidad y el hospital y comenzará la fase de desarrollo del estudio en el ámbito empresarial.		
Tarea	Categoría	Estimación	Dedicación real:
Acuerdo de objetivos y alcance del proyecto	Reunión	5 horas	5 horas
Investigación sobre Broadsea	Investigación	10 horas	20 horas
Instalación y despliegue de Broadsea	Desarrollo	30 horas	35 horas
Configuración de Broadsea	Desarrollo	40 horas	35 horas

Tabla 3.7: Planificación y dedicación real del segundo sprint

Sprint 3			
Inicio: 12/03/2024	Fin: 12/04/2024	Esfuerzo estimado: 90 horas	Esfuerzo real: 100 horas
Resumen	Este sprint se dedicará a la documentación de los contenidos aprendidos durante el segundo sprint además del comienzo de redacción de los contenidos teóricos de la memoria una vez establecidos los objetivos, alcance y objeto de estudio del trabajo.		
Tarea	Categoría	Estimación	Dedicación real:
Revisión del proyecto	Reunión	10 horas	10 horas
Redacción del Anexo A	Documentación	35 horas	35 horas
Redacción de la memoria	Documentación	45 horas	55 horas

Tabla 3.8: Planificación y dedicación real del tercer sprint

Sprint 4			
Inicio: 12/04/2024	Fin: 20/05/2024	Esfuerzo estimado: 95 horas	Esfuerzo real: 100 horas
Resumen	Este último sprint se dedicará a la reproducción del estudio práctico del TFG y la finalización de la redacción de la memoria. Así como de la supervisión y repaso de todo lo tratado.		
Tarea	Categoría	Estimación	Dedicación real:
Revisión final del proyecto	Reunión	15 horas	10 horas
Reproducción del estudio del HUVR	Desarrollo	55 horas	45 horas
Redacción final de la memoria	Documentación	25 horas	45 horas

Tabla 3.9: Planificación y dedicación real del cuarto sprint

En la realidad la planificación se ha visto importantemente alterada debido a dificultades encontradas durante el desarrollo del proyecto, sobre todo relacionadas con la realización del caso práctico sobre el uso de la herramienta ATLAS para estandarizar un estudio clínico del HUVR. Estas dificultades se listan a continuación:

- **Instalación, despliegue y configuración de ATLAS Broadsea.** En primer lugar, cabe destacar la complejidad que ha tenido la tarea de implementar la herramienta debido al surgimiento de numerosos problemas y conflictos, necesitándose más tiempo de lo esperado en la planificación.
- **Montaje de la base de datos.** Por otro lado, hubo un largo período de espera entre que el departamento del HUVR eligió el conjunto de datos adecuado para el proyecto y hasta que finalmente un compañero del equipo estructuró los datos en una base de datos accesible.

- **Aprobación del uso secundario de los datos.** Posteriormente, el proceso de aprobación del uso secundario de la base de datos del HUVR para la realización del TFG también se extendió más de lo previsto, no pudiendo tener acceso a los datos hasta recibir el aprobado del Comité de Ética a principios de abril.
- **Preparación de los datos.** Finalmente, una vez recibido el aprobado y el acceso al servidor del HUVR con la base de datos, se requería de la conversión previa de los datos al Modelo Común de Datos de OMOP, trabajo realizado por mi compañero Francisco Rey Garduño para su Trabajo Fin de Grado.

En conclusión, de forma ajena a lo previsto, los datos fueron recibidos de forma útil la semana del 13 de mayo, a dos semanas de la fecha prevista de entrega del proyecto. Por tanto, aunque las horas total de trabajo no han cambiado, siendo 300 horas en ambos casos, el esfuerzo dedicado a cada sprint sí se ve alterado, modificando el peso de cada tarea en el proyecto. No obstante, esto no ha supuesto dificultades mayores y, animado por los tutores el proyecto ha sido entregado en plazo según la fecha de la primera convocatoria de TFG.

3.3. Planificación financiera

La planificación financiera se realiza de forma similar a la elaboración de un presupuesto sobre el proyecto. Para ello se realizará el cálculo de dos tipos de coste: personal y material. Por último se estimará el coste total y el beneficio.

Coste de personal

Para el coste de personal se tendrán en cuenta los roles definidos previamente (véase [3.1 "Participantes del proyecto"](#)). Concretamente, intervendrán los roles ejercidos por la alumna y se omitirán los roles de tutorización y supervisaje para el cómputo del presupuesto del proyecto.

Por tanto, el proyecto requiere del ejercicio fundamental de tres roles: jefe de proyecto, desarrollador y analista. El jefe de proyecto asume las tareas de comunicarse con los tutores (de la universidad y del hospital), tomar decisiones y acordar objetivos y elaboración de la investigación y desarrollo teórico de la memoria del proyecto. El desarrollador asume las tareas de instalar, desplegar y configurar el sistema así como gestionar y administrar las bases de datos, asegurar el correcto funcionamiento de la herramienta y reflejarlo en la memoria. Por último, el analista realiza las tareas meramente analíticas, se encarga de la reproducción del estudio clínico y su redacción en la memoria haciendo uso de la herramienta una vez instalada.

Los costes de cada rol se calculan por hora, utilizando como referencia el precio medio publicado en la consulta preliminar para perfiles profesionales del ámbito informático [\[27\]](#), considerando la categoría junior para cada uno.

A continuación se presenta en negrita el rol definido en el proyecto seguido de la categoría a la que se ha asociado según el informe de la Junta y el coste total asociado a las horas reales invertidas en sus tareas, según lo estipulado en la planificación temporal (véase 3.2 "Planificación temporal").

- **Jefe de proyecto.** Jefe de proyecto y coordinador junior: 39.16€/h.
- **Desarrollador.** Administrador de la base de datos junior: 35.18€/h.
- **Analista.** Analista funcional de aplicaciones junior: 33.12€/h

Rol	Salario	Tiempo estimado	Coste estimado
Jefe de Proyecto	39.16 €/h	105 h	4111.80 €
Developer	35.18 €/h	115 h	4045.70 €
Analista	33.12€/h	80 h	2649.60 €
Coste total:	10807.10€		

Tabla 3.10: Coste estimado de personal del proyecto

Por tanto, el **coste estimado de personal del proyecto es 11052.30€**.

Coste material

En cuanto a los costes materiales, se distinguen otras tres categorías: costes de amortizaciones, de licencias y de servicios. Es importante recordar que la planificación temporal marca una duración estimada del proyecto de cuatro meses.

En primer lugar, el coste de amortizaciones tendrá en cuenta únicamente el equipo portátil utilizado para el desarrollo del proyecto. Se realizará una amortización lineal en 5 años, con un coste inicial de 1000 € y un valor residual del 20 de este coste inicial que da lugar a un coste de 13.33€/mes.

- **Equipo portátil.** Ordenador con procesador 7th generation y 8 gb de RAM.

$$\text{valor residual} = 1000\text{€} \times 0,20 = 200\text{€} \quad (3.1)$$

$$\text{valor amortización} = \frac{1000\text{€} - 200\text{€}}{60 \text{ meses}} = 13,33\text{€}/\text{mes} \quad (3.2)$$

Por tanto, con una duración de cuatro meses, el **coste total de amortizaciones es 53.32€**

En segundo lugar, el coste de licencias tendrá en cuenta el uso de software de pago. La mayoría de las herramientas utilizadas durante el proyecto poseen un plan gratuito o son gratuitas en sí mismas, a excepción de las siguientes:

- **Licencia de Windows 11 Pro** [28]: 259€
- **Licencia profesional de Enterprise Architect** [29]: 229€
- **Licencia de Latex Estándar** [30]: 19€/mes

$$19\text{ €/mes} \times 4 \text{ meses} = 76\text{ €}. \quad (3.3)$$

Por tanto, el **coste total de licencias es 564€.**

En tercer lugar, los costes de servicios incluyen los gastos por suministro eléctrico, el cual tiene un coste promedio de 0,182 € / KWh (OCU, 2022) . Se estima un consumo medio de 0,3 KWh de los dispositivos electrónicos usados durante el desarrollo.

- **Suministro eléctrico.**

$$0,182\text{ €/kWh} \times 0,3 \text{ kWh} \times 300 \text{ h} = 16,38\text{ €}. \quad (3.4)$$

También es necesario tener en cuenta el servicio de internet, para el que se tiene contratado un servicio de fibra óptica simétrica de 100 megabytes con un coste mensual de 25,70 €:

- **Suministro de internet.**

$$25,70\text{ €/mes} \times 4 \text{ meses} = 102,8\text{ €}. \quad (3.5)$$

Por tanto, el **coste total de servicios es 119,18€**

En total se obtiene un **coste material estimado de 736,50€.**

Coste total y beneficio

Sumando los costes de personal y materiales, se tiene que el **coste total estimado del proyecto asciende a la cifra de 11781,80 €.**

$$11052,30\text{ €} + 729,50\text{ €} = 11781,80\text{ €} \quad (3.6)$$

En cuanto al beneficio que se estima obtener de este proyecto, se computa como el coste total más un 15 % de beneficio íntegro para la empresa. Esto hace un total de 13549,07 € de beneficio total del proyecto, de los cuales 1767.27 € son beneficio íntegro.

Por último, se añadirá un fondo de contingencia frente a riesgos del proyecto que quedará excluido de este coste total. Su finalidad será evitar que alguna desviación o problema pueda provocar una finalización temprana del proyecto. Correspondrá al 10 % del coste total estimado, por lo que se contará con un fondo de contingencia de 1354.90 €.

Desviaciones

El proyecto no ha sufrido desviaciones en cuanto al tiempo de dedicación total, aunque sí ha sufrido desviaciones en el tiempo de desarrollo de las tareas diferentes tareas, provocando desviaciones en el coste de personal.

En cuanto al coste de personal, el coste real ascendería a 11290.00€, es decir una desviación del 2 % sobre el coste estimado.

Rol	Salario	Tiempo real	Coste real
Jefe de Proyecto	39.16 €/h	110 h	4307.60 €
Developer	35.18 €/h	125 h	4397.50 €
Analista	33.12€/h	90 h	2980.80 €
Coste total:	11685.90€		

Tabla 3.11: Coste real de personal del proyecto

Por tanto, el **coste total real del proyecto ascendería 12019.50 €.**

$$11290,00\text{€} + 729,50\text{€} = 12019,50\text{€} \quad (3.7)$$

Esto supone una **desviación de 237.70 €** sobre el coste estimado del proyecto, que correspondía a 11781.80 €. No obstante, esta desviación no ha supuesto ningún riesgo para el proyecto debido a que el valor económico de la desviación no supera el valor reservado en el fondo de contingencia (1354.90€). Por tanto se puede concluir que el proyecto se ha concluido de forma exitosa.

3.4. Identificación de riesgos y planes de contingencia

Por último, para la gestión exitosa de un proyecto es importante identificar los posibles riesgos durante la elaboración del proyecto y premeditar planes de contingencia para actuar contra ellos.

A continuación se muestra en una tabla el conjunto de posibles riesgos identificados, acompañado de una descripción y un plan de contingencia frente al mismo.

ID	Riesgo	Descripción	Plan de contingencia
R-001	Trabajar sin conexión a internet	Puede causarse por problemas técnicos o necesidades circunstanciales que a la hora de trabajar en el proyecto no haya conexión a internet.	Acceder a la última versión de la memoria guardada en el repositorio local de github y trabajar sobre ella sobre un editor de texto plano. Luego subirla al repositorio.
R-002	Caída del servidor de Latex	Puede causarse por problemas técnicos del propio servidor de la organización. En este caso el trabajo que se estaba realizando se vería obligatoriamente interrumpido.	Seleccionar el texto sobre el que se estaba trabajando y abrirlo en un editor de texto plano. Cuando se termine de trabajar subirlo al repositorio de github.
R-003	Cambios sin guardar en Latex	Puede causarse por problemas técnicos. En este caso al reanudar el trabajo se habría perdido el trabajo realizado desde la última conexión a Latex.	Cada vez que se termina de trabajar descargar el archivo zip y subirlo a github para llevar un correcto control de versiones.
R-004	Caída del servidor de Docker	Puede causarse por problemas técnicos. En este caso el servidor docker sería inutilizable y no podría ejecutarse Broadsea.	Realizar las tareas deseadas en ATLAS demo y luego exportar los fragmentos de código del análisis para importarlos a ATLAS Broadsea.
R-005	Solapamiento en el servidor interno de PostgreSQL	Puede causarse debido a la administración interna del sistema, que automáticamente se desasigne el servidor de Broadsea a Postgre y haya conflicto entre varias entidades.	Parar el servicio interno del sistema de Postgre para permitir acceso total a Broadsea en el servidor.

Tabla 3.12: Posibles riesgos y planes de contingencia

Una vez se han identificado los posibles riesgos se pueden ordenar y evaluar en una matriz de impacto, según el impacto que ocasionaría el riesgo y la frecuencia con la que se produce.

		Impacto				
Frecuencia		Insignificante	Menor	Moderado	Mayor	Catastrófico
	Frecuente					
	Probable	R-001				
	Ocasional				R-005	
	Possible	R-002	R-003	R-004		
	Improbable					

Tabla 3.13: Matriz de impacto

Puede observarse que no se ha identificado ningún riesgo catastrófico que pueda suponer un problema real para el desarrollo del proyecto, por lo que este se desarrollara potencialmente de forma segura.

4. Metodología

A continuación se presentan las metodologías empleadas para desarrollar el proyecto. Para la redacción de la documentación y catálogo de requisitos se ha utilizado el software [4.1 SofIA](#) y para la planificación temporal se ha utilizado la metodología de gestión de proyectos [4.2 Scrum](#). El desarrollo completo de la memoria se ha desarrollado mediante [4.3 Control de versiones](#).

4.1. SofIA

SofIA [31] es una metodología web dirigida por modelos, cuyo propósito inicial consistió en brindar respaldo a los requisitos del desarrollo web. Fue implementada por primera vez como NDT aunque actualmente ha evolucionado para ofrecer soporte a todo el ciclo de desarrollo, abarcando fases como estudio de viabilidad, requisitos, análisis, diseño, implementación, pruebas y mantenimiento.

Este proyecto se ha beneficiado enormemente del uso de SofIA especialmente en la fase de requisitos, que es el núcleo de esta metodología y la razón principal para seguir sus técnicas. Estas facilitarán la captura, definición y validación de una amplia variedad de requisitos.

SofIA no solo ofrece técnicas tradicionales como trazabilidad o prototipos, sino que también aborda otros aspectos como la navegación entre componentes. Esto garantiza una conexión entre todos los elementos y evita inconsistencias en el catálogo de requisitos después de una modificación.

Es importante destacar que SofIA cuenta con un alto grado de automatización y se basa en la herramienta profesional Enterprise Architect [29]. En los últimos años, la metodología NDT ha sido ampliamente utilizada como enfoque principal en numerosos proyectos reales de importantes compañías. Entre ellas se destacan entidades públicas como la Consejería de Salud de Andalucía o la Consejería de Cultura de la Junta de Andalucía, así como empresas privadas como Airbus o Everis.

Esta amplia adopción refleja un alto grado de confianza en la metodología, y se garantiza que su aplicación conducirá al proyecto a desarrollarse en un entorno comparable al de cualquier otro proyecto real, como es el caso de este proyecto.

4.2. Scrum

Scrum [1] consiste en una metodología ágil para la gestión y planificación de proyectos informáticos. En el campo de la informática las metodologías ágiles están en auge, cada vez se opta menos por metodologías tradicionales y se apuesta por

estas nuevas metodologías disruptivas. Los motivos y beneficios de ello son muy numerosos, las metodologías ágiles abogan por el cambio continuo y la adaptabilidad frente a la rigurosidad tradicional.

A continuación se presenta una tabla esquemática de beneficios en el uso de metodologías ágiles.

Característica	Metodologías Tradicionales	Metodologías Ágiles
Planificación	Planificación detallada y rígida al inicio del proyecto.	Planificación adaptable y flexible, se adapta a cambios constantes.
Entrega de valor	Entregas al final del proyecto.	Entregas frecuentes de funcionalidades, permitiendo feedback temprano.
Cambio	Cambios difíciles de gestionar, conllevan retrasos y costos adicionales.	Cambios bienvenidos y gestionados de manera eficiente, se incorporan fácilmente al proyecto.
Cliente	Interacción limitada con el cliente.	Colaboración estrecha con el cliente, involucrado en todo el proceso.

Tabla 4.1: Comparación de características entre metodologías tradicionales y ágiles en proyectos informáticos.

Entre las diversas metodologías ágiles, concretamente se ha seleccionado Scrum, que es una solución que se basa en numerosos ciclos iterativos, denominados *sprints*, para el desarrollo incremental del producto final.

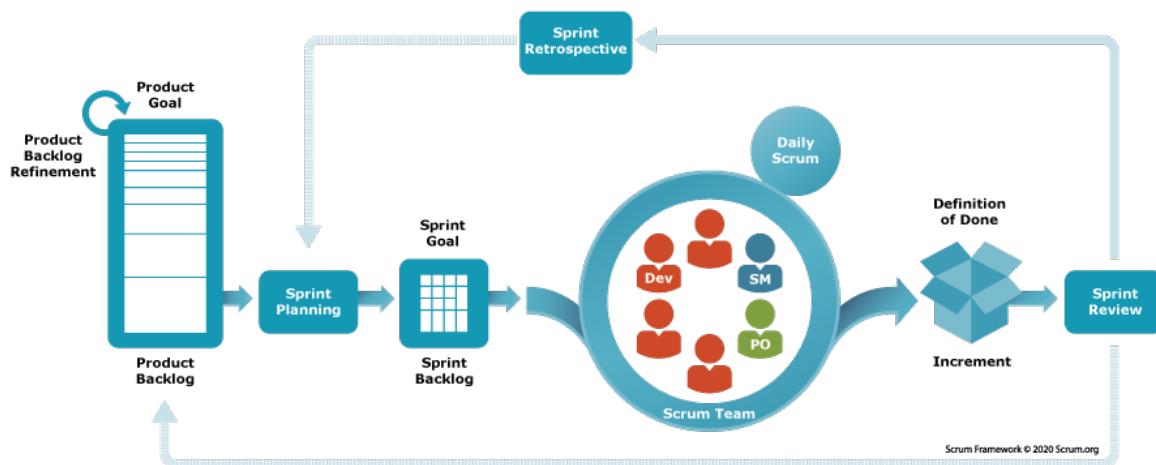


Figura 4.1: Esquema de metodología *Scrum*. Extraída de [1]

En la Figura 4.1 "Esquema de metodología Scrum" se presentan algunos los elementos que intervienen en un proyecto que utiliza la metodología Scrum. A continuación, se identifican los elementos más relevantes de Scrum en este proyecto:

- **Daily Standup:** Reuniones diarias para el seguimiento del proyecto, donde se exponen los avances y problemas o dificultades que se han tenido en el transcurso del ciclo diario del proyecto. En el proyecto, esta práctica se llevó a cabo a través de la monitorización continua realizada por los tutores del HUVR.
- **Product Owner:** Rol responsable de las características del producto y de asegurar que el equipo aporte valor a la empresa. En el proyecto, este rol es desempeñado por el tutor del HUVR, Carlos Parra (véase 3.1 "Participantes del proyecto").
- **Product Backlog:** Lista priorizada de características que debe tener el producto a desarrollar. En el proyecto, este elemento corresponde al catálogo de requisitos definido conjuntamente con los tutores de la universidad, Julián García y María José Escalona (véase 3.1 "Participantes del proyecto").
- **Sprint Backlog:** Conjunto de características escogidas del Product Backlog para implementar en el sprint. En el proyecto, en cada sprint se realizó una división de tareas a realizar, y dentro de cada una se incluyeron un conjunto de subtareas. En esta selección de tareas interviene mayoritariamente la alumna, que es Jefe del Proyecto.
- **Sprint Review:** Reunión en la que se presenta y se evalúa el trabajo realizado durante el sprint, con el objetivo de conseguir la aprobación por parte del cliente. En el proyecto, se considera el cliente a los tutores de la universidad, que son los evaluadores del proyecto.

El resto de los elementos que aparecen en la Figura 4.1 "Scrum" y no se han definido anteriormente, se debe a que no han tenido una aplicación práctica real en el transcurso del proyecto.

4.3. Control de versiones

La memoria del trabajo ha sido redactada empleando LaTeX [32], un sistema de composición de textos de alta calidad que facilita la creación de documentos estructurados y profesionales, ofreciendo herramientas poderosas fácilmente usables e insertables de manera eficiente.

Para gestionar eficazmente las diferentes versiones del documento, se ha utilizado Github como sistema de control de versiones, subiendo diariamente la última versión del trabajo al repositorio de github del proyecto [26]. GitHub proporciona un entorno seguro donde los cambios realizados por el autor pueden ser registrados, rastreados y revertidos si fuera necesario. Esto garantiza una gestión transparente y organizada del proceso de escritura.

La combinación de LaTeX y GitHub no solo ha facilitado la redacción y edición del Trabajo Fin de Grado, sino que también promueve buenas prácticas en cuanto a la gestión de documentos académicos, asegurando la integridad y trazabilidad de cada versión del mismo.

5. Observational Health Data Sciences and Informatics (OHDSI)

Este capítulo presenta el marco teórico sobre OHDSI y se divide en cinco secciones: [5.1 Introducción](#), [5.2 ¿Qué es OHDSI?](#), [5.3 ¿Qué es OMOP?](#) [5.4 ¿Cómo generar evidencia?](#) y [5.5 Conclusión](#).

5.1. Introducción

La organización Observational Health Science and Informatics (OHDSI) es muy importante para el TFG porque es la organización proveedora de la herramienta de análisis ATLAS, núcleo central del trabajo, y por la relevancia que ha adquirido a nivel europeo en los últimos años.

En este capítulo se da a conocer la organización y se identifican los conceptos, ideas y valores fundamentales de la misma. **Es necesario conocer OHDSI para comprender el proyecto en su totalidad y de forma profunda.** Además, satisface el Obj-002 del proyecto (véase [2.1 "Objetivos del TFG"](#)).

A continuación, en la sección [5.2 "¿Qué es OHDSI?"](#) se presenta la visión, misión y valores de la organización y una serie de características fundamentales que la definen.

En la sección [5.3 "¿Qué es OMOP?"](#) se presenta OMOP, la organización predecesora de OHDSI y creadora del conocido *Modelo Común de Datos (CDM)*.

Por último, en la sección [5.4 "¿Cómo generar evidencia?"](#) se presenta la metodología común que promueve la organización para alcanzar la finalidad principal de generar evidencia a partir de datos observacionales. Es muy importante conocer estos conceptos a la hora de conducir un estudio utilizando herramientas OHDSI.

5.2. ¿Qué es OHDSI?

OHDSI, pronunciado en inglés "Odyseey", son las siglas de **Observational Health Data Science and Informatics**. El Libro de OHDSI [3] define la organización 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". En la Figura [5.1 "Banner de OHDSI"](#) se muestra el logo de la organización.



Figura 5.1: Banner de OHDSI. Extraído de web oficial [2]

La **misión** de la comunidad consiste en "mejorar la salud empoderando a una comunidad para generar de manera colaborativa evidencia que promueva mejores decisiones de salud y una mejor atención", y la **visión** consiste en "un mundo en el que la investigación observacional produzca una comprensión integral de la salud y la enfermedad" [2][3].

La organización nació en 2014, como continuación del concluido proyecto OMOP (veáse a continuación 5.3 "¿Qué es OMOP?") y en la actualidad, cuenta con la participación de más de tres mil colaboradores distribuidos globalmente en 80 países.



Figura 5.2: Mapa de colaboradores de OHDSI. Extraído de la web oficial [2]

Haciendo referencia a la Figura 5.2 "Mapa de colaboradores de OHDSI", la presencia en Europa de la organización es innegable. Desde que inició en 2020 su colaboración con la red europea de datos EHDEN (*European Health Data Evidence*), está adquiriendo cada vez mayor relevancia. Ejemplo de ello es la celebración este mes de junio en Rotterdam del quinto Symposium Europeo de OHDSI, con 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.

5.2.1. Características de la organización

Más allá de los aspectos técnicos de la organización, en esta sección se presentan cuatro características inferidas de la investigación sobre OHDSI, que proveen una visión comprensiva de la misma. De esta forma, OHDSI se caracteriza por ser: (i) una comunidad o red colaborativa, (ii) de ciencia abierta, (iii) que promueve la estandarización en salud y (iv) la extracción de evidencia a partir de datos clínicos.

- **Una comunidad o red colaborativa.** La organización es una comunidad abierta a la incorporación de cualquiera que esté comprometido con su misión y valores. Este interés en la incorporación de nuevos colaboradores se muestra constantemente con el eslogan "*Join the Journey*", en español, "únete a la aventura".

La organización distribuye a sus colaboradores en nodos por países y en grupos de trabajo según los diferentes componentes de OHDSI. Por tanto, no se trata de una organización estrictamente burocratizada sino de una unión colaborativa de distintos equipos multidisciplinares que comparten un fin común.

- **Ciencia abierta (*Open science*).** La forma de trabajar de la organización es muy importante, puesto que promueve la colaboración y participación de las organizaciones a través de 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 [2] y el *Libro de OHDSI* [3].

Otras vías de divulgación son publicaciones científicas [33], tutoriales para principiantes, grabaciones de las reuniones semanales de la comunidad o las conferencias anuales a través de su canal de youtube [34]; canales de mensajería abierta como discord [35] o MS Teams [36], cientos de repositorios de github con información técnica de cada herramienta [37] y los foros de la comunidad [38] para solventar dudas y preguntas, entre otros.

