



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

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

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

---

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

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

# **Abstract**

---

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

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

# Índice general

---

<b>1 Descripción del Proyecto . . . . .</b>	<b>1</b>
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
<b>2 Objetivos del Proyecto . . . . .</b>	<b>9</b>
2.1. Objetivos del TFG . . . . .	9
2.2. Objetivos personales . . . . .	9
<b>3 Gestión del Proyecto . . . . .</b>	<b>11</b>
3.1. Participantes del proyecto . . . . .	11
3.2. Planificación temporal . . . . .	12
3.3. Planificación financiera . . . . .	14
3.4. Identificación de riesgos y planes de contingencia . . . . .	18
<b>4 Metodología . . . . .</b>	<b>20</b>
4.1. SofIA . . . . .	20
4.2. Scrum . . . . .	20
4.3. Control de versiones . . . . .	22
<b>5 Observational Health Data Sciences and Informatics (OHDSI) . . . . .</b>	<b>24</b>
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 . . . . .	31
5.4.3. Vías de implementación del análisis . . . . .	33
5.5. Conclusiones . . . . .	34
<b>6 Documento de Requisitos . . . . .</b>	<b>35</b>
6.1. Introducción . . . . .	35
6.2. Requisitos funcionales . . . . .	35
6.2.1. Diagrama de casos de uso . . . . .	35
6.2.2. Casos de uso del sistema . . . . .	36
6.3. Requisitos no funcionales . . . . .	48
6.4. Conclusiones . . . . .	50

<b>7 Entorno de Trabajo . . . . .</b>	<b>51</b>
7.1. Introducción . . . . .	51
7.2. Estándares de OHDSI . . . . .	51
7.2.1. Modelo de Datos Común de OMOP . . . . .	51
7.2.2. Vocabulario . . . . .	55
7.3. Herramientas de OHDSI . . . . .	56
7.3.1. ATLAS . . . . .	56
7.3.2. Otras herramientas . . . . .	61
7.4. Programas informáticos empleados . . . . .	62
7.5. Conclusiones . . . . .	64
<b>8 Arquitectura del Sistema . . . . .</b>	<b>65</b>
8.1. Introducción . . . . .	65
8.2. Arquitectura teórica del sistema . . . . .	66
8.3. Arquitectura de Broadsea . . . . .	67
8.4. Arquitectura de ATLAS Broadsea . . . . .	69
8.5. Conclusiones . . . . .	72
<b>9 Caso práctico . . . . .</b>	<b>73</b>
9.1. Introducción . . . . .	73
9.2. Estudio realizado por el HUVR . . . . .	73
9.3. Estandarización del estudio con ATLAS . . . . .	76
9.3.1. Datos . . . . .	77
9.3.2. Metodología . . . . .	79
9.3.3. Resultados . . . . .	83
9.4. Discusión de resultados . . . . .	83
9.5. Conclusiones . . . . .	83
<b>10 Resultados . . . . .</b>	<b>84</b>
10.1. Resultados . . . . .	84
10.2. Trazabilidad de objetivos . . . . .	84
10.3. Lecciones aprendidas . . . . .	84