Además, en su compromiso con la ciencia abierta, OHDSI asegura la fiabilidad, accesibilidad, interoperabilidad y reproducibilidad de sus estudios a través del cumplimiento de los principios FAIR. Este tema se desarrolla en mayor extensión en la sección 3.7 "OHDSI and the FAIR Guiding Principles" del Libro de OHDSI [3].

- **Promoción de estándares en salud.** OHDSI aboga por estandarizar a un modelo común no solo los modelos de datos sino también la metodología de la investigación médica, con la finalidad de aumentar la interoperabilidad entre los sistemas y organizaciones sanitarias a nivel mundial.

En el Symposium de 2023 se presentó un ejemplo muy intuitivo para divulgar este concepto tan importante: la conexión a la corriente eléctrica a través de

una plancha.

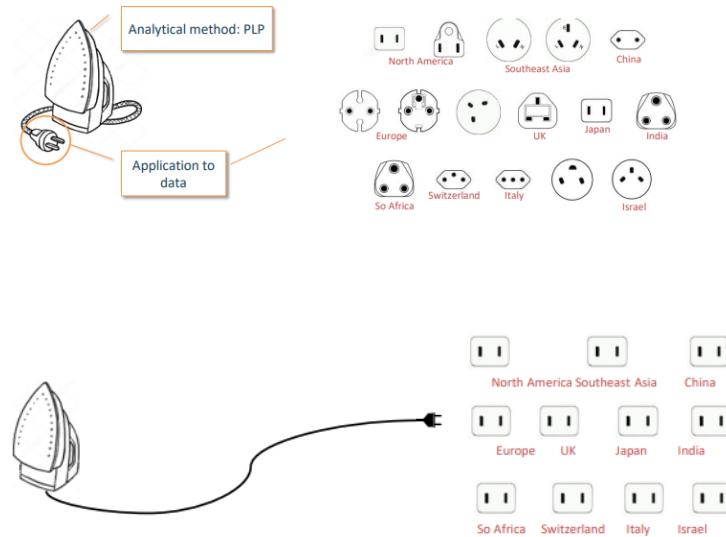


Figura 5.3: Ejemplo de la plancha. Extraído de la web oficial [2]

Como se muestra en la Figura 5.3 “Ejemplo de la plancha”, la plancha sería el diseño de un estudio observacional y el enchufe de pared, la base de datos. En el dibujo de arriba se presenta la problemática actual: un mismo estudio no se puede realizar o “enchufar” a distintas bases de datos porque no comparten la misma estructura. El objetivo de la organización se muestra abajo: estandarizar las bases de datos con una misma estructura para que un mismo estudio pueda aplicarse a diferentes bases de datos.

Con este fin OHDSI promueve el uso del Modelo de Datos Común de OMOP (véase en mayor extensión en 7.2.1 “Modelo de Datos Común”) para estandarizar las bases de datos observacionales. Por otro lado, para conducir los diferentes estudios de forma estandarizada, con el objetivo de fomentar su trazabilidad y reproducibilidad, se ofrecen marcos e instrucciones teóricas sobre cómo conducir los estudios (véase a continuación 5.4 “¿Cómo generar evidencia?”) y herramientas de análisis estandarizadas, como es el caso de ATLAS y otras herramientas (vease en mayor extensión en 7.3 “Herramientas”).

Por tanto, OHDSI se trata de un ecosistema de herramientas y estándares de salud. Este ecosistema se describe en mayor detalle en el capítulo 7 “Entorno de Trabajo”.

- **Extracción de evidencia a partir de datos observacionales.** Es importante destacar que la finalidad de OHDSI no es solo recopilar y almacenar la información clínica de forma estándar, sino también la extracción de información o evidencia de la misma.

El proceso de extracción de evidencia no es sencillo, como se muestra en la Figura 5.4 “Dibujo del proceso de extracción de evidencia”, y parte en un

extremo de las diferentes bases de datos del mundo real (RWD) hacia la obtención fiable de evidencia del mundo real (RWE).



Figura 5.4: Dibujo del proceso de extracción de evidencia. Extraído de la web oficial [2]

La organización se compromete fielmente con este cometido de facilitar la extracción de evidencia a partir de datos observacionales y para facilitar este proceso ofrece de forma abierta todos los estándares y herramientas mencionados anteriormente. Esta es idea es fundamental en OHDSI y se describe en mayor detalle a continuación, en la sección 5.4 “¿Cómo generar evidencia?”.

5.3. ¿Qué es OMOP?

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? Pues bien, **la relación que guardan estas dos entidades es filial, OHDSI (2014-Actualidad) es la sucesora de OMOP (2008-2013)**.

OMOP nació en 2008 como una asociación público-privada presidida por la Administración de Alimentos y Medicamentos de EE. UU. y administrada 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 [39]. El propósito inicial de OMOP fue impulsar la ciencia de la vigilancia activa de la seguridad de los productos médicos mediante el análisis de datos observacionales de atención médica [39]. 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.

Frente a esta problemática, 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 [40]. 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 un enfoque de ciencia abierta [3]. 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 de todo el sistema de salud, no solo de la industria farmacéutica. Finalmente, el equipo de investigación acordó que el fin de dicho proyecto debía ser el origen de uno nuevo y a partir de esta idea nació OHDSI [3].

5.4. ¿Cómo generar evidencia?

La extracción de evidencia a partir de estudios de datos clínicos observacionales es la finalidad fundamental de OHDSI (véase 5.2 "¿Qué es OHDSI?").

Por ello, no es casualidad que la invitación que hace OHDSI a sus colaboradores lleve el slogan "*Join the Journey*" (véase anteriormente 5.2.1 "Características de la organización"), sino que es un guiño al propósito al que se unen: al camino desde los datos hacia la evidencia o, en inglés, "*The Journey from data to evidence*".

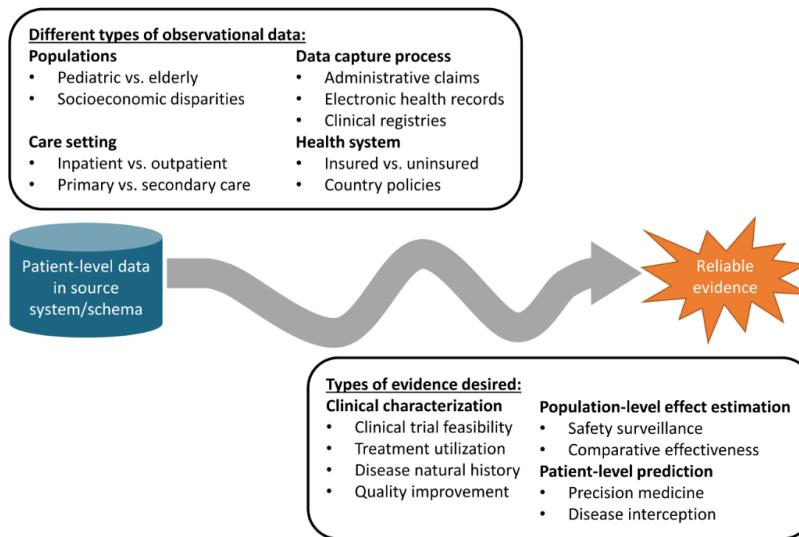


Figura 5.5: The Journey from Data to Evidence. Extraído del Libro de OHDSI [3]

La Figura 5.5 "The Journey from Data to Evidence" complementa a la anterior Figura 5.4 "Dibujo del proceso de extracción de evidencia" añadiendo mayor información en los extremos del recorrido, definiendo cuatro tipos distintos de bases de datos observacionales y tres tipos de evidencia que se quiere generar: la

caracterización clínica (*clinical characterization*), la estimación de efectos a nivel de población (*Population-level effect estimation*) y la predicción a nivel de paciente (*Patient-level prediction*). Estos tres "casos de uso" se presentan en mayor profundidad a continuación en [5.4.2 "Casos de uso para la investigación"](#).

Con ello la organización define un marco para llevar a cabo **estudios observacionales o fenotípicos** sobre datos. Un estudio observacional es una investigación que observa y recopila información sobre individuos o fenómenos sin intervenir en ellos. En el caso del estudio sobre datos, en vez de realizar seguimientos de estudios clínicos en vivo, se simulan estos estudios sobre una base de datos. Cuando la evidencia se extrae sobre datos del mundo real (*RWD*), se denomina evidencia del mundo real (*Real World Evidence, RWE*).

En OHDSI la conducción de estos estudios observacionales se realiza mediante el diseño y estudio de cohortes en la base de datos. Concretamente, se trata de **estudios de cohortes retrospectivos** porque los sujetos se estudian después de haberse producido la enfermedad, utilizando para ello bases de datos que tengan registrada información histórica de la enfermedad y de los factores de riesgo que hayan podido provocar dicha enfermedad.

A continuación en [5.4.1 "Cohortes"](#) se presenta este concepto en profundidad.

5.4.1. Cohortes

El componente central de cualquier investigación en OHDSI es el paciente, del que se recopilan las denominadas "historias del paciente". **Para cada evento clínico que sucede se recoge una historia del paciente o *Patient Journey***. Las investigaciones observacionales se diseñan para extraer información sobre la recopilación de todas las historias de paciente registradas en la base de datos.

La historia del paciente, como se muestra en la Figura [5.6 "The patient journey"](#), es por tanto, una ventana temporal que recoge un evento clínico que le sucede a un paciente en un período de tiempo concreto. El evento se describe mediante tres períodos de tiempo: la enfermedad (rojo), el tratamiento (naranja) y el efecto (verde); y a partir de distintas características como enfermedades (*conditions*), medicamentos (*drugs*), procedimientos (*procedures*) y pruebas (*measurements*).

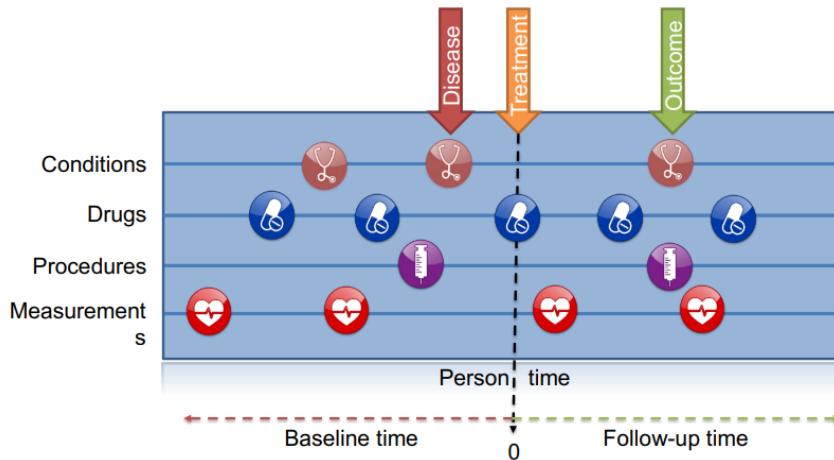


Figura 5.6: The patient journey. Extraído de la página web oficial [3]

Los pacientes se pueden agrupar en **cohortes** cuando comparten historias y características similares, al igual que a la hora de realizar un estudio clínico en vivo. Las diferentes prácticas para los análisis de cohortes darán lugar a los diferentes tipos de evidencia deseada (caracterización, estimación a nivel de población, predicción a nivel de paciente). Por tanto, el componente central para generar evidencia en OHDSI es el diseño de cohortes.

En OHDSI una cohorte es un “conjunto de personas que satisface uno o más criterios de inclusión durante un periodo de tiempo concreto” [3]. Definir correctamente la cohorte es fundamental a la hora de realizar cualquier estudio fenotípico en OHDSI y es crucial para realizar un buen análisis [41]. A continuación se presenta esquemáticamente la estructura fundamental de una cohorte, denominada en OHDSI “anatomía de una cohorte”.



Figura 5.7: “Anatomía de una cohorte”. Extraída del Tutorial 2022 publicado en la web oficial [2]

La investigación observacional comprende un intervalo temporal delimitado por el comienzo del período de observación (*Start of the observation period*, en verde) y el fin del período de observación (*End of the observation period*, en verde).

Dentro del período de observación, la cohorte se define con un evento de entrada a la cohorte (*Cohort Entry Event*, en azul) y un evento de salida de la cohorte (*Cohort Exit*, en azul).

- **Evento de entrada.** Define el evento que cualifica al paciente para entrar a la cohorte. El conjunto de pacientes que satisfacen el evento de entrada conforman la cohorte inicial.
- **Evento de salida.** Define el evento de salida de la cohorte, cuando el paciente ya no es elegible para formar parte de la cohorte.

Adicionalmente, la cohorte puede definirse más específicamente mediante una serie de **criterios de inclusión**. La cohorte que satisface todos los criterios de inclusión se denomina cohorte cualificada. La elección de los criterios de inclusión de la cohorte es fundamental en el diseño del estudio observacional.

La terminología que se emplea para describir los eventos clínicos que definen la cohorte se debe agrupar en grupos de conceptos. En OHDSI, los **grupos de conceptos son expresiones reusables que representan un listado de conceptos pertenecientes al Vocabulario que definen un evento clínico concreto**. Son el equivalente a las "listas de códigos" que se utilizan en los estudios observacionales [3].

5.4.2. Casos de uso para la investigación

Con el fin de estandarizar y proveer un marco metodológico en el camino hacia la generación de evidencia, OHDSI define tres casos de usos que establecen los diferentes tipos de estudio que se pueden realizar: (i) la caracterización, (ii) la estimación a nivel de población (iii) la predicción a nivel de paciente.

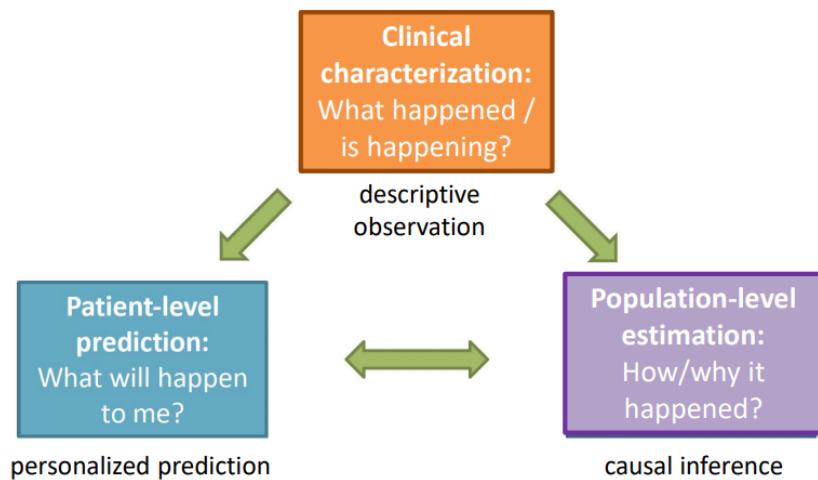


Figura 5.8: Esquema simplificado de los casos de uso para la investigación en OHDSI. Extraído del Symposium 2023, publicado en la web oficial [2]

Anteriormente se definieron las historias de paciente como marco fundamental de la investigación (véase 5.4.1 "Cohortes"). Cada caso de uso extrae un tipo de evidencia distinto a partir de la historia del paciente, tal y como se muestra a continuación en la Figura 5.9 "Esquema de los casos de uso encuadrado en la historia del paciente".

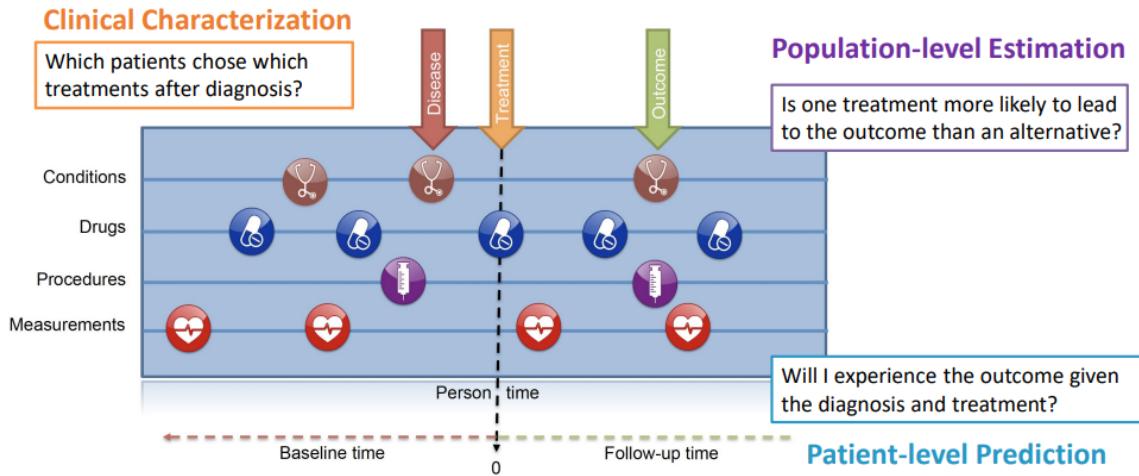


Figura 5.9: Esquema de los casos de uso encuadrado en la historia del paciente. Extraído del Symposium 2023 publicado en la web oficial [2]

La historia del paciente definirá la pertenencia o no del paciente a una cohorte y sobre esa cohorte se realizarán los distintos tipos de estudio. **El conjunto de estos tres casos de uso y la definición de cohortes conforma la metodología OHDSI para la generación de evidencia.** A continuación se describe brevemente cada uno de los casos de uso.

Caracterización

La caracterización busca la caracterización a nivel estadístico de una cohorte o una base de datos. Es una mera descripción estadística de los datos, sin realizar inferencias, predicciones o análisis más complejos, simplemente observando la base de datos.

Responde a la pregunta de investigación: **¿Qué les ha pasado?**

Obtiene como resultados: recuentos y porcentajes, medias, estadísticas descriptivas, ratios de incidencia...

Estimación a nivel de población

La estimación a nivel de población busca realizar inferencias causales sobre los efectos de las intervenciones sanitarias en la población. Se pretende entender los efectos causales para comprender las consecuencias de las acciones.

Responde a la pregunta de investigación: **¿Cuáles son los efectos causales?**

Obtiene como resultados: riesgos relativos, efectos causales, correlación entre variables, comparaciones de efectividad, asociaciones...

Predicción a nivel de paciente

La predicción a nivel de paciente busca, en base a los datos obtenidos de los conjuntos de pacientes en la base de datos, realizar predicciones concretas para individuos concreto.

Responde a la pregunta: **¿Qué me pasará a mí como paciente?**

Obtiene como resultados: probabilidades para un individuo, fenotipos probables, grupos de riesgo...

5.4.3. Vías de implementación del análisis

Para realizar un análisis, 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 5.10 "Tres vías para la implementación de un análisis observacional", extraída del capítulo 8 del Libro de OHDSI.

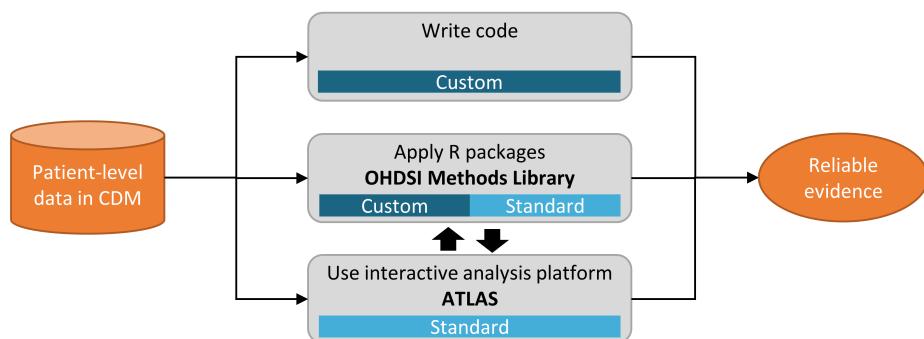


Figura 5.10: Tres vías para la implementación de un análisis observacional. Extraído del Libro de OHDSI [3]

Cada vía se evalúa en cuanto a lo personalizada (*custom*) o estandarizada (*standard*) que es. A estas alturas se debe conocer que la vía más recomendada para implementar el análisis será la más estandarizada, es decir, la tercera vía.

La problemática que presentan la primera y la segunda vía consiste en ser en mayor o menor medida vías customizadas, lo que genera problemas de interoperabilidad y reproducibilidad de los estudios. Si bien la primera vía consiste en la programación directa de código para realizar las consultas (no hay ningún tipo de estandarización, distintos lenguajes de programación, funciones personalizadas) al menos la segunda vía hace uso de librerías estándares en R que OHDSI ofrece (*OHDSI Methods Library*) pero, aunque se use el mismo lenguaje de programación y funciones, los scripts pueden ser tan distintos que aún dificultan la interoperabilidad.

Por tanto, la tercera vía se presenta como la alternativa óptima por ser la más estandarizada y es la que empleará el TFG en el estudio práctico. Esto es, usar la

herramienta interactiva *low-code* de análisis de datos que ofrece OHDSI, denominada **ATLAS**, sin necesidad de programar directamente código.

5.5. Conclusiones

En este capítulo se recogen las características fundamentales de OHDSI con el fin de comprender la relevancia de la organización en el panorama sanitario, como sucesora de OMOP y gran candidata para subsanar las dificultades en términos de interoperabilidad y estandarización de la investigación observacional.

Además se explora en mayor profundidad la metodología que promueve la organización en cuanto a la generación de evidencia clínica, a partir del concepto de cohorte y los estudios de caracterización, estimación a nivel de población y estimación a nivel de paciente.

6. Documento de Requisitos

Este capítulo se divide en cuatro secciones: [6.1 Introducción](#), [6.2 Requisitos funcionales](#), [6.3 Requisitos no funcionales](#) y [6.4 Conclusiones](#).

6.1. Introducción

Por la naturaleza informática del proyecto, bajo el convenio del departamento de Lenguajes y Sistemas Informáticos de la US, se presenta este capítulo donde se realiza un catálogo de requisitos del sistema.

En este caso, se ha realizado una adaptación de la ingeniería de requisitos debido a que no se está diseñando una herramienta o sistema de cero, sino que se está modelando un sistema ya existente, el ecosistema de ATLAS Broadsea. La arquitectura del sistema se presenta más detalladamente en el capítulo [8 "Arquitectura del Sistema"](#).

En este capítulo, en la sección [6.2 "Requisitos Funcionales"](#) se presenta el catálogo de requisitos funcionales del sistema.

En la sección [6.3 "Requisitos no Funcionales del Sistema"](#) se presenta el catálogo de requisitos no funcionales del sistema.

Por último, la sección [6.4 "Conclusiones"](#) recoge brevemente lo visto en el capítulo.

6.2. Requisitos funcionales

Los requisitos funcionales son declaraciones que especifican las acciones que un sistema debe realizar en respuesta a entradas específicas del usuario o del sistema.

Para el sistema de ATLAS Broadsea se han definido seis requisitos funcionales y dos actores o usuarios del sistema: el desarrollador y el analista de datos. Los requisitos funcionales hacen referencia a las tareas que puede realizar el usuario a la hora de conducir un análisis utilizando el sistema. A continuación se presentan: [6.2.1 "Diagrama de casos de uso"](#) y [6.2.2 "Casos de uso"](#).

6.2.1. Diagrama de casos de uso

El sistema distingue entre dos actores y las actividades que puede realizar cada uno de ellos. Mientras que desarrollador es el usuario encargado principalmente de gestionar el backend del sistema completo de Broadsea, el analista se encarga

más específicamente de realizar las tareas de análisis a través de la herramienta de ATLAS.

Debido a que el proyecto pone el foco mayoritariamente en el uso de la herramienta ATLAS, de los siete requisitos funcionales definidos, seis guardan relación con el analista y las tareas de análisis.

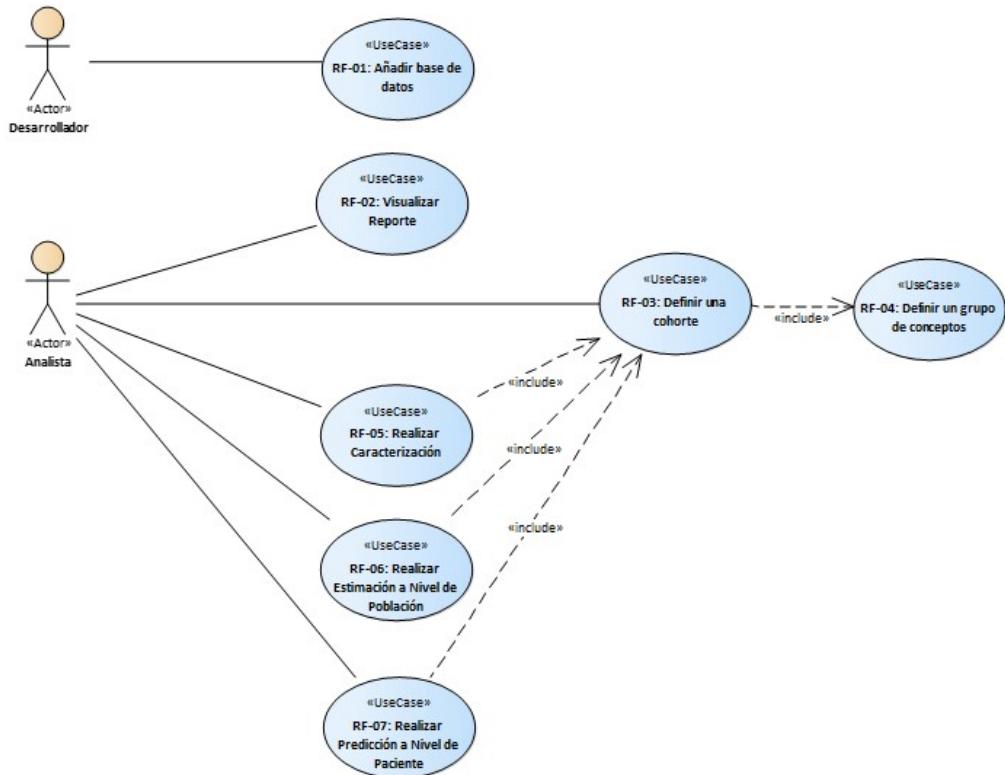


Figura 6.1: Diagrama de casos de uso

6.2.2. Casos de uso del sistema

A continuación se detalla un diagrama de actividad y una tabla descriptiva para caso de uso presentado en la anterior Figura 6.1 "Diagrama de casos de uso"

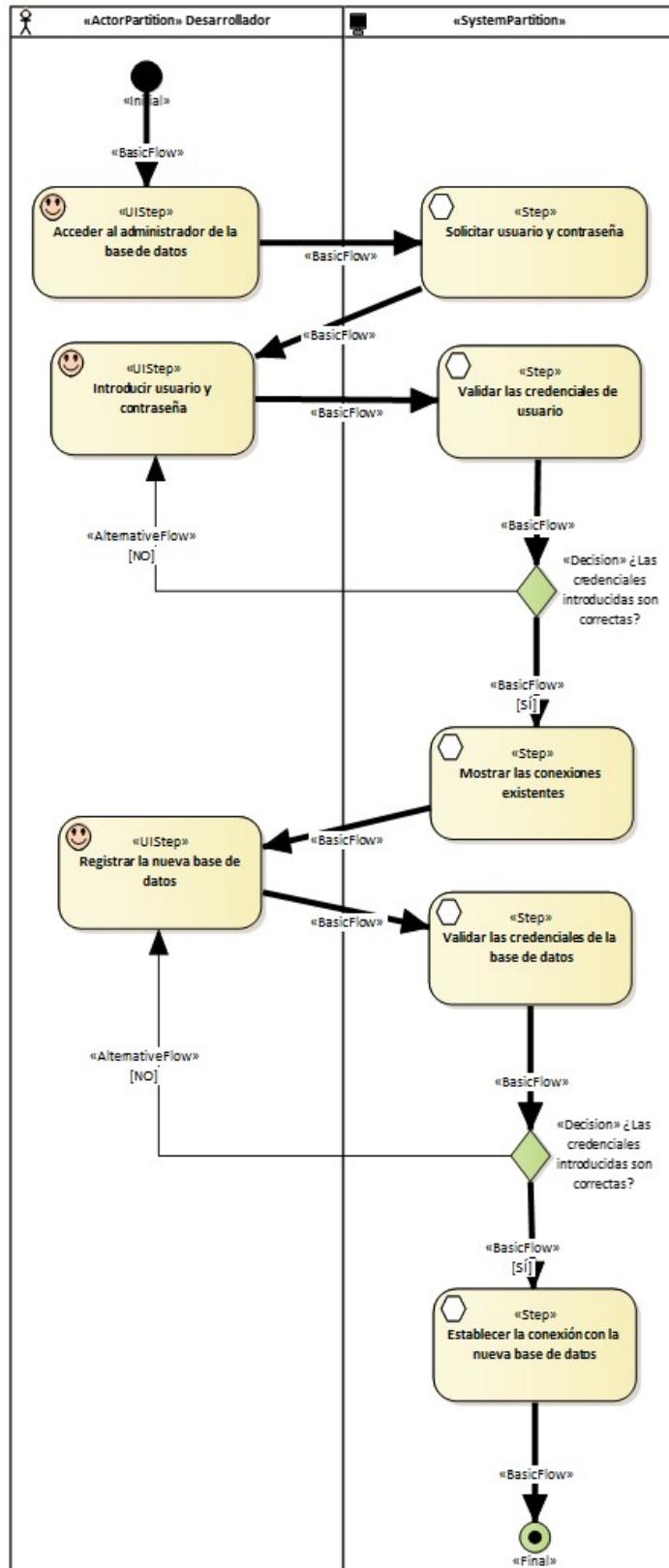


Figura 6.2: Diagrama de actividad de RF-01:Añadir base de datos

«UseCase» RF-01: Añadir base de datos					
Versión	1.0	08/04/2024 11:25			
Autor	MV Alonso de Caso O.				
Conexiones	Fuente	Estereotipo	Destino		
Desarrollador		<<Use>>	RF-01: Añadir base de datos		
Descripción		El desarrollador podrá añadir una base de datos a través de la configuración de la WebAPI del sistema			
Pre-condición y Post-condición	Pre-condición: La base de datos debe estar estandarizada al CDM de OMOP. Post-condición: La base de datos debe quedar registrada en el sistema.				
Estado	Implemented				

Tabla 6.1: Caso de uso de RF-01:Añadir base de datos

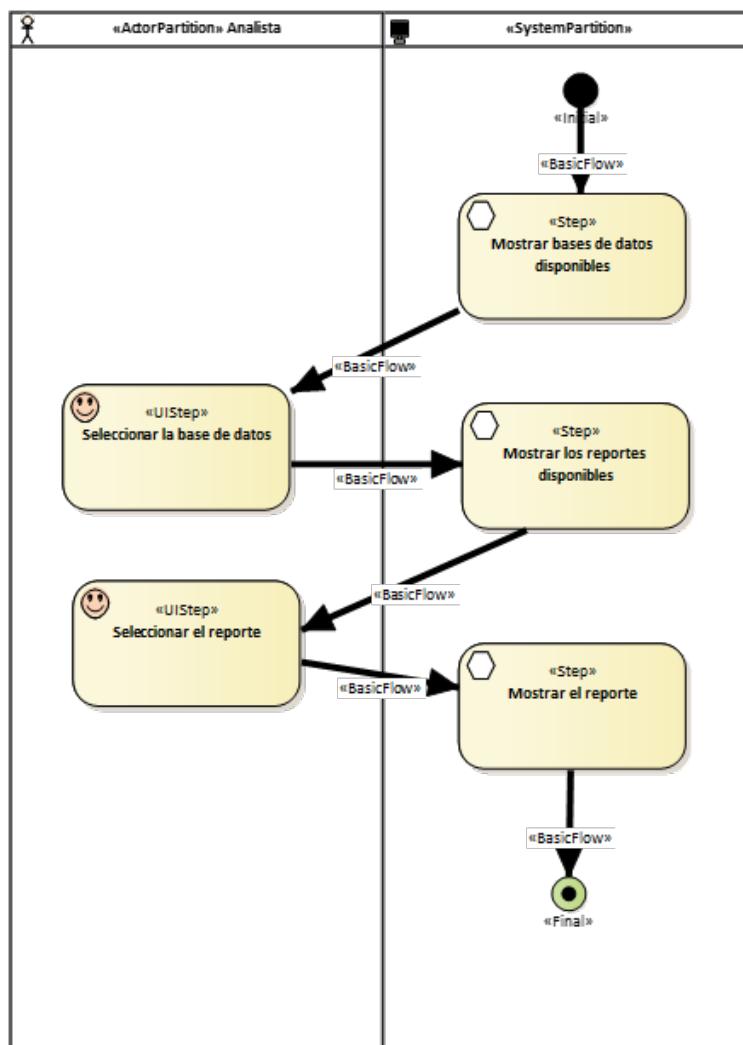


Figura 6.3: Diagrama de actividad de RF-02: Visualizar Reporte

CAPÍTULO 6. DOCUMENTO DE REQUISITOS

«UseCase» RF-01: Añadir base de datos			
Versión	1.0	08/04/2024 11:25	
Autor	MV Alonso de Caso O.		
	Fuente	Estereotipo	Destino
Conexiones	Desarrollador	<<Use>>	RF-01: Añadir base de datos
Descripción	El desarrollador podrá añadir una base de datos a través de la configuración de la WebAPI del sistema		
Pre-condición y Post-condición	Pre-condición: La base de datos debe estar estandarizada al CDM de OMOP. Post-condición: La base de datos debe quedar registrada en el sistema.		
Estado	Implemented		

Tabla 6.2: Caso de uso de RF-02:Visualizar Reporte

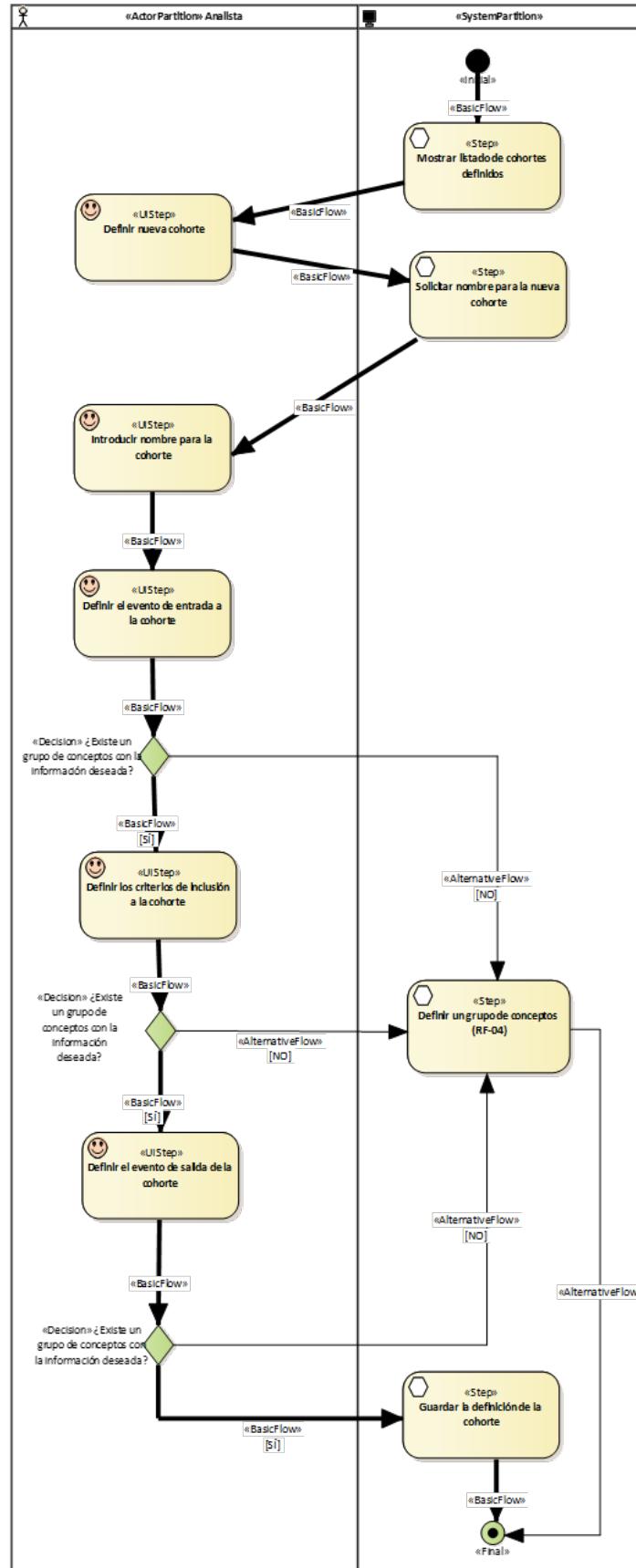


Figura 6.4: Diagrama de actividad de RF-03: Definir una cohorte

«UseCase» RF-03: Definir una cohorte			
Versión	1.0	08/04/2024 11:36	
Autor	MV Alonso de Caso O.		
Conexiones	Fuente	Estereotipo	Destino
	RF-03: Definir una cohorte	«include»	RF-04: Definir un grupo de conceptos
	Analista	«use»	RF-03: Definir una cohorte
	RF-07: Realizar Predicción a Nivel de Paciente	«include»	RF-03: Definir una cohorte
	RF-06: Realizar Estimación a Nivel de Población	«include»	RF-03: Definir una cohorte
Descripción	RF-05: Realizar Caracterización		
	Definir una cohorte o conjunto de personas que presentan unas características concretas sobre el que se va a realizar un estudio observacional durante un periodo concreto.		
Pre-condición y Post-condición	Pre-condición: Puede haber definido un grupo de conceptos que describa las características de la cohorte Post-condición: La cohorte debe quedar registrada en el sistema		
Estado	Implemented		

Tabla 6.3: Caso de uso de RF-03:Definir una cohorte

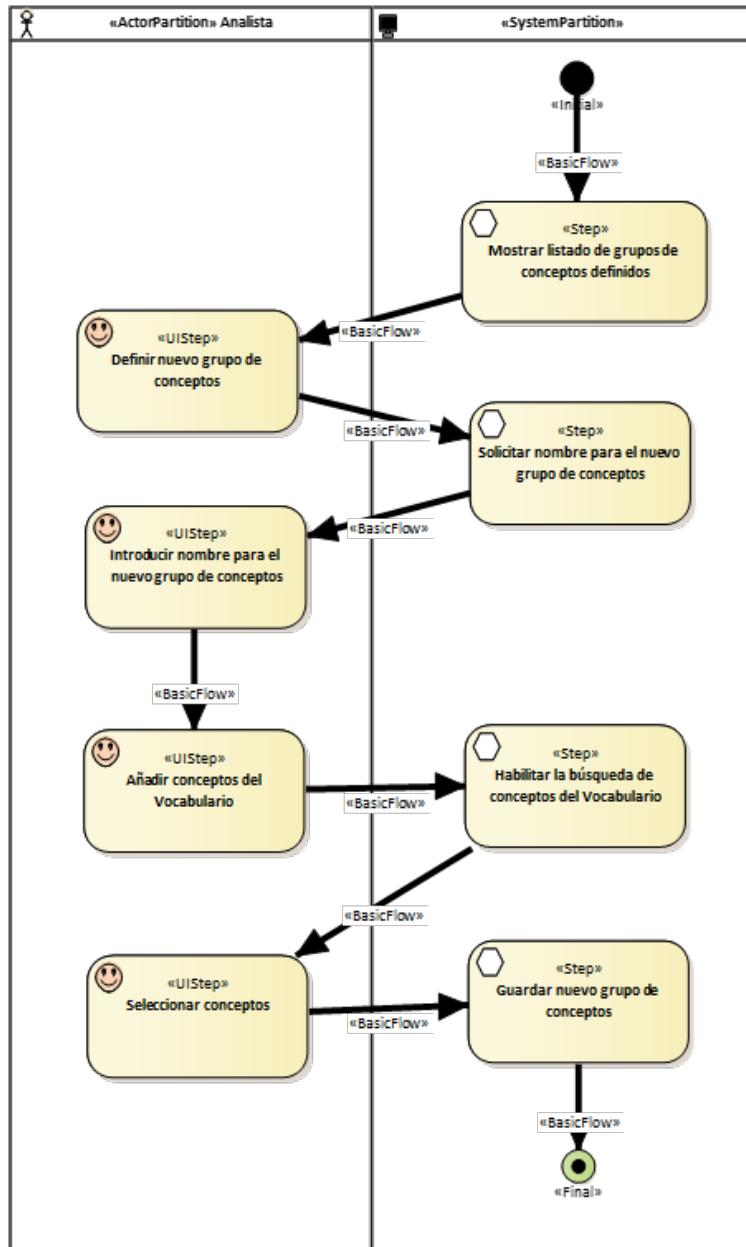


Figura 6.5: Diagrama de actividad de RF-04: Definir un grupo de conceptos

«UseCase» RF-04: Definir un grupo de conceptos			
Versión	1.0	08/04/2024 12:01	
Autor	MV Alonso de Caso O.		
Conexiones	Fuente	Estereotipo	Destino
	RF-03: Definir una cohorte	«include»	RF-04: Definir un grupo de conceptos
Descripción	Definir un grupo de conceptos del Vocabulario que reúna un conjunto de características fundamentales para el estudio.		
Pre-condición y Post-condición	Pre-condición: No procede Post-condición: El grupo de conceptos debe quedar registrado en el sistema		
Estado	Implemented		

Tabla 6.4: Caso de uso de RF-04:Definir un grupo de conceptos

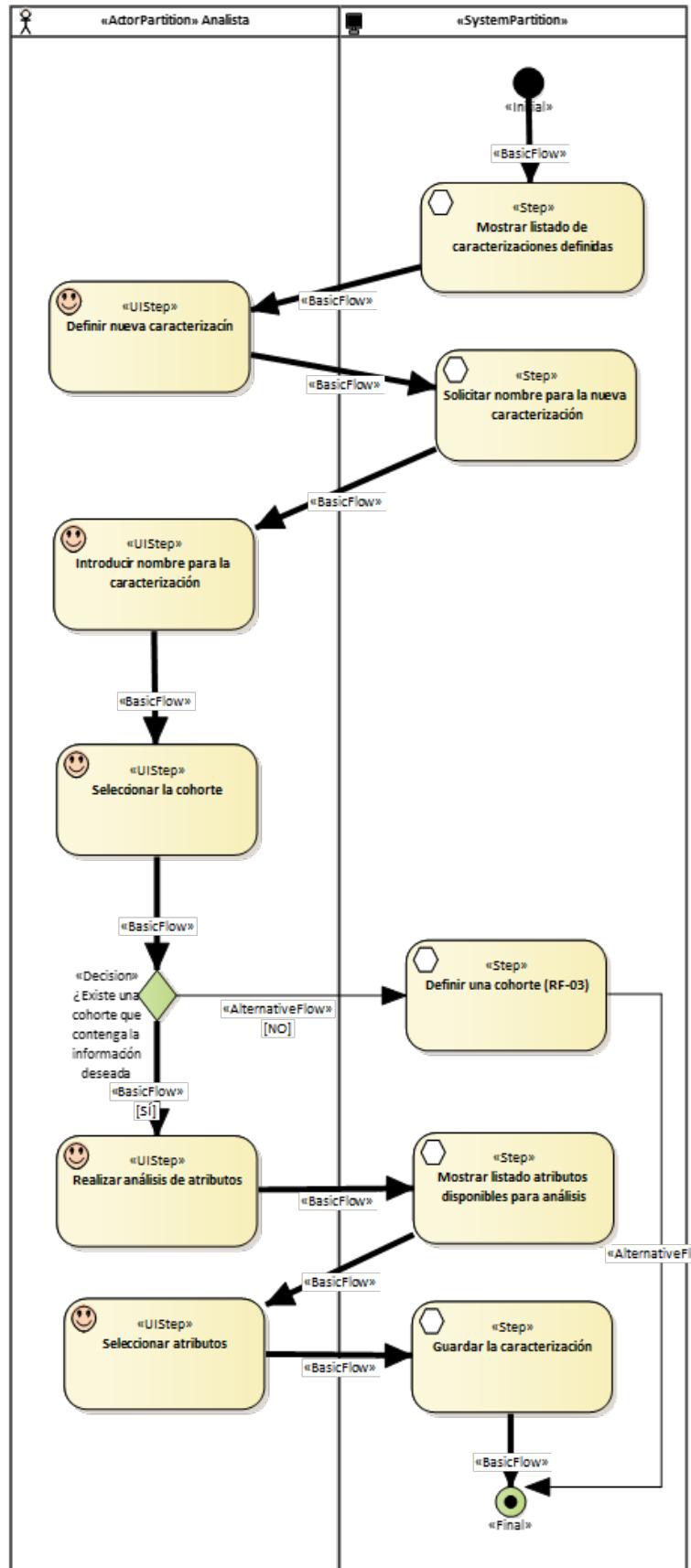


Figura 6.6: Diagrama de actividad de RF-05: Realizar Caracterización

CAPÍTULO 6. DOCUMENTO DE REQUISITOS

«UseCase» RF-05: Realizar Caracterización			
Versión	1.0	08/04/2024 11:37	
Autor	MV Alonso de Caso O.		
Conexiones	Fuente	Estereotipo	Destino
	RF-05: Realizar Caracterización	«include»	RF-03: Definir una cohorte
Descripción	Analista		
	«use»		
Pre-condición y Post-condición	Realizar un estudio de Caracterización de una cohorte para mostrar sus características estadísticamente más relevantes		
	Pre-condición: Puede haber una cohorte registrada que describa las características poblaciones del estudio Post-condición: La caracterización debe quedar registrada en el sistema		
Estado	Implemented		

Tabla 6.5: Caso de uso de RF-05: Realizar caracterización

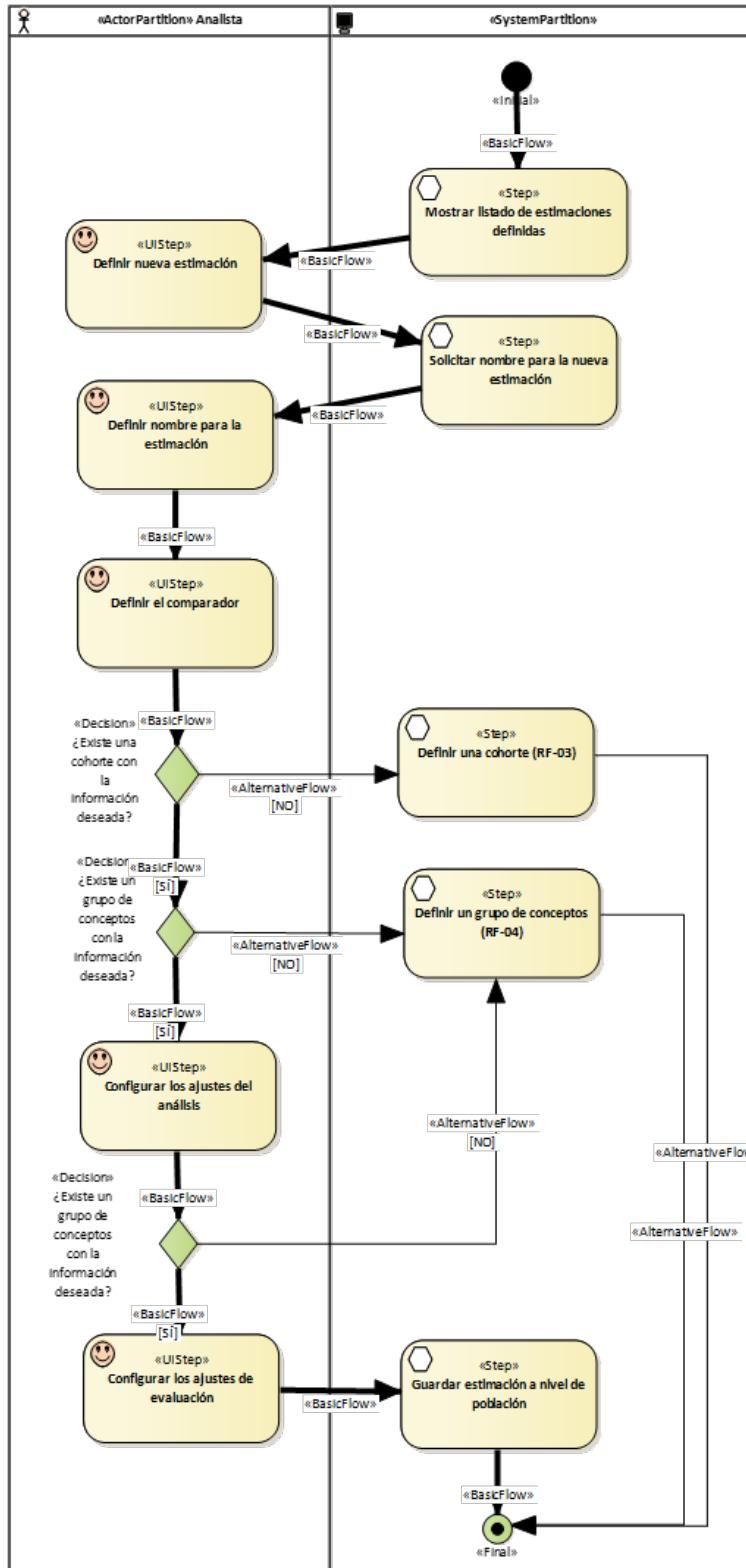
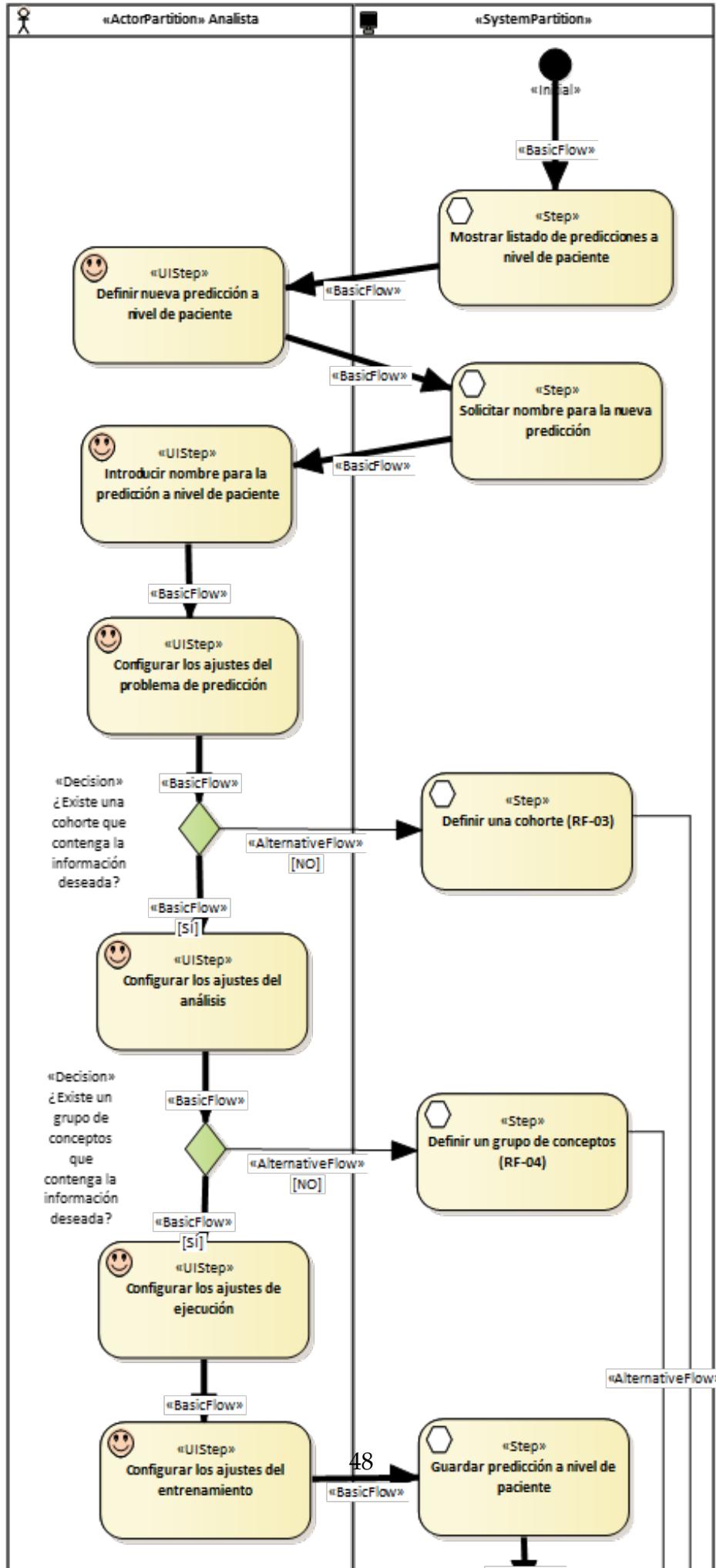


Figura 6.7: Diagrama de actividad de RF-06: Realizar caracterización

«UseCase» RF-06: Realizar Estimación a Nivel de Población			
Versión	1.0	08/04/2024 11:38	
Autor	MV Alonso de Caso O.		
Conexiones	Fuente	Estereotipo	Destino
	RF-06: Realizar Estimación a Nivel de Población	«include»	RF-03: Definir una cohorte
Descripción	Realizar un estudio de Estimación a Nivel de Población sobre unas cohortes previamente definidas para estimar efectos adversos que puede sufrir una población		
	Pre-condición: Puede haber una cohorte registrada que describa las características poblaciones del estudio Post-condición: La estimación a nivel de población debe quedar registrada en el sistema		
Estado	Implemented		

Tabla 6.6: Caso de uso de RF-06: Realizar Estimación a nivel de Población



«UseCase» RF-07: Realizar Predicción a Nivel de Paciente			
Versión	1.0	08/04/2024 11:39	
Autor	MV Alonso de Caso O.		
Conexiones	Fuente	Estereotipo	Destino
	RF-07: Realizar Predicción a Nivel de Paciente	«include»	RF-03: Definir una cohorte
Descripción	Analista		
	«use»		
Pre-condición y Post-condición	Pre-condición: Puede haber una cohorte registrada que describa las características poblaciones del estudio Post-condición: La predicción a nivel de paciente debe quedar registrada en el sistema		
Estado	Implemented		

Tabla 6.7: Caso de uso de RF-07: Realizar Predicción a nivel de Paciente

6.3. Requisitos no funcionales

Los requisitos no funcionales son restricciones o criterios de calidad que definen cómo debe comportarse un sistema, sin describir funciones específicas.

En base a lo aprendido sobre las características del sistema de ATLAS Broadsea, se han definido seis requisitos no funcionales.

A continuación se muestran estos requisitos de forma general en la Figura 6.9 "Diagrama de requisitos no funcionales" y posteriormente, se añade una tabla descriptiva para cada uno.



Figura 6.9: Diagrama de requisitos no funcionales

«NonfunctionalRequirement» RNF-01: Rendimiento		
Versión	1.0	09/05/2024 10:01:20
Autor	Maria del Valle Alonso de Caso Ortiz	
Descripción	El sistema debe funcionar eficientemente, proporcionando respuestas rápidas a las consultas y solicitudes de los usuarios, incluso cuando se trata con conjuntos de datos grandes o consultas complejas.	
Estado	Implemented	

Tabla 6.8: RNF-01: Rendimiento

«UserRequirement» RNF-02: Seguridad		
Versión	1.0	09/05/2024 10:03:11
Autor	Maria del Valle Alonso de Caso Ortiz	
Descripción	La herramienta debe ser respetuosa con los estándares de seguridad de la organización para proteger los datos sensibles de los pacientes.	
Estado	Implemented	

Tabla 6.9: RNF-02: Seguridad

«NonfunctionalRequirement» RNF-03: Usabilidad		
Versión	1.0	09/05/2024 10:04:38
Autor	Maria del Valle Alonso de Caso Ortiz	
Descripción	El sistema debe ser fácil de usar, con una interfaz intuitiva que permita a los usuarios navegar y realizar fácilmente.	
Estado	Implemented	

Tabla 6.10: RNF-03: Usabilidad

«NonfunctionalRequirement» RNF-04: Portabilidad		
Versión	1.0	09/05/2024 10:06:48
Autor	Maria del Valle Alonso de Caso Ortiz	
Descripción	El sistema debe ser capaz de ser implementado o transferido entre distintos entornos de programación, servidores y/o sistemas.	
Estado	Implemented	

Tabla 6.11: RNF-04: Portabilidad

«NonfunctionalRequirement» RNF-05: Interoperabilidad		
Versión	1.0	09/05/2024 10:08:44
Autor	Maria del Valle Alonso de Caso Ortiz	
Descripción	El sistema debe ser capaz de intercambiar información con otros sistemas, herramientas, lenguajes de programación y estándares o bases de datos.	
Estado	Implemented	

Tabla 6.12: RNF-05: Interoperabilidad

«NonfunctionalRequirement» RNF-06: Mantenimiento	
Versión	1.0
Autor	Maria del Valle Alonso de Caso Ortiz
Descripción	El sistema debe contar con servicios sólidos de soporte y mantenimiento, que incluyan actualizaciones oportunas, correcciones de errores, documentación y soporte al usuario para abordar las consultas y problemas de los usuarios de manera efectiva
Estado	Implemented

Tabla 6.13: RNF-06: Mantenimiento

6.4. Conclusiones

De este capítulo se concluye que, aunque el objetivo del proyecto no sea específicamente diseñar un sistema, el análisis de requisitos es de gran relevancia y utilidad para esquematizar y comprender las funcionalidades del sistema y sus propiedades.

Gracias a este análisis se abstrae de forma más sencilla el funcionamiento y las tareas que realiza el sistema de Broadsea, que en realidad es bastante más complejo.

7. Entorno de Trabajo

Este capítulo se divide en cinco secciones [7.1 Introducción](#), [7.2 Estándares de OHDSI](#), [7.3 Herramientas de OHDSI](#), [7.4 Programas informáticos empleados](#) y [7.5 Conclusiones](#).

7.1. Introducción

En este capítulo se presenta el entorno de trabajo utilizado durante el desarrollo del proyecto. Consiste principalmente en la utilización de los estándares y herramientas que provee OHDSI para conducir estudios observacionales.

No se puede entender la herramienta ATLAS sin entender el ecosistema de herramientas y estándares OHDSI que la acompañan.

Por tanto, en la sección [7.2 "Estándares de OHDSI"](#) se presentan los dos estándares fundamentales: el Modelo de Datos Común de OMOP y el Vocabulario.

En la sección [7.3 "Herramientas de OHDSI"](#) se presenta el conjunto de herramientas que ofrece la organización, prestando especial atención a la herramienta ATLAS.

Por último, en la sección [7.4 "Programas informáticos empleados"](#) se presentan los programas informáticos utilizados para desplegar el entorno de trabajo del proyecto.

7.2. Estándares de OHDSI

En términos de estandarización, OHDSI realiza una labor muy importante para paliar las dificultades de la investigación con datos de salud a causa de la heterogeneidad de los datos y estudios. Debido a la amplia colaboración internacional de la organización se reconoce la necesidad de estándares que permitan el intercambio de información sin pérdida entre los distintos sistemas de información de los miembros de la comunidad.

OHDSI ofrece dos estándares: el Modelo de Datos Común de OMOP y el Vocabulario. A continuación se describe cada uno de ellos en mayor detalle.

7.2.1. Modelo de Datos Común de OMOP

El Modelo de Datos Común o *Common Data Model (CDM)* 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” [4], en definitiva, es un modelo semántico estándar para estructurar los datos de salud. La información más relevante y actualizada sobre el CDM se encuentra en su página de github [4] y en el capítulo 4 del Libro de OHDSI [3].

Características

El modelo de datos de OMOP presenta características importantísimas para hacer frente a las necesidades del panorama socio-sanitario actual presentado en la sección 1.2 “Marco Contextual”. A continuación se presentan las características más relevantes del modelo (extraídas de la sección 4.1 del Libro de OHDSI [3]), según las necesidades identificadas previamente.

- **Estructura diseñada para la investigación.** El modelo presenta una estructura única y óptima para un propósito concreto: el de facilitar la realización de estudios observacionales. Por tanto reduce notoriamente los desafíos relativos a las diferentes estructuras y propósitos con los que se recogen los datos clínicos.
- **Modelo centrado en el paciente.** Es un modelo centrado en el paciente (alineado con la misma característica de la Sanidad 4.0). Estructuralmente esto significa que todos los eventos y tablas están relacionados con la tabla central del paciente, denominada *Person*.
- **Protección y privacidad.** El modelo limita el acceso a la información personal de los pacientes, evitando en la medida de lo posible el acceso a información personal sensible como nombres o apellidos, para fomentar la protección y privacidad de los datos. Mayor información sobre las técnicas empleadas para ello se encuentran en el apartado *Privacidad del paciente y OMOP* de la página de github [4].
- **Reutilización de estándares.** Un aspecto importantísimo es que el modelo propone su propio estándar pero sin olvidar los estándares globalmente utilizados, de manera que integra y reutiliza los conceptos provenientes de estándares ya existentes (ej. SNOMED, LOINC...) referenciándolos en su propio Modelo de Datos Común. El conjunto de todos los estándares conforma el Vocabulario.
- **Neutralidad tecnológica.** 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.

Modelo de Datos Lógico

Actualmente el CDM ha lanzado ya su 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 o 5.3 indistintivamente, que son las últimas versiones completamente funcionales.

A la hora de realizar un estudio en ATLAS o cualquier otra herramienta del ecosistema OHDSI la base de datos estará necesariamente estandarizada a este modelo por lo que es importante conocer su estructura fundamental. A continuación, en la Figura 7.1 "Estructura del CDM v5.4" se presenta la estructura lógica de este modelo y en la Figura 7.2 "Modelo Entidad-Relación del CDM v5.4", la estructura del modelo Entidad-Relación. Adicionalmente existe una página web que proporciona un modelo interactivo para facilitar su estudio [42].

Aunque el modelo de datos común de OMOP es muy complejo, incluso existen grupos de trabajo de la comunidad (*workshops*) especializados sólo en este ámbito, en esta subsección del trabajo tan solo se van a presentar los conceptos considerados estrictamente necesarios para la comprensión del contenido del mismo.

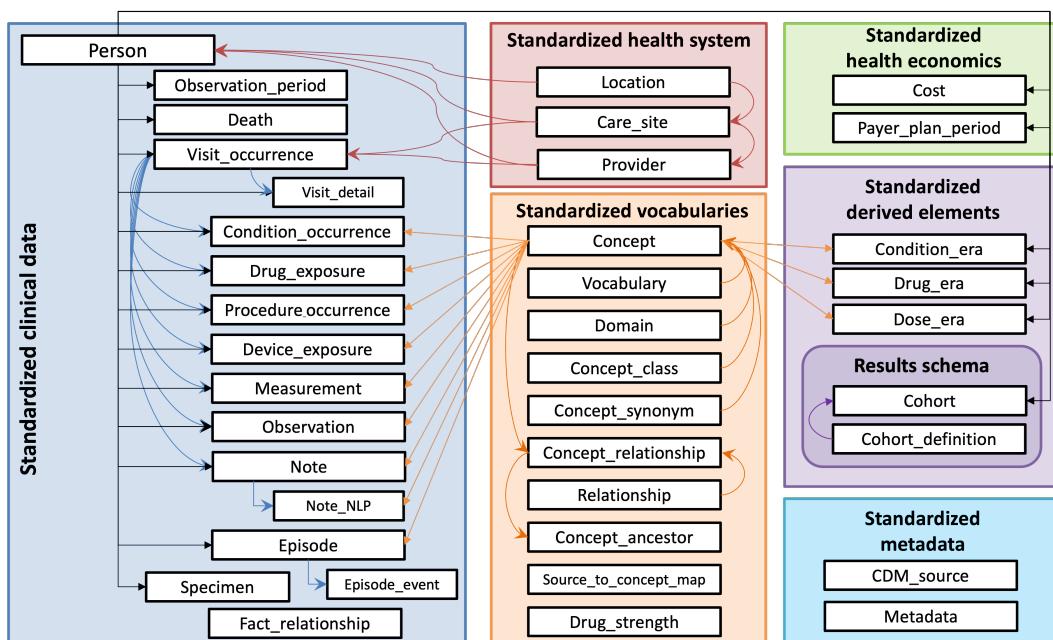


Figura 7.1: Estructura del CDM v5.4. Extraída de la página de github del CDM [4]

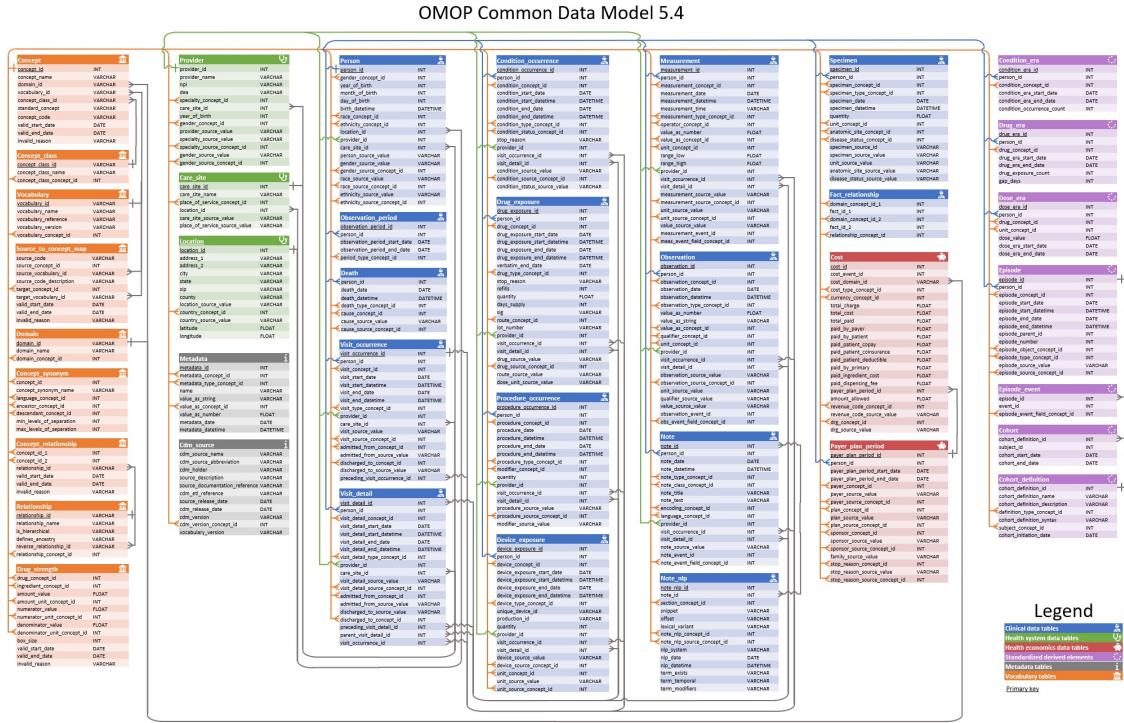


Figura 7.2: Modelo Entidad-Relación del CDM v5.4. Extraída de la página de github del CDM [4]

El modelo se comprende de 39 tablas agrupadas en seis grupos, de los cuales se destaca la importancia de tres: Datos clínicos estandarizados (*Standardized clinical data*, en azul), Vocabularios estandarizado (*Standardized vocabularies*, en naranja) y Elementos derivados estandarizados (*Standardized derived elements*, en morado). El grupo más importante es el de datos clínicos, que contiene la tabla Persona (*Person*), por la característica centrada en el paciente del modelo.

Cada evento clínico se registra en el modelo como un **Concepto** (*Concept*), perteneciente al grupo del Vocabulario estandarizado (en naranja). Además, cada concepto está ligado a un **Dominio** que especifica a qué tipo de información clínica corresponde dicho concepto. A continuación se muestra una tabla con los 30 dominios existentes y la cantidad de conceptos que tiene asociado cada uno.

Table 4.1: Number of standard concepts belonging to each domain.

Concept Count	Domain ID	Concept Count	Domain ID
1731378	Drug	183	Route
477597	Device	180	Currency
257000	Procedure	158	Payer
163807	Condition	123	Visit
145898	Observation	51	Cost
89645	Measurement	50	Race
33759	Spec Anatomic Site	13	Plan Stop Reason
17302	Meas Value	11	Plan
1799	Specimen	6	Episode
1215	Provider Specialty	6	Sponsor
1046	Unit	5	Meas Value Operator
944	Metadata	3	Spec Disease Status
538	Revenue Code	2	Gender
336	Type Concept	2	Ethnicity
194	Relationship	1	Observation Type

Tabla 7.1: Dominios del CDM v5.4. Extraída del Libro de OHDSI [3]

La información que contiene cada dominio se puede inferir fácilmente de la traducción al español del nombre por lo que no se va a hacer hincapié en ello. No obstante, se puede encontrar más información en [3], [4] o [42].

7.2.2. Vocabulario

El Vocabulario es otro de los elementos centrales del Modelo de Datos Común de OMOP y una gran herramienta de estandarización e interoperabilidad entre sistemas. Como se ha comentado en varias ocasiones, actualmente hay muchos estándares distintos en funcionamiento que establecen las terminologías de los eventos clínicos (por ejemplo LOINC, SNOMED CT, RxNorm...). El beneficio del Vocabulario de OMOP es que integra todos los vocabularios ya existentes en un único **Vocabulario estándar**, a través de la referenciación entre **conceptos estándar** (pertenecientes a OMOP) y conceptos no estándar (pertenecientes a vocabularios alternativos).

El Vocabulario de OHDSI, por tanto, impone sobre un conjunto de vocabularios, respetando las diversas procedencias de cada término pero mapeándolos a un único vocabulario estándar. **Cada concepto no estándar está asociado a un concepto estándar** y esta es la clave del Vocabulario.

Como todas las herramientas de la comunidad, la información acerca de este está disponible online de forma pública en el capítulo 5 del Libro de OHDSI [3] y en la

página de github del CDM [4]. Por otra parte, existe un buscador online de términos en el Vocabulario de OMOP denominado ATHENA [43].

Figura 7.3: 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 7.3 “Captura de pantalla del menú principal de ATHENA”, y 155 vocabularios distintos coexisten juntos en el estándar, de los cuales al menos 30 son vocabularios internos de OMOP.

7.3. Herramientas de OHDSI

OHDSI proporciona un conjunto de herramientas para facilitar la realización de los estudios e investigaciones a raíz de los datos clínicos y fomentar la interoperabilidad entre estos, aportando un estándar de herramientas.

Las herramientas que proporciona la organización están disponibles públicamente online y de forma gratuita y son desarrolladas por los propios miembros de la comunidad. Entre todas las herramientas, para la realización de este proyecto se destaca la herramienta de análisis de datos clínicos ATLAS, aunque también existen otras herramientas importantes de forma indirecta que se describen a continuación.

7.3.1. ATLAS

ATLAS es la herramienta de OHDSI por excelencia porque es la que estandariza el análisis observacional una vez que la base de datos está convertida al modelo OMOP. La documentación oficial sobre ATLAS se encuentra en el capítulo 8 del Libro de OHDSI y en su repositorio de github [5]. Además, aparte de la documentación oficial, hay montones de información esparcidas por la red sobre

ATLAS, en publicaciones científicas, foros de OHDSI, videotutoriales en youtube y un largo etcétera.



Figura 7.4: Logo de ATLAS. Extraída del repositorio de github [5]

Un importante promotor del uso de ATLAS es la red europea de datos EHDEN [?] (véase 1.3 “Estado del arte”). En esta línea, también la plataforma EHDEN Academy también ofrece cursos gratuitos sobre el uso de ATLAS y otras herramientas OHDSI.

Características y beneficios de su uso

El uso de ATLAS es beneficioso para la comunidad científica debido principalmente a su naturaleza *open-source*, *low-code* y la reproducibilidad que ofrece para los estudios:

- I. **Open source.** ATLAS se presenta como una herramienta disponible públicamente online, configurable gracias a su característica de código abierto, que expone toda su información y el propio código que la compone en los repositorios de github de la organización y, por si fuera poco, cuenta con el apoyo de un equipo de desarrolladores pendiente en los foros e *issues* que se reportan vía github para solucionar las dudas que tengan los implementadores.
- II. **Low-code.** Por otro lado, no requiere de conocimientos expertos de programación, puesto que es *low-code*. La herramienta se implementa sobre la Biblioteca de Métodos de OHDSI, con soporte para análisis en R, pero no requiere programación directa sino que ofrece una interfaz gráfica e intuitiva para el analista de datos. Además, el código que subyace al análisis es fácilmente exportable, siempre estructurado según el mismo estándar, favoreciendo la interoperabilidad del mismo.



Figura 7.5: Biblioteca de Métodos OHDSI. Extraída del Libro de OHDSI [3]

Todo ello no solo facilita la tarea del analista de datos sino que además favorece la interoperabilidad entre los estudios, puesto que todos los estudios que utilizan ATLAS implementan (en una capa inferior) los mismos métodos, el mismo lenguaje de programación y la misma estructura de análisis (véase [5.4 “¿Cómo generar evidencia?”](#)).

- III. **Reciclabilidad.** Por último, otro beneficio es que gracias a estas características ATLAS permite diseñar estructuras para el estudio de los datos que puedan utilizarse en diferentes bases de datos distintas. Volviendo al ejemplo de la plancha en [5.2 “¿Qué es OHDSI?”](#), esto quiere decir que una misma plancha (o estudio) puede conectarse a cualquier enchufe de cualquier región (a cualquier base de datos). ATLAS está intrínsecamente configurada para diseñar análisis reproducibles, por lo que los elementos que se configuran durante un análisis de datos (grupos de cohortes, estimadores, predictores, grupos de conceptos...) se pueden exportar fácilmente a modo de estructura general e implementarse sobre otro estudio que, aunque posea datos distintos execute ATLAS. Por tanto las estructuras más eficientes que se utilicen en un análisis remoto, pueden compartirse en la red de la comunidad y ser utilizados en cualquier nodo y cualquier estudio, favoreciendo la reciclabilidad, reproducibilidad e interoperabilidad del estudio.

Aspectos técnicos

En cuanto a los aspectos técnicos, ATLAS se despliega como una herramienta basada en web, normalmente alojada en un servidor Apache, combinada con la WebAPI de OHDSI. Generalmente se recomienda su despliegue en Google Chrome. Además la herramienta puede implementarse de forma pública a través de internet o tras el

firewall de la red privada de una organización, según las necesidades de la entidad que lo implementa.

Sin embargo, es importante recalcar que tanto ATLAS como la mayoría de las herramientas de OHDSI no consiste en un archivo ejecutable aislado sino en una aplicación contenida y dependiente de un ecosistema completo basado en web. La dependencia principal y red que sostiene a ATLAS es la **WebAPI**.

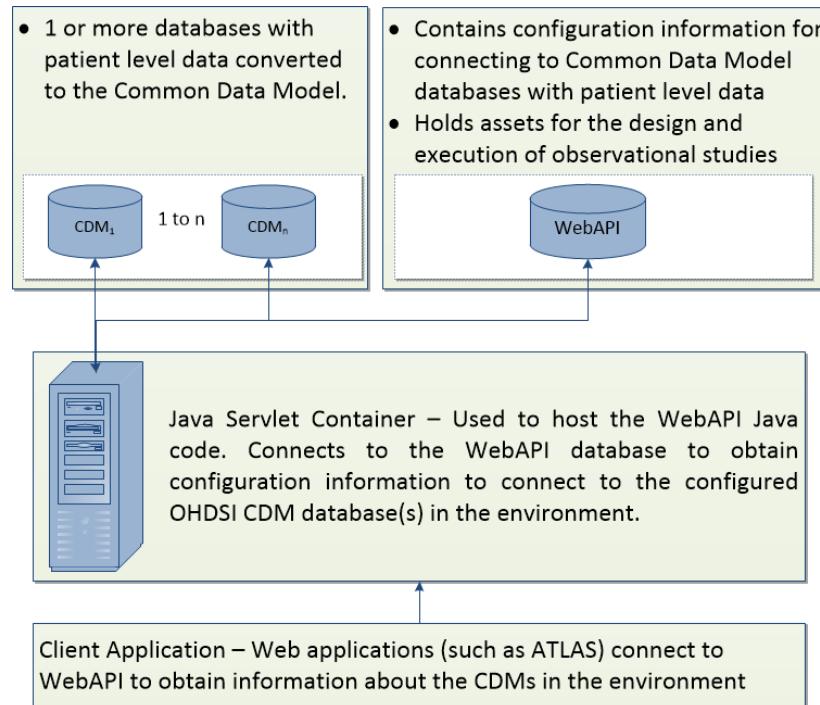


Figura 7.6: Estructura de la WebAPI. Extraída de la wiki de github [6]

Tal y como se muestra en la Figura 7.6 "Estructura de la WebAPI", la WebAPI es la aplicación que proporciona los servicios RESTful para que la herramienta pueda interactuar con las bases de datos [6]. Por tanto su relación con ATLAS es estrechamente necesaria. ATLAS no es una herramienta aislada sino un eslabón del ecosistema OHDSI.

Por otra parte, la herramienta en sí se muestra a través de una interfaz gráfica, que proporciona un estrecho menú lateral con 15 herramientas para el análisis de datos. La interfaz de la herramienta seleccionada se muestra en el lado derecho, como se muestra en la Figura 7.7 "Captura de pantalla del menú principal de ATLAS demo".

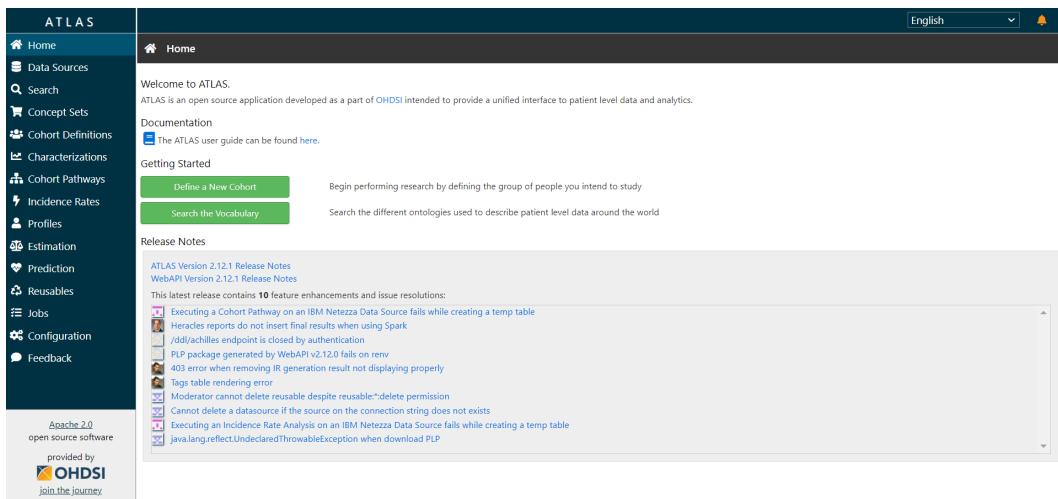


Figura 7.7: Captura de pantalla del menú principal de ATLAS demo

Recientemente, en diciembre de 2023, ATLAS lanzó su versión 2.14.1 que está en correcto funcionamiento y es la que se utiliza en el desarrollo del Trabajo Fin de Grado. Más información sobre los aspectos técnicos de la herramienta se encuentran en el repositorio de github [5].

Estrategias de Implementación

La implementación de ATLAS en una organización puede ser una tarea complicada por su dependencia con la WebAPI, la Biblioteca de Métodos y otras dependencias al ecosistema OHDSI.

No obstante, la organización ha desarrollado varias iniciativas que facilitan su implementación y accesibilidad, para no crear obstáculos en la promoción del uso de la herramienta. Estas iniciativas se describen a continuación.

- ATLAS demo** [44]. En primer lugar, esta es una herramienta muy fácilmente accesible que proporciona la comunidad científica para tomar un primer contacto con la herramienta. En este caso, la herramienta es accesible a través del navegador web, públicamente a través de Internet. Cualquier usuario de internet tiene acceso a la herramienta demo. Se le denomina demo porque se sobreentiende que su uso es principalmente educativo o formativo, aunque verdaderamente ofrece todas las capacidades de la herramienta y los análisis que con ella se realizan, podrían reutilizarse en estudios más complejos o de organizaciones privadas.
- ATLAS Docker**. Por otro lado, también muy fácilmente implementable se presenta **Broadsea** [45], que consiste en la virtualización del ecosistema OHDSI en un multicontenedor Docker. Gracias a la facilidad del uso de las tecnologías Docker, esta forma de implementar el ecosistema es bastante sencilla, permitiendo además añadir nuevas configuraciones más complejas (si fuese necesario) añadiendo o eliminando contenedores. Para realizar la parte práctica de este trabajo se emplea la tecnología Docker de Broadsea

para implementar ATLAS. A la herramienta ATLAS desplegada con Broadsea, frecuentemente se le denominará a lo largo del documento *ATLAS Broadsea*. El TFG presenta un documento anexo bastante complejo enteramente dedicado a la instalación, despliegue y configuración del entorno Broadsea (véase anexo [A](#) "Manual de instalación, despliegue y configuración de ATLAS Broadsea"). Adicionalmente, la arquitectura de Broadsea también se presenta en ?? "Arquitectura de Broadsea".

- c. **ATLAS Amazon Web Services.** Otra alternativa que propone la organziación, en colaboración con Amazon, es la virtualización del ecosistema en el entorno de computación en la nube de Amazon Web Services (AWS). Para ello se ofrecen los entornos *OHDSI-in-a-Box* [\[46\]](#) y *OHDSIonAWS* [\[47\]](#). OHDSI-in-a-Box se crea específicamente como un entorno de aprendizaje y se utiliza en la mayoría de los tutoriales proporcionados por la comunidad OHDSI mientras que OHDSIonAWS es una arquitectura de referencia para entornos OHDSI de clase empresarial, multiusuario, escalables. Por las restricciones intrínsecas al uso de AWS, estas alternativas han sido rechazadas para ser empleadas en el TFG.
- d. **ATLAS Azure.** Por último, *OHDSI on AZURE* [\[48\]](#) es otra alternativa de virtualización pero a través de la plataforma Microsoft de Azure. No obstante, esta alternativa es la menos común.

Herramientas embebidas

Si bien las herramientas del ecosistema de OHDSI no son totalmente aisladas, ATLAS presenta en su propia interfaz acceso a dos de estas herramientas de forma íntegra, para facilitar la eficiencia y rapidez en el análisis. Estas herramientas son las siguientes:

- **ACHILLES** [\[49\]](#). Esta herramienta, de las siglas *Automated Characterization of Health Information at Large-Scale Longitudinal Evidence Systems*, en español Caracterización automatizada de la información sanitaria en sistemas de evidencia longitudinal a gran escala, sirve para caracterizar y/o obtener un reporte estadístico de la base de datos estandarizada que se va a utilizar para el estudio. Intrínsecamente es una librería de R que se implementa como una opción del menú lateral *Data Sources* de ATLAS.
- **ATHENA** [\[50\]](#). Esta herramienta sirve para realizar búsquedas dinámicas en el Vocabulario de OMOP (véase [7.2.2](#) "El Vocabulario"). Está implementada en ATLAS en la opción *Search* del menú lateral. Además, se puede acceder a ella online de forma externa a través de su propia página web [\[43\]](#).

7.3.2. Otras herramientas

El ecosistema de OHDSI presenta gran cantidad de herramientas adicionales. A continuación se presentan otras herramientas que aunque no se utilizan directamente, son importantes para realizar un análisis de datos completo.

- **HADES** [51]. HADES, del inglés *Health Analytics Data-To-Evidence Suite* y en español Suite de análisis sanitario de datos a evidencia, es el nombre con el que se denomina a la herramienta que implementa el paquete R con la Biblioteca de Métodos de OHDSI (ver Figura 7.5 "Biblioteca de Métodos OHDSI"). Se puede instalar como un entorno independiente mediante Java y Rtools para implementar análisis mediante código estandarizado (véase Figura 5.10 "Tres vías para la implementación de un análisis observacional"). No se utiliza en el TFG más allá de la implementación subyacente de las bibliotecas en ATLAS.
- **Rabbit tools y Usagi** [52]. Estas herramientas en conjunto llevan a cabo el proceso de ETL, para omopizar las bases de datos al Modelo de Datos Común de OMOP. Las herramientas son tres: White-Rabbit, Rabbit-In-Hat y Usagi. No se utiliza directamente en el TFG porque el dataset utilizado para el análisis ya estaba previamente omopizado.
- **Data Quality Dashboard** [53]. Esta herramienta, en español Panel de control de calidad de los datos, pertenece a un paquete de HADES aunque implementado como una interfaz gráfica aparte para facilitar su acceso online. Tal y como su nombre indica sirve para automatizar la tarea de comprobación de la calidad de los datos, un paso previo fundamental antes de realizar un análisis de datos. Tampoco se utiliza directamente para el TFG porque este estudio se llevó a cabo durante la omopización del dataset.

7.4. Programas informáticos empleados

Los programas informáticos que han permitido el despliegue de este entorno tecnológico que envuelve al sistema son los siguientes: Google Chrome, Docker, PostgreSQL y Github.

Google Chrome

Google Chrome es el navegador web de Google que permite el acceso a internet y la búsqueda en la web a través de una interfaz amigable e intuitiva [54].

Chrome es el navegador recomendado por OHDSI para desplegar las herramientas de su ecosistema y más especialmente en el despliegue de Broadsea, permitiendo el acceso al servidor donde se aloja el sistema. Por tanto su uso ha sido muy relevante como portal de acceso a las herramientas OHDSI.

Docker

Docker es una plataforma abierta para desarrollar, enviar y ejecutar aplicaciones. Docker le permite separar sus aplicaciones de su infraestructura para que pueda entregar software rápidamente. Con Docker, puede administrar su infraestructura de la misma manera que administra sus aplicaciones [55].

De esta forma, Docker permite empaquetar y ejecutar aplicaciones en contenedores, entornos poco aislados pero seguros. Esto posibilita la ejecución de múltiples contenedores simultáneamente en un mismo host, sin depender de lo instalado en él. Los contenedores son ligeros y contienen todo lo necesario para la aplicación, facilitando su compartición y asegurando consistencia entre usuarios.

El uso de Docker en el desarrollo del proyecto es evidente, es la herramienta que despliega Broadsea y, por consiguiente, ATLAS. El proceso concreto de instalación, despliegue y configuración de Docker así como la explicación detallada de su estructura y archivos más importantes se presenta en el anexo A "Manual de instalación, despliegue y configuración de ATLAS Broadsea".

PostgreSQL

PostgreSQL es un potente sistema de base de datos relacional de objetos de código abierto que utiliza y amplía el lenguaje SQL combinado con muchas funciones que almacenan y escalan de forma segura las cargas de trabajo de datos más complicadas [56].

El uso de postgres es fundamental para la implementación correcta de Broadsea, puesto que su base de datos se implementa según PostgreSQL. Las bases de datos externas que se interactúan con la WebAPI pueden estar en otros lenguajes relacionales, pero el sistema de Broadsea intrínsecamente solo se sostiene sobre Postgre.

El proceso concreto de instalación, despliegue y configuración de la base de datos Postgre se realiza a través de la interfaz visual de pgAdmin 4.0 [57] y la explicación detallada de su estructura y archivos más importantes se presenta en el anexo A "Manual de instalación, despliegue y configuración de ATLAS Broadsea".

Github

GitHub es una plataforma para desarrolladores que les permite crear, almacenar, gestionar y compartir su código. Utiliza el software Git, proporcionando control de versiones distribuido, además de control de acceso, seguimiento de errores, solicitudes de funciones de software, gestión de tareas, integración continua y wikis para cada proyecto [58].

El uso de Github es muy recomendado debido a que la mayor parte de la información sobre OHDSI y sus herramientas se encuentran en internet disponibles en repositorios de Github (véase 5.2 "¿Qué es OHDSI?").

Además, siguiendo esta iniciativa de OHDSI, para desarrollar este Trabajo Fin de Grado se ha creado un repositorio de Github específico [26] que contiene toda la documentación relevante a su desarrollo (archivos latex, pdf...) y archivos de variables de entorno o scripts utilizados durante la configuración del entorno del sistema o la realización del análisis de datos.

7.5. Conclusiones

En esta sección se concluye que el entorno de trabajo del proyecto está enmarcado en el entorno de estándares y herramientas de la organización OHDSI, desplegados a través de una serie de programas informáticos.

Conocer el entorno de trabajo del proyecto y de OHDSI es gran relevancia puesto que no puede entenderse ATLAS sin conocer estos otros.

8. Arquitectura del Sistema

Este capítulo se divide en cuatro secciones: [8.1 Introducción](#), [8.2 Arquitectura teórica del sistema](#) y [8.3 Arquitectura de Broadsea](#) y [8.5 Conclusiones](#).

8.1. Introducción

La implementación del ecosistema de herramientas OHDSI y ATLAS puede ser una ardúa tarea. En el contexto de desarrollo del Trabajo Fin de Grado junto a las prácticas en empresa en el Hospital Virgen del Rocío, la dificultad de la tarea se ve exponencialmente aumentada debido a los grandes protocolos de seguridad y privacidad de la administración pública. Por ello, se ha seleccionado el despliegue de las herramientas OHDSI a través del sistema Docker de Broadsea, que presenta una vía sencilla para realizar esta labor.

Broadsea es un proyecto basado en Docker que permite desplegar todo el entorno de herramientas, configuraciones y dependencias OHDSI de la manera más sencilla hasta el momento. Por tanto, **el sistema se trata de una virtualización en Docker del entorno de herramientas OHDSI**.



Figura 8.1: Esquema sencillo de Broadsea. Extraída de [7].

En la sección [8.2 "Arquitectura teórica del sistema"](#) se presentan los aspectos teóricos fundamentales sobre virtualización y componentes de los sistemas docker y en la sección [8.3 "Arquitectura tecnológica de Broadsea"](#) se presenta la arquitectura específica de Broadsea.

No obstante, la arquitectura del sistema también se presenta en mayor profundidad técnica en el Anexo A "Manual de instalación, despliegue y configuración de ATLAS Broadsea".

8.2. Arquitectura teórica del sistema

El sistema se implementa mediante virtualización en Docker y una arquitectura en tres niveles o *three-tier*, donde se diferencian al cliente, frontend y backend. Esta arquitectura se describirá de forma general utilizando el esquema de la Figura 8.2.

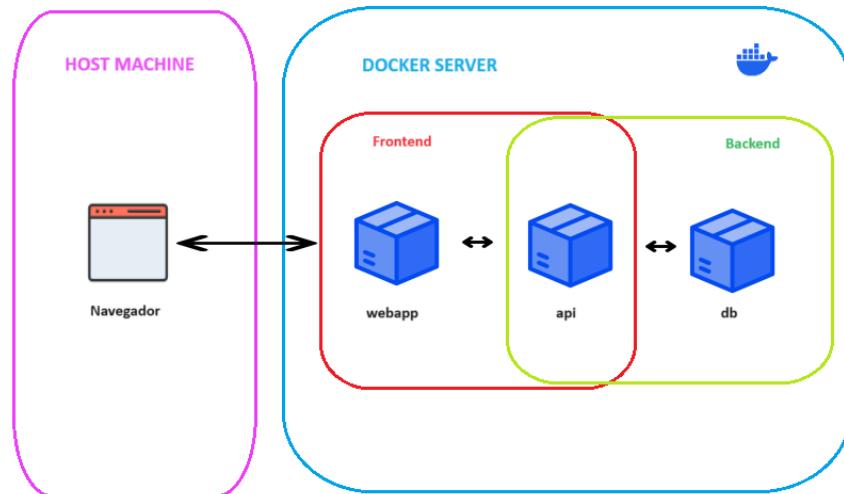


Figura 8.2: Esquema de arquitectura *three-tier* en Docker.

En primer lugar, la virtualización obliga a diferenciar entre una maquina local o anfitriona (*host machine*, en rosa) y una maquina virtual que provee el servicio docker (*docker service*, en azul).

1. **La máquina local.** La máquina local es la propia máquina del usuario. Se le denomina anfitriona porque aloja en su interior a la máquina virtual. La máquina local cede un servidor y un puerto a la máquina virtual para que el usuario final pueda acceder al sistema a través de la dirección del servidor en que se aloja, típicamente accediendo mediante un navegador web. El acceso mediante el navegador web es lo que se denomina la capa cliente, pues es la interfaz que permite al usuario acceder al sistema.
2. **La máquina virtual.** La máquina virtual es el sistema virtualizado en Docker. Es el sistema que contiene toda la lógica de la aplicación y los datos empaquetado en un multicontenedor Docker, en este caso el multicontenedor es el propio sistema Broadsea. Está compuesto por tres nodos la *webapp*, la *api* y la *db* que conforman las dos capas restantes de la arquitectura: el frontend y el backend.

Por tanto, a nivel de aquitectura del sistema en sí, se encuentra la capa cliente (en el *host machine*, en rosa), el frontend (*network-frontend*, en rojo) y el backend (*network-backend*, en verde).

1. **El cliente.** El cliente está alojado en la máquina anfitriona y proporciona el acceso a los servicios virtualizados del sistema a través de la conexión internet

con el servidor docker.

En el caso de Broadsea el navegador deberá ser Google Chrome y la dirección por defecto será <http://127.0.0.1:5432>.

2. **El frontend.** El frontend está alojado en la máquina virtual, es el servicio que guarda la lógica de la aplicación que se muestra al usuario. Se compone de la *webapp*, que contiene la aplicación como tal, y la *api*, que es la red que permite establecer interconexiones entre la aplicación lógica y la base de datos; entre el frontend y el backend.

En el caso de Broadsea la *webapp* y la *api* se combinan en el componente de la WebAPI, que permite el acceso a la aplicación de ATLAS y maneja las conexión con las bases de datos del backend.

3. **El backend.** El backend está alojado en la máquina virtual, es el servicio que aloja la base de datos sobre la que se sostiene la aplicación. Se compone de la *api* y la *db*. De igual forma que en el frontend, la *api* es la red que permite la interconexión entre los componentes del sistema, en este caso con la base de datos, que puede ser una o varias.

En el caso de Broadsea, las bases de datos deberán estar estandarizadas a OMOP y podrán encontrarse en el propio servidor Docker, como es el caso de Eunomia, o en servidores externos. No obstante, la relación entre cualquier base de datos y ATLAS se realiza a través de la WebAPI.

8.3. Arquitectura de Broadsea

Broadsea es un sistema muy complejo, contenido en un multicontenedor Docker que alberga el ecosistema completo de herramientas OHDSI y sus interconexiones en distintos contenedores. Además, se definen distintos perfiles (*profiles*) para facilitar la instalación de los distintos contenedores. Por ello se le denomina *a-la-carte*.

Broadsea es el *docker server* al que se refiere la anterior Figura 8.2 "Esquema de arquitectura three-tier en Docker". A continuación se muestran todos los contenedores que alberga el sistema de Broadsea.

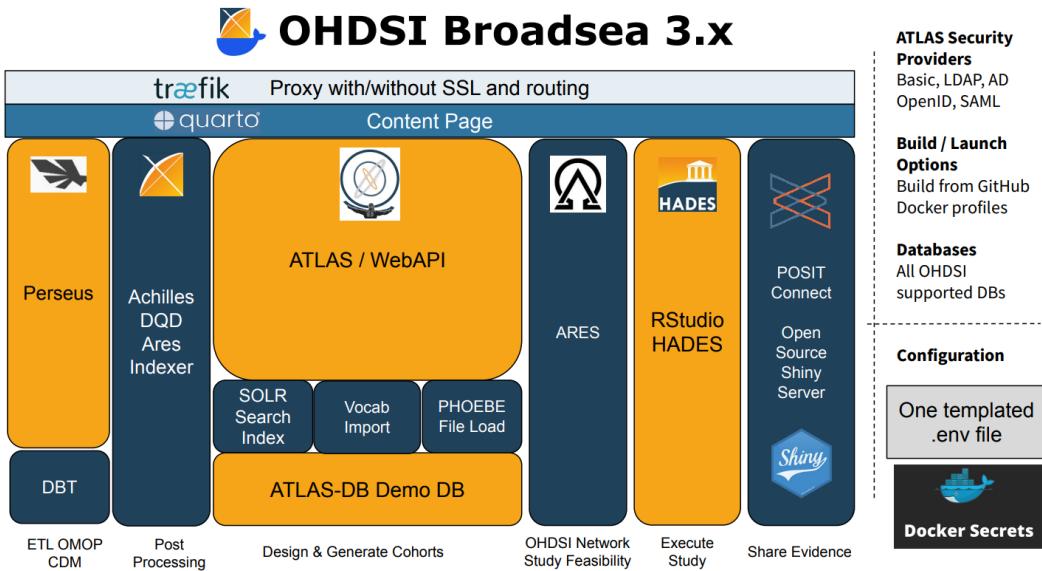


Figura 8.3: Vista general de todos los componentes de Broadsea. Extraída de [7].

El despliegue por defecto de Broadsea genera una interfaz de usuario con acceso a tres aplicaciones: ATLAS, HADES y ARES. Para acceder a esta interfaz de usuario basta con buscar en el navegador el servidor y puerto donde se aloja broadsea. Tipicamente el servidor corresponde al *localhost* y el puerto 5354, correspondiente a Postgre. La figura a continuación muestra la interfaz principal de herramientas disponibles al acceder a Broadsea desde Chrome.

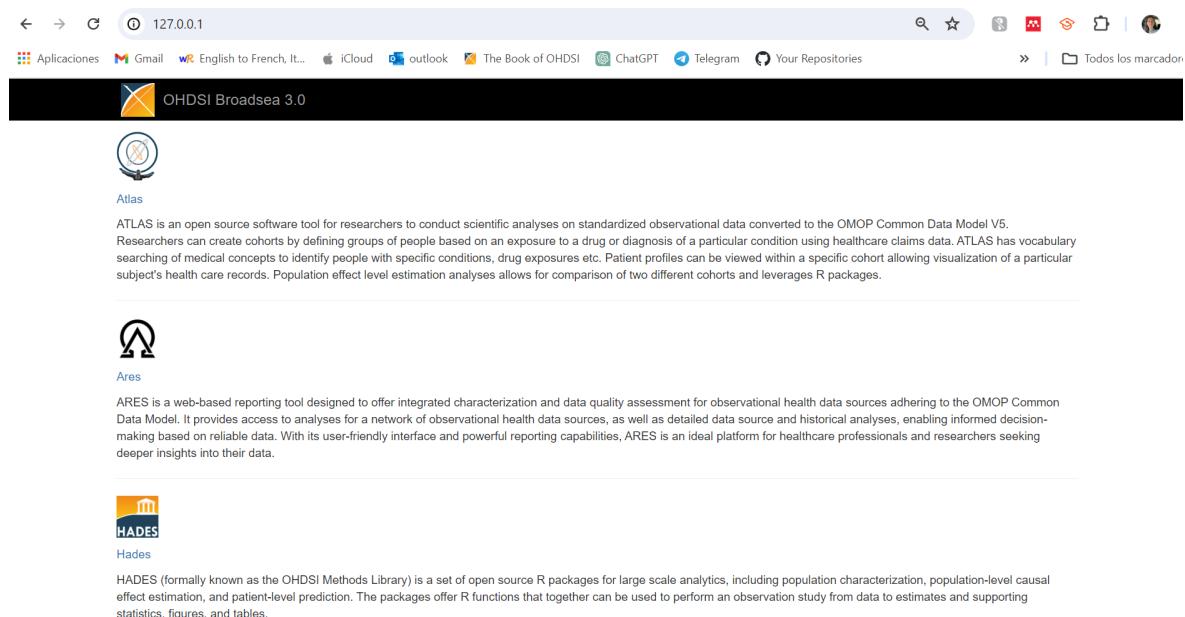


Figura 8.4: Captura de pantalla del menú principal de Broadsea

Por tanto, haciendo referencia a ambas figuras, se presenta a continuación una breve descripción de cada una de las herramientas accesibles desde el menú

principal de Broadsea. En el caso de ATLAS, por su relevancia, se describe la relación de contenedores de Broadsea que participan en su despliegue.

1. **ATLAS.** ATLAS Broadsea despliega todas las funcionalidades de la herramienta de forma local. ATLAS se sostiene sobre la WebAPI y cuenta con la base de datos de Eunomia.
 - **WebAPI.** La WebAPI se despliega como un contenedor docker y como un volumen de datos. Además, también se construirá un esquema en la base de datos del servidor Postgre que aloja al contenedor, denominado webapi. A través de la modificación de este esquema se podrán agregar o eliminar las diferentes fuentes de datos a la herramienta.
 - **BD.** Para facilitar el correcto funcionamiento de ATLAS se implementa una base de datos demo que es Eunomia. Esta base de datos cuenta con un pequeño registro de datos normalizados a OMOP y también crea varios esquemas en la base de datos del servidor Postgre que permiten su configuración, o la realización de consultas directamente desde el administrador de la base de datos.
2. **HADES.** HADES Broadsea despliega todas las funcionalidades de la herramienta de forma local. Se sostiene sobre una virtualización del IDE de RStudio que tiene preinstalada y preconfiguradas todas las librerías de la Librería de Métodos. Su uso no es relevante en el TFG.
3. **ARES.** ARES Broadsea despliega todas las funcionalidades de la herramienta de forma local. Su uso tampoco es relevante en el TFG.

8.4. Arquitectura de ATLAS Broadsea

ATLAS Broadsea hace referencia a la herramienta ATLAS desplegada a través de Broadsea. Como se ha mencionado previamente, ATLAS Broadsea es accesible a través del navegador Chrome, y se muestra de forma similar a ATLAS demo pero implementada localmente (recuerde Figura 7.7 “Captura de pantalla del menú principal de ATLAS demo”).

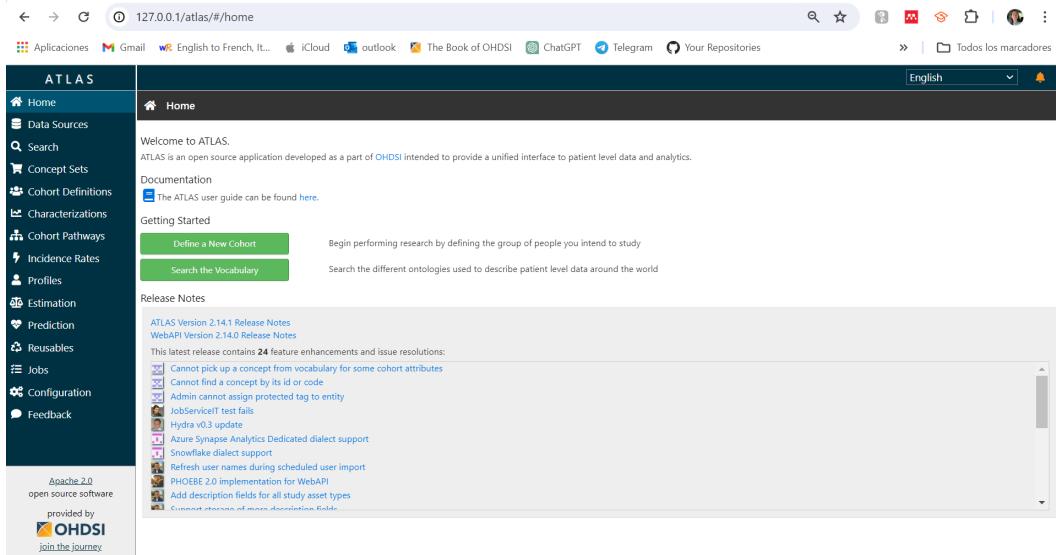


Figura 8.5: Captura de pantalla del menú principal de ATLAS Broadsea

El menú lateral de ATLAS presenta 15 herramientas de análisis, de las cuales en el proyecto se utilizan las siguientes:

- **Home.** Es el menú principal de ATLAS. Se muestra por defecto al abrir la herramienta.
- **Data Sources.** Es la herramienta para obtener reportes de las bases de datos integradas en la herramienta.
- **Search.** Es la herramienta para realizar búsquedas de conceptos en el Vocabulario.
- **Concept Sets.** Es la herramienta para definir grupos de conceptos que se utilizarán en la realización de análisis.
- **Cohort Definitions.** Es la herramienta para definir las cohortes que intervienen en los estudios y análisis.
- **Characterization.** Es la herramienta para realizar estudios estadísticos de caracterización de las cohortes definidas.
- **Estimation.** Es la herramienta para realizar estudios de estimación a nivel de población.
- **Prediction.** Es la herramienta para realizar estudios de predicción a nivel de paciente.

ATLAS Broadsea despliega por defecto la base de datos de Eunomia, que es una pequeña base de datos sintética estructurada al Modelo Común de Datos de OMOP que sirve de ayuda para la toma de contacto con la herramienta. La base de datos de Broadsea es accesible a través de un gestor de bases de datos PostgreSQL como pgAdmin. A continuación se muestra la estructura del servidor y base de datos de Broadsea.

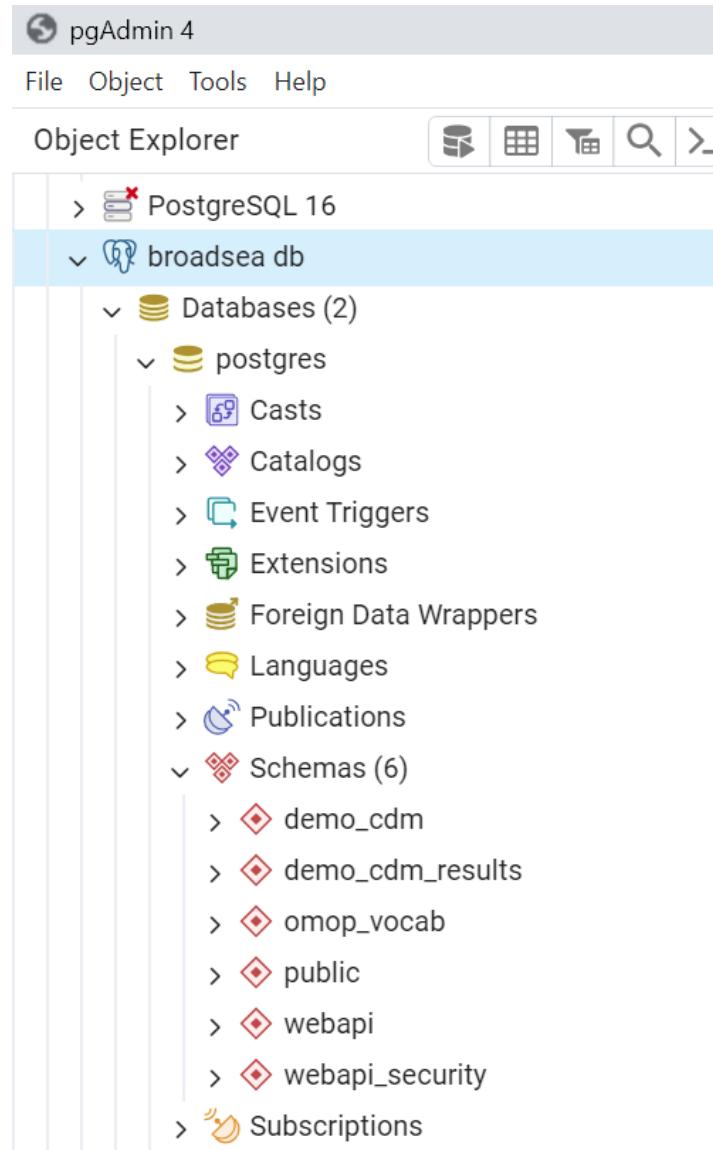


Figura 8.6: Captura de pantalla de pgAdmin de la estructura postgre del servidor de Broadsea

La base de datos presenta seis esquemas, siguiendo la configuración del Modelo de Datos Común de OMOP [59]. Para mayor información sobre la estructura postgre de Broadsea se recomienda consultar el anexo A “Manual de instalación, despliegue y configuración de ATLAS Broadsea”. No obstante, a continuación se describe brevemente la función de cada uno de estos esquemas:

- **demo_cdm:** Contiene toda la información de eventos clínicos y pacientes registrados en la base de datos. Es el grueso del contenido de la base de datos.
- **demo_cdm_results:** Contiene información generada de la ejecución de ACHILLES (véase 7.3 “Herramientas de OHDSI”) sobre la base de datos.
- **omop_vocab:** Este esquema no venía preinstalado pero es fundamental para el correcto funcionamiento de la herramienta. Contiene todo el vocabulario que va a ejecutar ATLAS. Su instalación se detalla en el manual.

- **public:** Este esquema no pertenece al CDM sino que se genera por defecto al crear una base de datos postgre. No contiene información relevante.
- **webapi:** Es el esquema de la WebAPI. Desde este esquema se establecen y gestionan las conexiones con bases de datos externas.
- **webapi_security:** Este esquema contiene ajustes para configurar la seguridad de la WebAPI. No se utiliza en el proyecto. Mayor información sobre la seguridad de la WebAPI en el manual.

8.5. Conclusiones

En este capítulo se concluye que la arquitectura tecnológica del sistema es compleja puesto que involucra una virtualización del ecosistema OHDSI a través de Docker, denominado Broadsea. No obstante, la implementación del sistema en Docker facilita bastante la tarea de configurar el ecosistema completo, gracias al empaquetamiento de las funcionalidades en contenedores *a-la-carte*.

9. Caso práctico

En este capítulo se divide en cinco secciones: [9.1 Introducción](#), [9.2 Estudio](#) realizado por el HUVR, [9.3 Estandarización del estudio con ATLAS](#), [9.4 Discusión](#) de resultados y [9.5 Conclusiones](#).

9.1. Introducción

Este capítulo pretende demostrar la relevancia de OHDSI (Observational Health Data Science and Informatics) y la utilidad de sus herramientas, concretamente el uso de ATLAS para la estandarización y reproducibilidad de los análisis clínicos observacionales sobre bases de datos estandarizadas al Modelo de Datos Común de OMOP.

Para ello, bajo la tutela de D. Carlos Parra y Da. Silvia Rodríguez (tutores de las prácticas en empresa, véase [3.1 "Participantes del proyecto"](#)), se ha seguido la reproducción de un estudio realizado por investigadores del hospital sobre predicción mediante modelos de ML de efectos adversos en el tratamiento radioterápico de pacientes con cáncer de pulmón.

Este estudio, se encuentra públicamente accesible en Pubmed en dos artículos, el primero publicado en el año 2019 titulado "*Comparison of Feature Selection Methods for Predicting RT-Induced Toxicity*" [60] y el segundo, en 2023 titulado "*Benchmarking machine learning approaches to predict radiation-induced toxicities in lung cancer patients*" [8]. Ambos estudios están también publicados en la ruta Thesis-ATLAS-OHDSI/documentation/pdf/estudioHUVR del repositorio de github del TFG [26].

El objetivo es promover el uso de ATLAS para la investigación observacional, reproduciendo mediante ATLAS un estudio que fue realizado sin hacer uso de la herramienta, para demostrar con un caso práctico los beneficios de utilizar la herramienta en términos de reproducibilidad y estandarización.

9.2. Estudio realizado por el HUVR

El estudio consiste en la comparación de 300 modelos de ML sobre un dataset de 875 pacientes de cancer de pulmón con el objetivo de predecir los efectos adversos a corto (esofagitis, tos, disnea y neumonitis) y a largo plazo (disnea y neumonitis) que producirá el tratamiento radioterápico sobre estos pacientes.

Contexto

La radioterapia, aunque beneficia el tratamiento oncológico, puede ocasionar efectos perjudiciales a corto y largo plazo, de forma personalizada según cada paciente [60, 8]. La medicina centrada en el paciente (véase 1.2 "Marco contextual") destaca la importancia de planificar individualmente cada tratamiento, dado que las respuestas varían entre individuos. Por tanto, la gestión personalizada de los efectos adversos es crucial en la planificación radioterápica para facilitar la toma de decisiones médico-paciente en términos de calidad de vida y supervivencia.

Objetivo

El objetivo del estudio es utilizar un conjunto de datos del mundo real (RWD) para facilitar la toma de decisiones clínicas, estudiando para cada efecto adverso del tratamiento radioterápico, el modelo de ML que provee una mejor predicción en términos de precisión del modelo (AUC).

Datos

El estudio utiliza datos del mundo real (RWHD) obtenidos de la combinación entre los datos almacenados en el registro S31 del HUVR y otros datos recogidos en consultas oncológicas rutinarias del hospital. La descripción de los datos del registro S31 se encuentra en el apéndice A del artículo "Benchmarking machine learning approaches to predict radiation-induced toxicities in lung cancer patients", también disponible en la ruta del repositorio de github Thesis-ATLAS-OHDSI/documentation/pdf/estudioHUVR.

En resumen los datos consisten en una recopilación de pacientes de cáncer de pulmón, con datos oncológicos descriptivos, datos de los tratamientos recibidos por el paciente y los efectos adversos sufridos.

Metodología

Para conformar los 300 modelos de ML se han entrenado y testeado 5 modelos de ML combinados con 10 métodos de selección de atributos (*Feature Selection, FS*) sobre 6 efectos adversos (*outcomes o clinical endpoints*), de la siguiente forma:

- **5 Modelos de ML.** Se utilizaron cinco clasificadores basados en aprendizaje automático:
 - Máquina de Vectores de Soporte (*Support Vector Machine, SVM*).
 - Vecinos más Cercanos (*k-Nearest Neighborhood, kNN*).
 - Red Neuronal Artificial (*Artificial, Neural Network, ANN*) de alimentación directa.
 - Modelo Lineal Generalizado (*Generalized Linear Model, GLM*).

- Clasificador de Naïve-Bayes (NB).

Los hiperparámetros de los modelos se optimizaron automáticamente siguiendo "las recomendaciones de la literatura".

- **10 Métodos de Selección de Atributos (FS).** Para reducir la dimensionalidad de los conjuntos de datos, se implementaron los siguientes métodos:

- Selección de Características Basada en Correlación (*Correlation-based Feature Selection, CFS*).
- Chi-cuadrado
- Boruta.
- Mínima Redundancia - Máxima Relevancia (*Minimum Redundancy-Maximum Relevance, mRMR*).
- Relief.
- Ganancia de Información (*Information Gain, IG*).
- Bosque Aleatorio (*Random Forest, RF*).
- 2 métodos de ensamblaje a partir de métodos de FS individuales y de subconjuntos.
- Subconjuntos de variables determinadas por un oncólogo experto para predecir las toxicidades seleccionadas basadas en la evidencia clínica.

- **6 Efectos adversos.** Se seleccionaron seis efectos adveros a estudiar, clasificados según si su duración fue a corto plazo y a largo plazo. A corto plazo:

- Esofagitis.
- Tos.
- Disnea.
- Neumonitis.

A largo plazo:

- Disnea.
- Neumonitis.

Se consideran efectos adversos crónicos o a largo plazo si los efectos se mantuvieron presente más de tres meses a partir del inicio del tratamiento.

Para la validación interna de los modelos se ha utilizado una estrategia de validación cruzada de 10 pliegues (*10-fold Cross-Validation*) en la que se aplicó una técnica de submuestreo aleatorio para generar un conjunto de datos equilibrado. Para la validación externa, se han utilizado los datos generados con los casos registrados después del 31 de mayo de 2018, que no fueron utilizados para la validación interna.

Por último, el rendimiento de los modelos se ha medido en términos del AUC logrado por cada modelo predictivo.

Resultados

Los resultados del estudio resaltan para cada outcome el mejor modelo de ML y selección de atributos, con la valoración de AUC en validación interna y externa. Los resultados se muestran de forma muy intuitiva en la siguiente tabla, extraída del artículo del HUVR [8].

Table 2
Best performing models in terms of AUC for each clinical endpoint and clinical variables considered by the models.

Clinical endpoint	Best model	AUC (internal validation)	AUC (external validation)	N Features	Clinical variables
Acute esophagitis	mRMR + GLM	0.85	0.81	69	Age, socioeconomic level, HIV, ethnicity, smoking status, primary symptom, anorexia, weight loss, KPS, height, tumor location, histology, EGFR, ALK, TNM, creatinine, hematocrit, familiar cancer history, QoL, concurrent CT, RT dose (lung, esophagus, heart, GTV, CTV).
Acute cough	IG + ANN	0.90	0.77	13	Socioeconomic level, QoL, RT dose (lung, esophagus, heart, GTV, CTV)
Acute dyspnea	mRMR + GLM	0.81	0.57	32	Socioeconomic level, COPD, oxygen therapy, primary symptom, anorexia, KPS, height, histology, ALK, TNM, Pulmonary function test, familiar cancer history, QoL, RT dose (lung, esophagus, GTV)
Acute pneumonitis	χ^2 + NB	0.81	0.85	24	Socioeconomic level, dyspnea, cough, histology, TNM, QoL, RT dose (CTV, GTV)
Chronic dyspnea	mRMR + GLM	0.87	0.97	19	Socioeconomic level, primary symptom, dysphagia, PET, TNM, ALK, familiar cancer history, QoL, GTV
Chronic pneumonitis	mRMR + ANN	0.90	0.73	32	Socioeconomic level, primary symptom, dyspnea, pleuritic pain, PET, tumor location, ALK, TNM, Pulmonary function test, familiar cancer history, QoL, RT dose (CTV, GTV, heart)

Tabla 9.1: Recopilación de resultados del estudio del HUVR. Extraída de [8]

9.3. Estandarización del estudio con ATLAS

Este capítulo presenta el caso práctico realizado por la alumna en el que se aplican todos los contenidos teóricos y herramientas presentadas a lo largo de la memoria para realizar un análisis de datos real.

El objetivo de este estudio se alinea con el objetivo de OHDSI: estandarizar la investigación clínica observacional, mediante el Modelo de Datos Común de OMOP y la herramienta de análisis de datos ATLAS. No se pretende meramente reproducir el estudio haciendo uso del ecosistema de OHDSI sino que se destaca que el fin último del proyecto es estandarizar el estudio, adaptarlo al marco de investigación OHDSI para que cualquier nodo de la organización pudiera procesarlo, analizarlo y reproducirlo fácilmente.

En el marco de OHDSI, el estudio realizado por el HUVR corresponde al caso de uso de Predicción a nivel de Paciente (recuerde 5.4.2 "Casos de uso para la investigación"). Tiene el objetivo de construir modelos que predigan la probabilidad de experimentar un efecto concreto en función de las características concretas de los pacientes.

No obstante, para poder aplicar en el análisis con ATLAS los otros dos casos de uso estudiados (Caracterización y Estimación a Nivel de Población), se ha adaptado el estudio bajo las consideraciones necesarias para que tuviera sentido. Por ello, la

finalidad del caso práctico no es una reproducción fiel del estudio sino la estandarización del estudio al marco de investigación metodológica de OHDSI.

9.3.1. Datos

En cuanto a los datos empleados, se ha utilizado una base de datos montada por el equipo de investigadores del HUVR en un servidor PostgreSQL

La estructura original de la base de datos proporcionada no correspondía con el Modelo de Datos Común de OMOP por lo que una tarea crucial de preprocesamiento ha sido el **OMOPizado de la base de datos y la comprobación de calidad de los datos**, realizado por mi compañero Francisco Rey Garduño como objeto de su Trabajo de Fin de Grado "Análisis de datos sanitarios mediante herramientas OHDSI y modelo de datos OMOP".

Por otra parte, es importante destacar que **la base de datos que se ha utilizado no es exactamente igual a la del estudio original** debido a motivos internos del equipo de investigación. La base de datos utilizada es una combinación y adaptación entre los datos del registro S31 y S32 del HUVR con una modificación importante: no incluye las variables relacionadas con las toxicidades experimentadas por los pacientes. Por tanto, las partes del estudio que involucran estas variables solo serán diseñadas pero no probadas en la herramienta.

9.3.2. Metodología

El estudio se ha realizado utilizando la herramienta ATLAS Broadsea. A estas alturas se conoce que el componente central de los estudios observacionales en OHDSI es el estudio de cohortes (recuerde [5.4.1 "Cohortes"](#)). Según el caso de uso que se diseñe sobre la cohorte se obtiene un tipo de evidencia u otro. En este caso, se va a realizar un estudio de cada caso de uso, utilizando en la medida necesaria las configuraciones recomendadas por defecto que ofrece la herramienta.

Análisis exploratorio

En primer lugar, es interesante realizar un breve análisis exploratorio de la base de datos. Para ello la herramienta Data Sources de ATLAS proporciona una interfaz muy intuitiva para generar reportes automáticamente según las características más relevantes de la base de datos.

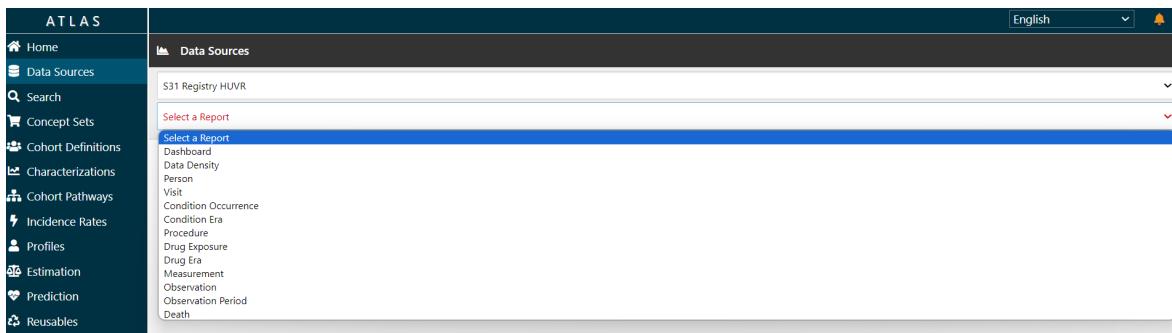


Figura 9.1: Seleccionador del reporte que se desea generar sobre sobre la base de datos S31/32 Registry HUVR

A continuación se adjuntan los gráficos más relevantes del conjunto total de reportes. Todos los gráficos del conjunto total de reportes se han descargado y subido al repositorio de github en la ruta Thesis-ATLAS-OHDSI/atlas/reports.



Figura 9.2: Reporte general de la información de la base de datos S31/32 Registry HUVR

La base de datos contiene 1332 registros de pacientes, es decir, un número mayor que la base de datos utilizadas en el estudio del HUVR. Esto se debe a que las bases de datos no son exactamente las mismas y la base de datos utilizada en el proyecto es una combinación del registro S31 y S32 procedente de la fuente original denominada omop_oncologia. Por otra parte, el 84.1 % de la población registrada es masculina y el 15.9 % es femenina.

CAPÍTULO 9. CASO PRÁCTICO

Concept Id	Name	Person Count	Prevalence	Records per person
4163446	Distant metastasis present	571	42.87%	1.71
255573	Chronic obstructive lung disease	564	42.34%	1.00
4110705	Squamous cell carcinoma of lung	535	40.16%	1.00
4198434	Local recurrence of malignant tumor of lung	451	33.86%	1.00
4112738	Adenocarcinoma of lung	427	32.06%	1.00
46284863	Malignant melanoma stage IIIB	401	30.10%	1.00
46287026	Malignant melanoma stage IIIA	373	28.00%	1.00
201820	Diabetes mellitus	358	26.88%	1.00
321588	Heart disease	337	25.30%	1.00
4140471	Epidermal growth factor receptor negative non-small cell lung cancer	296	22.22%	1.00
4110591	Small cell carcinoma of lung	250	18.77%	1.00
46284865	Malignant melanoma stage IV M1a	199	14.94%	1.00
46284859	Malignant melanoma stage IB	149	11.19%	1.00
46284860	Malignant melanoma stage IIA	62	4.66%	1.00
46284858	Malignant melanoma stage IA	60	4.50%	1.00
4143825	Epidermal growth factor receptor positive non-small cell lung cancer	49	3.68%	1.00
4264626	Grade not determined	39	2.93%	1.00
4115276	Non-small cell lung cancer	35	2.63%	1.00
46284861	Malignant melanoma stage IIB	32	2.40%	1.00
443388	Malignant tumor of lung	13	0.98%	1.00
46284864	Malignant melanoma stage IIIC	10	0.75%	1.00
37165673	Undifferentiated carcinoma of lung	8	0.60%	1.00

Tabla 9.2: Reporte de las condiciones registradas en la base de datos S31/32 Registry HUVR

Las condiciones registradas en la base de datos son mayoritariamente conceptos distintos que describen distintas características del cáncer de pulmón. Estos conceptos posteriormente se agruparán en un único grupo de concepto para facilitar el estudio. También aparecen las condiciones de diabetes mellitus y enfermedad cardíaca, aunque estas no son muy relevantes para el estudio.

Es importante recordar que no aparece en el reporte ningún tipo de condición relacionada con alguno de los efectos adversos identificados en el estudio original (disnea aguda, disnea crónica, neumonitis aguda, neumonitis crónica, esofagitis o tos) debido a que no están incluidos en la base de datos facilitada por el HUVR.

Concept Id	Name	Person Count	Prevalence	Records per person
1617298	Common news sources	1,308	98.20%	1.00
3630374	Karnofsky Performance Status [interpretation]	1,304	97.90%	1.00
4056418	Radiological tumor control	1,300	97.60%	1.08
44804077	Survival rate	1,297	97.37%	2.25
4177258	Intention values	1,283	96.32%	1.00
4056678	Radiotherapy started	835	62.69%	1.00
4057771	Radiotherapy completed	834	62.61%	1.00
37165785	Radiotherapy treatment plan	817	61.34%	1.00
4303574	Area restriction	817	61.34%	1.00
44793207	Preparation for simple radiotherapy with imaging and dosimetry	815	61.19%	1.00
4036402	Planning target volume	766	57.51%	2.00
4041434	Finding by auscultation	634	47.60%	1.00
4310250	Ex-smoker	622	46.70%	1.00
4298794	Smoker	611	45.87%	1.00
4058776	Radiotherapy stopped	527	39.56%	1.68
4030314	Neoplasm	278	20.87%	1.00
4056679	Radiotherapy changed	22	1.65%	1.00

Tabla 9.3: Reporte de las observaciones registradas en la base de datos S31/32 Registry HUVR

Las observaciones registradas en la base de datos hacen referencia a características observables en el paciente. Las observaciones más destacables son aquellas relacionadas con la planificación del tratamiento radioterápico.

Concept Id	Name	Person Count	Prevalence	Records per person
4311405	Systematic Nomenclature of Medicine - Clinical Terms (IHTSDO) 86273004: Biopsy	944	70.87%	1.00
37156151	Systematic Nomenclature of Medicine - Clinical Terms (IHTSDO) 1162782007: Three dimensional external beam radiation therapy	566	42.49%	1.00
42872834	Systematic Nomenclature of Medicine - Clinical Terms (IHTSDO) 450827009: Induction chemotherapy	492	36.94%	1.00
45766299	Systematic Nomenclature of Medicine - Clinical Terms (IHTSDO) 703423002: Combined chemotherapy and radiation therapy	264	19.82%	1.00
603135	Systematic Nomenclature of Medicine - Clinical Terms (IHTSDO) 1156530009: Volumetric modulated arc therapy	212	15.92%	1.00
4216177	Systematic Nomenclature of Medicine - Clinical Terms (IHTSDO) 394895009: Postoperative chemotherapy	136	10.21%	1.00
4070879	Systematic Nomenclature of Medicine - Clinical Terms (IHTSDO) 173171007: Lobectomy of lung	127	9.53%	1.00
40480519	Systematic Nomenclature of Medicine - Clinical Terms (IHTSDO) 441799006: Intensity modulated radiation therapy	34	2.55%	1.00
4119250	Systematic Nomenclature of Medicine - Clinical Terms (IHTSDO) 287310004: Lung tumor excision	19	1.43%	1.00
4172438	Systematic Nomenclature of Medicine - Clinical Terms (IHTSDO) 49795001: Total pneumonectomy	14	1.05%	1.00
44790293	Systematic Nomenclature of Medicine - Clinical Terms (IHTSDO) 231711000000108: Radiotherapy delivery	12	0.90%	1.00
603132	Systematic Nomenclature of Medicine - Clinical Terms (IHTSDO) 1156526006: Two dimensional external beam radiation therapy	8	0.60%	1.00

Tabla 9.4: Reporte de los procedimientos registrados en la base de datos S31/32 Registry HUVR

Los procedimientos registrados en la base de datos presentan una característica común y es que todos están mapeados a SNOMED CT, el estándar por excelencia para definir este tipo de información. Los procedimientos se pueden agrupar en tres conjuntos: radioterapia, quimioterapia y cirugía torácica. Estos tres conjuntos formarán a continuación otros tres grupos de conceptos.

Grupos de Conceptos

Una vez se conocen los conceptos más relevantes en el estudio, estos se puede asociar en grupos según los eventos clínicos a los que hacen referencia, tal y como se sugría en el apartado anterior. Esta tarea se realiza a través del menú Concept Sets de ATLAS.

Se han identificado y definido 12 grupos de conceptos relevantes en el estudio. A continuación se muestra el listado de los grupos de conceptos empleados en el estudio.

Id	Name	Created	Updated	Author
10	neumonitis aguda	04/29/2024 2:30 PM	05/17/2024 6:55 PM	anonymous
55	neumonitis crónica	05/17/2024 6:55 PM	05/17/2024 6:55 PM	anonymous
54	disnea crónica	05/17/2024 11:20 AM	05/17/2024 11:24 AM	anonymous
53	disnea aguda	05/17/2024 11:19 AM	05/17/2024 11:24 AM	anonymous
51	radioterapia	05/15/2024 2:25 PM	05/17/2024 11:02 AM	anonymous
45	radioterapia y quimioterapia	04/30/2024 10:21 AM	05/17/2024 11:02 AM	anonymous
52	cirugía torácica	05/15/2024 2:28 PM	05/15/2024 2:29 PM	anonymous
49	quimioterapia	05/15/2024 2:23 PM	05/15/2024 2:25 PM	anonymous
47	cáncer de pulmón	05/15/2024 1:28 PM	05/15/2024 2:09 PM	anonymous
7	tos	04/29/2024 2:23 PM	05/06/2024 9:50 AM	anonymous
9	esofagitis	04/29/2024 2:28 PM	05/06/2024 9:49 AM	anonymous
4	diabetes mellitus	04/29/2024 2:17 PM	04/29/2024 2:20 PM	anonymous

Tabla 9.5: Listado de los 12 grupos de conceptos definidos en ATLAS Broadsea

Los grupos de conceptos se pueden clasificar en tres temáticas según la utilidad que tienen en el estudio:

- **Condición prevalente.** La condición que prevalece en las base de datos es el cáncer de pulmón (CP).
- **Tratamientos oncológicos.** El estudio distingue tres tipos de tratamiento oncológico: radioterapia (RT), quimioterapia (QT) y cirugía torácica (IQ).

- **Toxicidades inducidas.** El estudio identifica seis tipos de toxicidades inducidas: disnea aguda (DA), disnea crónica (DC), neumonitis aguda (NA), neumonitis crónica (NC), tos (T) y esofagitis (E).

En este caso, aunque la base de datos no contiene información sobre las toxicidades inducidas, mediante la búsqueda genérica en el vocabulario (a través de la herramienta Search) se han definido grupos de conceptos relacionados con estas variables. De este modo se pueden hacer presente y continuar diseñando el estudio aunque no tengan presencia real en la base de datos.

La definición de cada grupo de concepto se ha exportado a archivos json y se encuentran accesibles en la ruta del repositorio de github Thesis-ATLAS-OHDSI/atlas/concept sets.

Cohortes

La definición de las cohortes es el componente central del estudio. Las cohortes son los componentes que luego combinándose entre sí y junto a otros parámetros conforman los distintos tipos de estudio (véase 5.4 “¿Cómo generar evidencia?”). La definición de cohortes se realiza a través del menú Cohort Definitions de ATLAS.

En el proyecto se han definido 14 cohortes combinando los grupos de conceptos definidos previamente.

Id	Name	Created	Updated	Author
19	[T5] Pacientes de CP que han recibido únicamente IQ	05/20/2024 5:44 PM	05/20/2024 5:45 PM	anonymous
18	Pacientes de CP	05/20/2024 11:06 AM	05/20/2024 1:05 PM	anonymous
12	Cirugía Torácica	05/20/2024 9:17 AM	05/20/2024 1:04 PM	anonymous
14	Radioterapia	05/20/2024 9:21 AM	05/20/2024 1:04 PM	anonymous
13	Quimioterapia	05/20/2024 9:18 AM	05/20/2024 1:04 PM	anonymous
5	[T4] Pacientes de CP que han recibido RT-QT sin IQ	05/15/2024 1:56 PM	05/20/2024 1:03 PM	anonymous
16	[T3] Pacientes de CP que han recibido únicamente RT	05/20/2024 10:54 AM	05/20/2024 1:03 PM	anonymous
17	[T2] Pacientes de CP que han recibido únicamente QT	05/20/2024 11:00 AM	05/20/2024 1:02 PM	anonymous
11	Tos	05/17/2024 6:59 PM	05/17/2024 6:59 PM	anonymous
10	Esofagitis	05/17/2024 6:59 PM	05/17/2024 6:59 PM	anonymous
9	Neumonitis Crónica	05/17/2024 6:58 PM	05/17/2024 6:58 PM	anonymous
8	Neumonitis Aguda	05/17/2024 6:57 PM	05/17/2024 6:57 PM	anonymous
7	Disnea Crónica	05/17/2024 6:44 PM	05/17/2024 6:45 PM	anonymous
6	Disnea Aguda	05/17/2024 12:49 PM	05/17/2024 6:44 PM	anonymous

Tabla 9.6: Listado de las 14 cohortes definidas en ATLAS

Se puede observar que hay dos formas diferentes de definir las cohortes en función de la utilidad que tienen en el estudio:

- **Cohorte principal (Target):** Este tipo de cohorte representa a la población sobre la que se realiza el estudio. Suelen ser poblaciones con características muy concretas y su definición incluye uno o varios criterios de inclusión más específicos.

En el proyecto, las cohortes principales son las cuatro cohortes cuyo nombre empieza por [T] y la cohorte primitiva Pacientes de CP. Se le denomina cohorte primitiva porque es la cohorte genérica que no aplica ninguna

restricción al conjunto de pacientes de cáncer de pulmón. Por otro lado, la etiqueta [T] en las cohortes es un identificador que la asocia con las cohortes definidas internamente en el estudio del HUVR.

- **Cohorte de resultado (*Outcome*):** Este tipo de cohorte representa a la población que experimenta un *outcome* o efecto adverso. Su definición es mucho más sencilla, simplemente representan la población que experimenta una condición concreta.

Su función es servir de parámetro para configurar los estudios posteriores que se realizan sobre las cohortes principales.

La definición de las cohortes se ha exportado a archivos json y sql y se encuentran accesibles en la ruta del repositorio de github Thesis-ATLAS-OHDSI/atlas/cohort definitions.

Caracterización

Con esta sección comienza el primero de los casos de uso definidos para la investigación observacional de OHDSI. Aunque estrictamente la generación de reportes de la base de datos también se considera caracterización, esta sección se centra en la caracterización de cohortes.

Para ello se emplean dos herramientas de ATLAS: Characterization y Cohort Pathway.

La caracterización consiste en la obtención de un reporte con las características estadísticas más relevantes de la cohorte.

En este caso se ha realizado una caracterización que compara tres cohortes principales del estudio: pacientes que han recibido únicamente quimioterapia, pacientes que han recibido únicamente radioterapia, y pacientes que han recibido ambas terapias sin sufrir cirugía torácica. Esta última cohorte es la que se utiliza para diseñar el resto de casos de uso de la investigación.

CAPÍTULO 9. CASO PRÁCTICO

Cohort definition

Import

Show 10 entries Filter: Search...

Id		Name	Actions
5	[T4]	Pacientes de CP que han recibido RT-QT sin IQ	Edit cohort Remove
16	[T3]	Pacientes de CP que han recibido únicamente RT	Edit cohort Remove
17	[T2]	Pacientes de CP que han recibido únicamente QT	Edit cohort Remove

Showing 1 to 3 of 3 entries Previous 1 Next

Feature analyses

Import

Show 10 entries Filter: Search...

Id		Name	Description	Actions
41		Distinct Condition Count Long Term	The number of distinct condition concepts observed in the long term window.	Remove
47		Distinct Procedure Count Short Term	The number of distinct procedures observed in the short term window.	Remove
48		Distinct Condition Count Short Term	The number of distinct condition concepts observed in the short term window.	Remove
51		Distinct Procedure Count Medium Term	The number of distinct procedures observed in the medium term window.	Remove
54		Distinct Condition Count Medium Term	The number of distinct condition concepts observed in the medium term window.	Remove
66		Distinct Procedure Count Long Term	The number of distinct procedures observed in the long term window.	Remove
93		Distinct Observation Count Medium Term	The number of distinct observations observed in the medium term window.	Remove
94		Distinct Observation Count Long Term	The number of distinct observations observed in the long term window.	Remove
97		Distinct Observation Count Short Term	The number of distinct observations observed in the short term window.	Remove

Tabla 9.7: Definición del análisis estadístico de las cohortes principales

Las variables seleccionadas para diseñar el estudio de caracterización han sido aquellas relacionadas con las distintas condiciones, procedimientos y observaciones que experimentan las cohortes. Los resultados y el diseño del análisis se han exportado y están disponibles en la ruta del repositorio de github Thesis-ATLAS-OHDSI/atlas/characterization s/characterization_1_execution_72_reports.

Otra forma de caracterizar una cohorte es a través de la ruta de la cohorte. En este caso se ha diseñado una ruta de la cohorte que estudie las diferentes rutas de tratamiento (quimioterapia, radioterapia y cirugía torácica) que experimenta la cohorte primitiva de cáncer de pulmón..

Target Cohorts

Each of the Target Cohorts will be analyzed for the pathways through the event cohorts.

Import

Show 10 entries Filter: Search...

Id		Name	Actions
18		Pacientes de CP	Edit cohort Remove

Showing 1 to 1 of 1 entries Previous 1 Next

Event Cohorts

Each Event Cohort defines the step in a pathway that may occur for a person in the Target Cohort.

Import

Show 10 entries Filter: Search...

Id		Name	Actions
12		Cirugía Torácica	Edit cohort Remove
13		Quimioterapia	Edit cohort Remove
14		Radioterapia	Edit cohort Remove

Showing 1 to 3 of 3 entries Previous 1 Next

Tabla 9.8: Definición del análisis de la ruta de la cohorte de cáncer de pulmón

Pathways Analysis for ChP pacientes CP que experimentan RT, QT o IQ

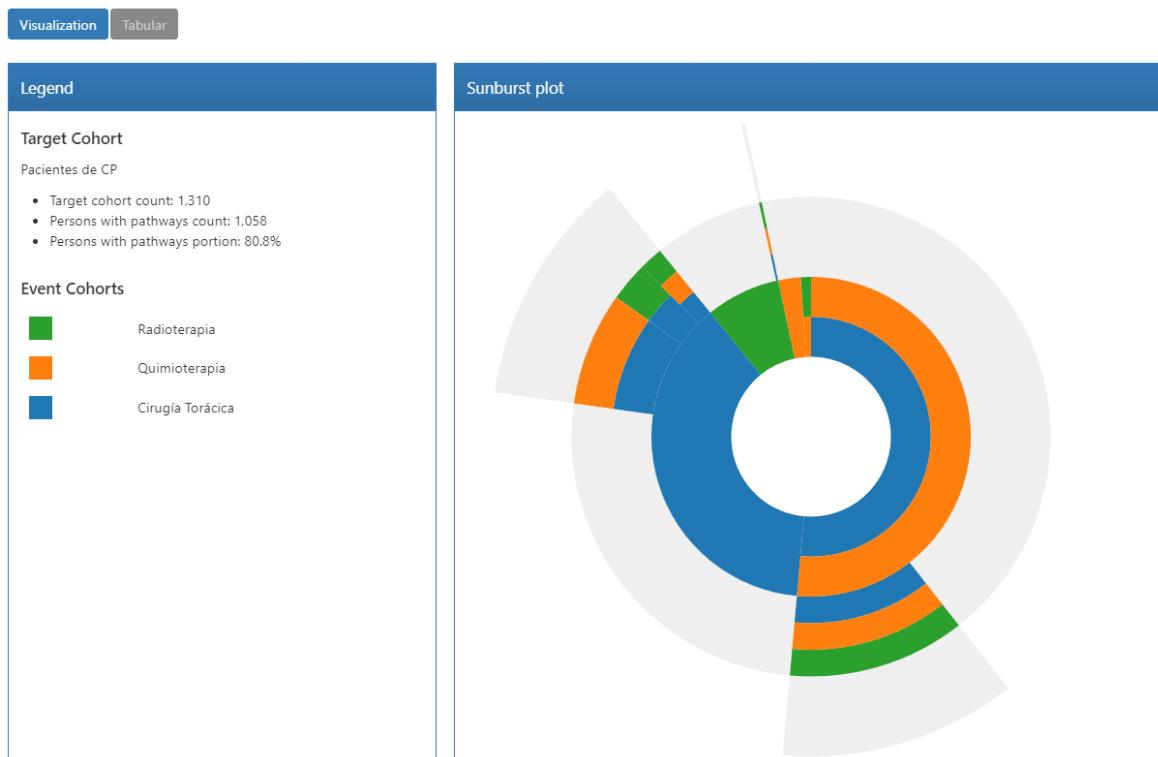


Figura 9.3: Análisis de las rutas de la cohorte de cáncer de pulmón

En este caso, el resultado se muestra en la Figura 9.3 que se obtiene es una "gráfica de explosión solar" o *sunburst* que muestra las diferentes rutas que experimenta la cohorte. No obstante, tanto la definición del estudio como los resultados también se encuentran accesibles a través de la ruta del repositorio de github Thesis-ATLAS-OHDSI/atlas/characterizations/statistics

Estimación a Nivel Población

La estimación a nivel de población consiste en la comparación entre los efectos adversos que sufrirá una cohorte principal en comparación con otra cohorte comparadora. En ATLAS, esta tarea se realiza a través del menú Prediction.

El estudio original del HUVR no realiza explícitamente una estimación a nivel de población aunque se podría adaptar el estudio para realizar un estudio entre los seis efectos adversos (DA, DC, NA, NC, E, T) que sufren los pacientes que han recibido únicamente tratamiento radioterápico (cohorte T2) en comparación con aquellos que han recibido únicamente tratamiento quimioterápico (cohorte T3).

A continuación se muestra la definición del problema de estimación mediante la interfaz gráfica de ATLAS.

The screenshot shows the 'Comparison' section of the ATLAS interface. It includes fields for selecting target cohorts (T1), comparator cohorts (T2), and outcome cohorts. A table lists various outcomes with columns for ID, Name, Edit cohort, and Remove. A negative control concept set is also listed.

ID	Name	Edit cohort	Remove
6	Disnea Aguda		
7	Disnea Crónica		
8	Neumonitis Aguda		
9	Neumonitis Crónica		
10	Esofagitis		
11	Tos		

Figura 9.4: Ajustes de la comparación para la estimación a nivel de población en ATLAS

Para configurar los ajustes del análisis se han seguido las recomendaciones del capítulo 12 "Population-Level Estimation" del Libro de OHDSI [3], correspondientes con los ajustes establecidos por defecto por la propia herramienta ATLAS.

The screenshots show the 'Analysis Settings' and 'Evaluation Settings' sections of the ATLAS interface. The analysis settings table includes rows for Propensity Score Matching with specific parameters like Time At Risk Start and End, and Outcome Model set to cox. The evaluation settings page details the 'Negative Control Outcome Cohort Definition'.

Figura 9.5: Ajustes predeterminados para la estimación a nivel de población en ATLAS

Una vez que se define el estudio se genera un paquete R para ejecutar el estudio en un entorno de programación más avanzado. **La tarea de ATLAS no es realizar este tipo de estudios sino estandarizar la forma de definirlos.** Dicho paquete está disponible en el repositorio de github en la ruta Thesis-ATLAS-OHDSI/atlas/estimation.

Predicción a Nivel de Paciente

Por último en este caso, el estudio realizado por el HUVR correspondía concretamente a una predicción a nivel de paciente, tal y como se justifica en apartados anteriores. Esta tarea en ATLAS se realiza a través del menú Prediction

Se pretende estudiar la probabilidad de experimentar alguno de los seis efectos adversos (DA, DC, NA, NC, T, E) en los pacientes que reciben tratamiento radioterápico o quimioterápico indistintivamente sin haber sufrido cirugía torácica (cohorte T4).

A continuación se muestra la definición del problema de predicción mediante la interfaz gráfica de ATLAS.

The screenshot shows the 'Prediction Problem Settings' interface in ATLAS. It is divided into two main sections: 'Target Cohorts' and 'Outcome Cohorts'.

Target Cohorts:

- Header: '+ Add Target Cohort' button.
- Table: Shows one entry: 'T4 Pacientes de CP que han recibido RT-QT sin IQ'. A red 'X' icon indicates it can be removed.
- Buttons: 'Show 10 entries' dropdown, 'Remove' button, 'Name' input field, 'Filter: Search...' input field, 'Previous' and 'Next' navigation buttons.
- Text: 'Showing 1 to 1 of 1 entries'.

Outcome Cohorts:

- Header: '+ Add Outcome Cohort' button.
- Table: Shows six entries: 'Tos', 'Esofagitis', 'Neumonitis Crónica', 'Neumonitis Aguda', 'Disnea Crónica', and 'Disnea Aguda'. Each has a red 'X' icon.
- Buttons: 'Show 10 entries' dropdown, 'Remove' button, 'Name' input field, 'Filter: Search...' input field, 'Previous' and 'Next' navigation buttons.
- Text: 'Showing 1 to 6 of 6 entries'.

Figura 9.6: Ajustes para la predicción a nivel de paciente en ATLAS

De nuevo, una vez que se define el estudio se genera un paquete R para ejecutar el estudio en un entorno de programación más avanzado debido a que la tarea de ATLAS no es realizar este tipo de estudios sino estandarizar la forma de definirlos. Dicho paquete está disponible en el repositorio de github en la ruta Thesis-ATLAS-OHDSI/atlas/prediction

9.3.3. Resultados

Los resultados del estudio son algo limitados porque no se pueden ejecutar la mayoría de los análisis debido a que la base de datos no contiene la información sobre los efectos adversos y los estudios giran alrededor de la caracterización, estimación o predicción de estos efectos en las cohortes. A continuación se listan algunos resultados:

- **Posibilidad de diseñar el estudio sin poseer los datos.** Es destacable y de gran interés la facilidad que proporciona ATLAS para diseñar estudios precisamente sin disponer de los datos exactos, como es este caso. Gracias a la abstracción lógica de los datos mediante los grupos de conceptos y cohortes,

se pueden diseñar los estudios sin necesidad de manipular los datos subyacentes. Esto es también un beneficio a la hora de realizar los estudios en diferentes bases de datos.

De hecho, si Eunomia (la base por defecto de ATLAS Broadsea) incluyese información oncológica, se podrían ejecutar los estudios a la vez sobre las dos bases de datos y comparar resultados de una forma muy sencilla.

- **Reproducibilidad del estudio.** También, como se ha mencionado durante la metodología, los resultados más destacables son la cantidad de archivos exportables que genera ATLAS con la finalidad de favorecer la reproducibilidad del estudio. Prácticamente cada paso que se da durante el diseño del estudio en ATLAS es exportable, y de igual forma importable, permitiendo el intercambio de información entre sistemas e investigadores.
- **Limitación en la ejecución del estudio.** No se obtienen resultados tangibles o concretos sobre la estimación o la predicción porque no se llegan a ejecutar estos estudios en el entorno R, debido a la carencia de la base de datos y a que su ejecución se puede considerar fuera del alcance del proyecto, si bien se definía anteriormente que el objetivo del estudio no es reproducirlo sino estandarizarlo. En clave de estandarización se ha demostrado que ATLAS es una herramienta muy poderosa para realizar esta tarea mediante una interfaz muy intuitiva y fácil de usar.

9.4. Discusión de resultados

Volviendo al objetivo inicial del estudio, a la hora de facilitar y apoyar la toma de decisiones durante la planificación radioterápica en pacientes oncológicos, ambos estudios podrían proporcionar resultados muy positivos.

Si bien es verdad que no se ha podido ejecutar completamente el estudio diseñado en ATLAS por las limitaciones encontradas, en mi opinión en este caso no es un factor que haya impedido demostrar que la herramienta está más que preparada para apoyar la decisión clínica y realiza dicha tarea de forma más beneficiosa que la programación directamente sobre un script de código, tal y como se presentaba en la sección 5.4.3 "Vías de implementación del análisis".

Una característica fundamental para ello que sí se ve reflejada durante el análisis es la interfaz gráfica intuitiva y las facilidades que proporciona la herramienta para conducir cualquier estudio. El hecho de que ATLAS sea *low-code* y esté estructurado de una forma preestablecida favorece la conducción del estudio según las opciones pre-establecidas de la herramienta, siendo esto mucho más fácil que enfrentarse directamente a un script de código en blanco, aparte de los beneficios en términos de estandarización del análisis que conlleva.

También aporta el beneficio de poder exportar el código subyacente a la herramienta, de modo que una tarea tan compleja a nivel de conocimiento informático como programar un problema de predicción o estimación mediante

técnicas de Machine Learning, queda simplificado a un "juego" de combinar cohortes y parámetros de análisis. ATLAS permite programar sin saber programar y eso es una ventaja absoluta sobre cualquier otra herramienta.

Al fin y al cabo, si el objetivo es ayudar al diagnóstico médico se debe tener en cuenta que el doctor no tiene por qué poseer conocimientos informáticos. Sería más fácil para un doctor entender ATLAS que aprender a programar en python o R. Por tanto, ATLAS puede favorecer también el interés por el análisis de datos en profesionales cuyos conocimientos informáticos sean más limitados pero quizás sus conocimientos clínicos sean más profundos, favoreciendo también la investigación clínica de esta forma.

9.5. Conclusiones

ATLAS es fundamental para la estandarización de los sistemas de informática clínica, mejorando la interoperabilidad y la calidad de la investigación. Su configuración *low-code* democratiza la investigación clínica observacional, permitiendo que profesionales sin conocimientos avanzados en programación participen en estudios complejos. Además, las guías del Libro de OHDSI son cruciales para realizar estudios rigurosos y reproducibles, permitiendo a investigadores sin experiencia informática llevar a cabo investigaciones de alta calidad, elevando la credibilidad y validez científica.

10. Resultados

Este capítulo pone fin al desarrollo teórico-práctico del proyecto, realizando una recopilación de los resultados obtenidos durante el mismo. A continuación se listan los resultados obtenidos:

- **Res-001: Conocimiento y participación activa en la comunidad de OHDSI.** Al término del proyecto, se ha obtenido un conocimiento teórico exhaustivo de OHDSI, de sus estándares y sus herramientas y por supuesto de la herramienta ATLAS en todos sus aspectos, tal y como se detalla en los aspectos teóricos del trabajo.

Sin embargo, lo más importante de haber conocido y comprendido la organización Observational Health Data Sciences and Informatics ha sido entender la importancia de la participación colaborativa en la comunidad, que he realizado mediante la participación en foros de la comunidad [38] y formando parte de los grupos de Discord [35] y MS Teams de la comunidad. Mi proyecto termina pero el camino con OHDSI (*"the journey"*) todavía continúa.

- **Res-002: Aplicación y redacción del manual de despliegue de ATLAS Broadsea.** Frente a todos los problemas encontrados durante la instalación, despliegue y configuración de ATLAS Broadsea, haber dejado por escrito una guía detallada del proceso es uno de los resultados más destacables del proyecto, por la relevancia que tendrá para el resto de investigadores poder poseer de un manual sencillo que comprenda todos los conocimientos teóricos y prácticos necesarios para realizar esta tarea.

Además, por la naturaleza práctica del proyecto realizado en colaboración con el HUVR, el interés de este manual es aún mayor, pues les facilitará enormemente realizar esta tarea si la requiriesen una vez cese mi periodo de prácticas.

- **Res-003: Desarrollo de un estudio reproducible a través de ATLAS y el repositorio de github.** Por último, la utilización de ATLAS para demostrar la viabilidad de la herramienta a la hora de conducir estudios de análisis de datos de forma sencilla, a través de guías de buenas prácticas y múltiples tutoriales online ha sido un resultado muy importante.

Sin embargo, lo más destacable del caso práctico no ha sido solo la facilidad con la que se ha podido reproducir el estudio sino la facilidad para generar código exportable a partir del mismo y compartirlo en el repositorio de github del proyecto, alineando el TFG a la visión *open-source* y *open-science* de la comunidad de OHDSI.

10.1. Trazabilidad de objetivos

Los resultados obtenidos durante el proyecto casan con los objetivos definidos en el capítulo 2 "Objetivos del Proyecto".

A continuación se muestra una tabla que señala los objetivos que satisface cada uno de los resultados obtenidos. Se contemplan los objetivos del proyecto y los objetivos personales de la alumna.

	Obj-001: Estudio teórico de organización OHDSI y herramienta ATLAS	Obj-002: Instalación, despliegue y configuración de ATLAS mediante Broadsea	Obj-003: Estandarización de caso práctico de análisis de datos clínicos proporcionados por el HUVR	Obj-Pers-001: Aumentar mi conocimiento sobre la comunidad OHDSI y sus herramientas	Obj-Pers-002: Aumentar mi conocimiento del mundo del análisis de datos.	Obj-Pers-003: Aumentar mi experiencia laboral analizando datos clínicos
Res-001: Conocimiento y participación activa en la comunidad de OHDSI	X	X	X	X	X	
Res-002: Aplicación y redacción del manual de despliegue de ATLAS Broadsea	X	X		X		
Res-003: Desarrollo de un estudio reproducible a través de ATLAS y el repositorio de github		X	X	X	X	X

Tabla 10.1: Trazabilidad de objetivos con resultados

- **Trazabilidad del Res-001.** Este resultado abarca prácticamente todos los objetivos, puesto que el conocimiento teórico de la organización es subyacente a cada paso en la elaboración del proyecto, tanto en la teoría como en la práctica, haciendo especial hincapié en la parte práctica a la hora de participar en los foros de la comunidad publicando y revisando preguntas sobre las herramientas empleadas durante el desarrollo del proyecto.
- **Trazabilidad del Res-002.** Este resultado abarca objetivos de aspecto teórico y de implementación. No se considera que contribuya a los objetivos de análisis porque realiza más bien las tareas de desarrollador, montando la estructura Docker y la base de datos del sistema (véase 7 "Entorno de Trabajo"). El estudio teórico de OHDSI ha sido fundamental para poder implementar correctamente el sistema.
- **Trazabilidad del Res-003.** Este resultado abarca los objetivos relacionados con el análisis y con la herramienta de ATLAS Broadsea, que al fin y al cabo es la que se ha utilizado para llevar a cabo el análisis. Además del análisis, el repositorio de github también recopila archivos importantes y documentación del manual por lo que también se relaciona con este objetivo.

10.2. Lecciones aprendidas

Este proyecto ha sido de gran relevancia para culminar mi formación académica y dar inicio a mi carrera profesional, permitiéndome aplicar los conocimientos adquiridos en el Grado de Ingeniería de la Salud y abriendo nuevas oportunidades en el ámbito del tratamiento de datos clínicos. La integración de teoría y práctica en este proyecto ha sido fundamental para mi desarrollo como profesional en esta disciplina.

La lección más significativa que he obtenido de este proyecto es **la importancia de aprender de los errores y ser capaz de enfrentar y resolver problemas inesperados**. Esta habilidad es esencial para cualquier ingeniero, ya que los desafíos imprevistos son comunes en el ámbito profesional y académico.

Aprender a desplegar y utilizar ATLAS de manera autodidacta ha sido un reto sumamente valioso. Este proceso no solo implicó una investigación exhaustiva para comprender el funcionamiento de la herramienta y sus configuraciones, sino también una inmersión en el ecosistema de herramientas y estándares de OHDSI. Esta experiencia me ha formado como una experta en el campo, destacando la importancia de la investigación independiente y el autoaprendizaje en el desarrollo profesional.

Estos conocimientos teóricos me han permitido comprender la necesidad crítica de la interoperabilidad entre los sistemas de información, especialmente en el ámbito sanitario, y los desafíos y limitaciones que esto implica. Reconozco la labor crucial que realiza OHDSI para estandarizar la investigación clínica, subrayando su relevancia a nivel mundial y europeo. Aunque actualmente OHDSI no es el estándar predominante, estoy convencida de que lo será en el futuro, y los conocimientos adquiridos ahora serán aún más valiosos en los próximos años.

Gracias a todo lo aprendido, actualmente colaboro con el grupo de Informática de la Salud del HUVR en el proyecto europeo IMPaCT-Data. En este proyecto, he contribuido a la redacción de secciones que proponen soluciones relacionadas con OHDSI, el Modelo de Datos Común de OMOP y ATLAS. Esta colaboración ha sido una oportunidad invaluable para aplicar mis conocimientos en un contexto real y contribuir a un proyecto de gran impacto.

La colaboración con el HUVR y con Francisco Rey Garduño, en relación con nuestros Trabajos de Fin de Grado, me han enseñado mucho sobre el trabajo en equipo, la importancia de la distribución de tareas, la planificación del tiempo y la asunción de responsabilidades. Estas habilidades son esenciales para el éxito en cualquier proyecto profesional.

Desde una perspectiva práctica, el despliegue de ATLAS me ha aportado nuevos conocimientos sobre Docker y ha mejorado mis habilidades en bases de datos. Nunca había trabajado con Docker antes de este proyecto, y aprender sobre la configuración de contenedores y perfiles, la identificación y resolución de errores en los logs, y la interacción con volúmenes de datos ha sido una experiencia enriquecedora. Además, al trabajar con bases de datos PostgreSQL, he perfeccionado mis habilidades en SQL y en la configuración de servidores y bases de datos.

En conclusión, este proyecto no solo ha sido un cierre significativo de mi formación académica, sino también una puerta de entrada a mi carrera profesional en el tratamiento de datos clínicos. He adquirido habilidades prácticas, desarrollado capacidades de investigación independiente y comprendido la importancia de la interoperabilidad y estandarización en la investigación clínica. Estas experiencias y conocimientos no solo son valiosos ahora, sino que también serán cruciales para mi futuro profesional, permitiéndome contribuir de manera significativa a este sector.

11. Conclusiones

El análisis de datos clínicos está adquiriendo una importancia cada vez mayor a nivel mundial, especialmente mediante estudios observacionales a gran escala que utilizan datos del mundo real. Estos estudios son fundamentales para generar evidencia real que pueda mejorar la toma de decisiones en el ámbito sanitario. Una de las necesidades cruciales en este contexto es la interoperabilidad entre los diferentes sistemas de información. La búsqueda de estándares que faciliten esta interoperabilidad es esencial, pero también conlleva desafíos significativos.

En este escenario, la iniciativa OHDSI (Observational Health Data Sciences and Informatics) juega un papel crucial. OHDSI se dedica a la estandarización de la investigación clínica y tiene una presencia destacada en numerosos proyectos europeos como EHDEN, IMPaCT-Data, y EUCAIM. Estas colaboraciones subrayan la importancia de OHDSI en el esfuerzo por crear un marco común para el análisis de datos clínicos, que permita a los investigadores de todo el mundo colaborar y compartir conocimientos de manera más eficiente.

Uno de los aspectos más destacados de OHDSI es su modelo de datos OMOP (Observational Medical Outcomes Partnership), que proporciona una estructura estandarizada para la recopilación y análisis de datos de salud. El modelo OMOP permite la armonización de datos provenientes de diversas fuentes, lo que es crucial para realizar estudios multicéntricos y comparativos. Por otro lado, la herramienta ATLAS, realiza una tarea muy importante en la exploración y análisis de datos clínicos a través de su interfaz *low-code*, de gran relevancia para facilitar la realización de estos análisis de forma más sencilla, eficiente y precisa.

La colaboración en redes como OHDSI, así como en espacios de trabajo más limitados como el grupo de Informática de la Salud del Hospital Universitario Virgen del Rocío es fundamental para el intercambio de conocimientos y experiencias, y fomenta la innovación y mejoran la calidad de los resultados de la investigación. Trabajar en equipo y compartir responsabilidades dentro de un entorno colaborativo es esencial para enfrentar los desafíos complejos del análisis de datos clínicos y para avanzar en la investigación sanitaria.

En conclusión, la implementación y uso del estándar OHDSI y sus herramientas asociadas han demostrado ser altamente beneficiosos para la investigación clínica y la práctica médica. La capacidad de estandarizar y analizar datos de salud de manera eficiente y coherente tiene un impacto significativo en la mejora de la calidad de la atención sanitaria y en el avance de la investigación médica. La creciente adopción de OHDSI en proyectos europeos subraya su relevancia y potencial para transformar el panorama de la salud pública y la investigación biomédica. Este proyecto ha proporcionado una valiosa experiencia práctica y ha resaltado la importancia de continuar promoviendo la estandarización y la interoperabilidad en los sistemas de salud.

Bibliografía

- [1] Scrum. Scrum home page, 2024. URL <https://www.scrum.org/>.
- [2] Observational Health Data Sciences and Informatics. Ohdsi.org. <https://www.ohdsi.org/>, 2024.
- [3] Observational Health Data Sciences and Informatics (OHDSI). *The Book of OHDSI*. OHDSI, 2022. URL <https://ohdsi.github.io/TheBookOfOhdsi/>.
- [4] Observational Health Data Sciences and Informatics (OHDSI). Common data model, 2023. URL <https://ohdsi.github.io/CommonDataModel/index.html>.
- [5] OHDSI github. Atlas, 2024. URL <https://github.com/OHDSI/Atlas>.
- [6] OHDSI github. Ohdsi webapi wiki, 2023. URL <https://github.com/OHDSI/WebAPI/wiki>.
- [7] A. Londhe. Slides from the 2023 ohdsi global symposium, 2023. URL <https://www.ohdsi.org/wp-content/uploads/2023/10/419-Londhe-Slides.pdf>.
- [8] F J Núñez-Benjumea, S González-García, A Moreno-Conde, J C Riquelme-Santos, and J L López-Guerra. Benchmarking machine learning approaches to predict radiation-induced toxicities in lung cancer patients. *Clinical and translational radiation oncology*, 41:100640, 2023. doi: 10.1016/j.ctro.2023.100640. URL <https://doi.org/10.1016/j.ctro.2023.100640>.
- [9] 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.
- [10] Chiehfeng 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.
- [11] 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.
- [12] 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.
- [13] 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.

- [14] 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.
- [15] 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.
- [16] 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.
- [17] Dimitrios G Katehakis and Angelina Kouroubali. A framework for ehealth interoperability management. *Journal of Strategic Innovation and Sustainability*, 14(5):51–61, 2019.
- [18] 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>.
- [19] Kécia Souza Santana Santos, Larissa Barbosa Leoncio Pinheiro, and Rita Suzana Pitangueira Maciel. Interoperability types classifications: A tertiary study. 2021. doi: 10.1145/3466933.3466952. URL <https://doi.org/10.1145/3466933.3466952>.
- [20] Gabriel da Silva Serapião Leal, Wided Guédria, and Hervé Panetto. Interoperability assessment: A systematic literature review. *Computers in Industry*, 106:111–132, 2019.
- [21] Rebeca C Motta, Káthia M de Oliveira, and Guilherme H Travassos. A conceptual perspective on interoperability in context-aware software systems. *Information and Software Technology*, 114:231–257, 2019.
- [22] 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>.
- [23] 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>.
- [24] DigitalHealthEurope. eHDSI - European Health Data Space, 2023. URL <https://digitalhealtheurope.eu/glossary/ehdsi/>.
- [25] European Genomic Data Infrastructure (GDI) project. European genomic data infrastructure (gdi) project, 2022. URL <https://gdi.onemilliongenomes.eu/>.
- [26] vallealonsodc. Thesis-ATLAS-OHDSI. <https://github.com/vallealonsodc/Thesis-ATLAS-OHDSI>, 2024.

- [27] Junta de Andalucía. Temas: Perfiles de contratante. https://www.juntadeandalucia.es/haciendayadministracionpublica/apl/pdc_sirec/perfiles-licitaciones/consultas-preliminares/detalle.jsf?idExpediente=000000078484, 2018.
- [28] Microsoft. Comprar windows 11 pro — microsoft store españa. <https://www.microsoft.com/es-es/d/windows-11-pro>, 2024.
- [29] Sparx Systems. Enterprise architect pricing, 2024. URL <https://sparxsystems.com/products/ea/shop/>.
- [30] Microsoft. Compra microsoft 365 personal — microsoft store españa. <https://www.microsoft.com/es-es/microsoft-365/p/microsoft-365-personal>, 2024.
- [31] MJ Escalona, L García, JA García-García, G López-Nicolás, and N Koch. Choose your preferred life cycle and sofia will do the rest. 2023.
- [32] Wikipedia. Latex, 2024. URL <https://es.wikipedia.org/wiki/LaTeX>.
- [33] Observational Health Data Sciences and Informatics. Publications - ohdsi.org. <https://www.ohdsi.org/publications/>, 2024.
- [34] Observational Health Data Sciences and Informatics (OHDSI). Ohdsi youtube channel, 2023. URL <https://www.youtube.com/c/OHDSIorg/videos>.
- [35] OHDSI discord server invitation. <https://discord.com/invite/xABFWShJYx>, 2024.
- [36] Observational Health Data Sciences and Informatics (OHDSI). Ohdsi google office forms, 2023. URL <https://docs.google.com/forms/d/1QVvNt8qap9QsNWwWw1Yt0vLqQhjh4sk>.
- [37] Observational Health Data Sciences and Informatics (OHDSI). Ohdsi github repository, 2023. URL <https://github.com/OHDSI/>.
- [38] OHDSI. Ohdsi forums, 2024. URL <https://forums.ohdsi.org/>.
- [39] P. E. et al Stang. Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership, 2010.
- [40] J. M. et al Overhage. Validation of a common data model for active safety surveillance research., 2012.
- [41] George Hripcsak and David J Albers. High-throughput phenotyping with electronic health records. *Journal of the American Medical Informatics Association*, 25(11):1392–1395, 2018. doi: 10.1093/jamia/ocy019. URL <https://doi.org/10.1093/jamia/ocy019>.
- [42] Vishnu Chandrabalan. Schemaspy analysis of omop, 2024. URL <https://omop-erd.surge.sh/index.html>.
- [43] OHDSI. Athena, 2024. URL <https://athena.ohdsi.org/search-terms/terms>.

- [44] OHDSI. Atlas demo, 2024. URL <https://atlas-demo.ohdsi.org/>.
- [45] OHDSI github. Broadsea, 2023. URL <https://github.com/OHDSI/Broadsea>.
- [46] OHDSI github. Ohdsi in a box, . URL <https://github.com/OHDSI/OHDSI-in-a-Box>.
- [47] OHDSI github. Ohdsi aws, . URL <https://github.com/OHDSI/OHDSIonAWS>.
- [48] OHDSI github. Ohdsi on azure, 2024. URL <https://github.com/microsoft/OHDSIonAzure>.
- [49] OHDSI github. Achilles, 2024. URL <https://github.com/OHDSI/Achilles>.
- [50] OHDSI github. Athena, 2024. URL <https://github.com/OHDSI/Athena>.
- [51] OHDSI github. Hades, 2024. URL <https://github.com/OHDSI/Hades>.
- [52] OHDSI. Software tools, 2024. URL <https://www.ohdsi.org/software-tools/>.
- [53] OHDSI github. Data quality dashboard, 2024. URL <https://github.com/OHDSI/DataQualityDashboard>.
- [54] Google. Google chrome, 2024. URL https://www.google.com/intl/es_es/chrome/.
- [55] Docker. Descripción general de docker, 2024. URL <https://docs.docker.com/get-started/overview/>.
- [56] PostgreSQL. Acerca de postgresql, 2024. URL <https://www.postgresql.org/about/>.
- [57] pgAdmin. pgadmin, 2024. URL <https://www.pgadmin.org/>.
- [58] Wikipedia. Github, 2024. URL <https://en.wikipedia.org/wiki/GitHub>.
- [59] OHDSI. Ohdsi common data model configuration (wiki), 2023. URL <https://github.com/OHDSI/WebAPI/wiki/CDM-Configuration>.
- [60] F J Núñez-Benjumea, J Moreno-Conde, S González-García, A Moreno-Conde, J L López-Guerra, M J Ortiz-Gordillo, and C L Parra-Calderón. Comparison of feature selection methods for predicting rt-induced toxicity. *Studies in health technology and informatics*, 258:253–254, 2019.

A. Manual de ATLAS Broadsea

El nombre completo de este anexo corresponde a **Manual de instalación, despliegue y configuración de ATLAS Broadsea**, aunque por motivos de extensión se ha reducido en el índice de la memoria a *Manual de ATLAS Broadsea*.

El manual se presenta a la convocatoria como un documento aparte debido a su larga extensión, de casi 40 páginas. No obstante, se utiliza este apartado de la memoria para presentar resumidamente sus contenidos básicos y cómo acceder a él. Su gran extensión se debe a que recopila en un único lugar una grandísima variedad de información que hasta ahora se encontraba esparcida de forma más o menos ordenada en la red, sobretodo en diferentes repositorios de github.

El anexo se adjunta a la documentación entregable de la convocatoria con el nombre "Anexo A - Manual de ATLAS Broadsea.pdf". Adicionalmente, también es accesible a través del repositorio de github del Trabajo Fin de Grado [26], concretamente en la ruta Thesis-ATLAS-OHDSI/docs/pdf/manual.pdf.

El manual trata cinco aspectos importantes de ATLAS Broadsea:

1. **Introducción y descripción de Broadsea.** Este capítulo explica contenidos sobre el entorno tecnológico necesario para seguir correctamente los procedimientos del manual.
2. **Despliegue por defecto.** Este capítulo presenta el despliegue más sencillo del entorno Broadsea, sin ningún tipo de configuración adicional.
3. **Conexión con la BD por defecto.** Este capítulo explica la conexión con el servidor Postgre del contenedor docker de Broadsea.
4. **Conexión con BD externa.** Este capítulo explica cómo añadir una conexión de una base de datos externa al servidor docker de Broadsea.
5. **Configuración del Vocabulario.** Este capítulo explica cómo configurar el Vocabulario desde ATHENA y se presentan otras configuraciones avanzadas.

Todo ello complementa la información del TFG de forma subyacente, es decir, durante la reproducción del estudio práctico (véase ?? "Caso práctico") se da por supuesto todo el proceso de instalación de la herramienta así como la configuración del servidor, base de datos, etc. En términos de roles del proyecto (véase 3 "Gestión del proyecto") se podría decir que mientras que el analista se encarga de reproducir el estudio haciendo uso de la interfaz de usuario de ATLAS, el developer habría sido el encargado de realizar toda el anexo, con toda la instalación, despliegue y configuración para que la herramienta funcione. No obstante, en este caso ambos roles son ejecutados por la misma persona que es la alumna. Además satisface explícitamente el **Obj-002: Instalación, configuración y despliegue de ATLAS mediante Broadsea** del Trabajo Fin de Grado (véase 2 "Objetivos del Proyecto").

B. Glosario

Aprendizaje automático (*Machine Learning, ML*): Campo de la inteligencia artificial que desarrolla algoritmos y modelos que permiten a las máquinas aprender a partir de datos, identificar patrones y tomar decisiones sin necesidad de ser programadas explícitamente para cada tarea específica.

ATLAS: Herramienta de código abierto desarrollada por la colaboración Observational Health Data Sciences and Informatics (OHDSI), diseñada para la visualización, exploración y análisis de datos de salud provenientes de diferentes fuentes y estándares, facilitando la investigación en salud pública y la toma de decisiones clínicas basadas en evidencia.

Contenedor Docker (*Docker container*): Tecnología de virtualización que permite empaquetar y ejecutar aplicaciones y sus dependencias en entornos aislados, proporcionando portabilidad, rapidez y consistencia en el despliegue de aplicaciones en diferentes sistemas operativos y entornos de ejecución.

Código abierto (*Open source*): Modelo de desarrollo de software que promueve el acceso abierto al código fuente de un programa, permitiendo su estudio, modificación y distribución por parte de la comunidad de desarrolladores, lo que fomenta la colaboración, la transparencia y la innovación en el desarrollo de software.

Computación en la Nube (*Cloud Computing*): Modelo de prestación de servicios de computación a través de internet, donde los recursos como almacenamiento, servidores y aplicaciones son proporcionados y gestionados por proveedores externos, permitiendo un acceso flexible y escalable según la demanda del usuario.

Cohorte (*Cohort*): Grupo de individuos que comparten una característica común o que han sido seleccionados para participar en un estudio de investigación, con el fin de observar y analizar los resultados de un evento o exposición específica durante un período de tiempo determinado.

Datos masivos (*Big Data*): Conjunto de datos extremadamente grandes y complejos que requieren tecnologías especializadas para su almacenamiento, procesamiento y análisis, con el objetivo de extraer información significativa y tomar decisiones informadas.

Datos del mundo real (*Real World Data, RWD*): Información sobre la salud y los resultados de atención médica recopilada de fuentes del mundo real, como registros médicos electrónicos, reclamaciones de seguros y dispositivos portátiles, utilizada para complementar los datos de ensayos clínicos y proporcionar información sobre la efectividad y seguridad de tratamientos en condiciones reales fuera del entorno controlado de un estudio clínico.

European Health Data & Evidence Network (EHDEN): Consorcio europeo que tiene como objetivo establecer una infraestructura escalable y sostenible para el

análisis de datos de salud del mundo real en Europa. EHDEN promueve la estandarización de datos y el uso de herramientas y métodos avanzados para facilitar la investigación clínica y epidemiológica.

Historial Clínico Electrónico (HCE): Registro digitalizado y centralizado de toda la información médica de un paciente, que incluye datos como diagnósticos, tratamientos, resultados de pruebas, alergias y antecedentes médicos, accesible por profesionales de la salud autorizados para mejorar la coordinación de la atención, la precisión diagnóstica y la seguridad del paciente.

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.

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.

Interoperabilidad: Capacidad de sistemas, dispositivos o aplicaciones para intercambiar datos y trabajar juntos de manera efectiva, garantizando que la información sea comprensible y utilizada de manera consistente entre diferentes plataformas, organizaciones o entornos. Se puede clasificar en tres grupos: semántica, técnica y organizacional.

Low-code: Enfoque de desarrollo de software que utiliza herramientas visuales y abstracciones de código para permitir a los usuarios crear aplicaciones de manera rápida y con menos necesidad de programación manual, acelerando el proceso de desarrollo y permitiendo a usuarios con menos experiencia técnica participar en la creación de aplicaciones.

Modelo de Datos Común de OMOP (*OMOP Common Data Model, OMOP CDM*): Estructura estandarizada de base de datos desarrollada por la colaboración Observational Medical Outcomes Partnership (OMOP), diseñada para representar datos de salud de manera uniforme y compatible, facilitando el análisis comparativo de datos clínicos y epidemiológicos provenientes de diferentes fuentes y sistemas de salud.

Observational Medical Outcomes Partnership (OMOP): Iniciativa colaborativa entre la industria, académicos y reguladores para mejorar la evaluación de medicamentos a través del análisis de datos de salud del mundo real. OMOP desarrolla métodos y estándares para el análisis de datos de salud, incluido el Modelo de Datos Común (CDM), que permite la armonización de datos para la investigación.

Omopizar: Proceso de transformar datos de salud de diferentes fuentes y formatos al Modelo de Datos Común de OMOP (CDM), para estandarizar la representación

de los datos y facilitar su análisis comparativo y la generación de evidencia científica en investigación clínica.

Observational Health Data Sciences and Informatics (OHDSI): Organización internacional que desarrolla y aplica métodos de análisis de datos de salud para generar evidencia a partir de datos del mundo real, con el objetivo de mejorar la toma de decisiones en salud pública y clínica, promoviendo el uso de estándares y herramientas abiertas para el intercambio y análisis de datos.

Salud digital (e-Salud): Utilización de tecnologías de la información y comunicación en el ámbito de la salud para mejorar la eficiencia, accesibilidad, calidad y seguridad de los servicios médicos, así como para fomentar la participación activa de los pacientes en su cuidado y la gestión de su salud.

Sistemas ciber-físicos (Cyber-Physical Systems, CPS): Sistemas que integran componentes físicos y computacionales, conectados a través de redes, para monitorear y controlar procesos físicos en tiempo real, utilizando tecnologías como sensores, actuadores, y sistemas de información y comunicación.

Sanidad 4.0 (Healthcare 4.0): También conocido como Salud 4.0, es la aplicación de tecnologías digitales como inteligencia artificial, Internet de las cosas y big data en el sector de la salud para mejorar la atención médica, la gestión de datos y la experiencia del paciente.

Tecnologías de la Información y Comunicación (TICs): Conjunto de herramientas, recursos y sistemas tecnológicos utilizados para adquirir, almacenar, procesar, transmitir y presentar información de manera digital, facilitando la comunicación y el intercambio de datos entre personas, organizaciones y dispositivos.

Telemedicina: Práctica médica que utiliza tecnologías de la información y comunicación para realizar consultas médicas, diagnósticos, tratamiento y seguimiento de pacientes a distancia, facilitando el acceso a la atención médica y la colaboración entre profesionales de la salud sin necesidad de encuentros físicos.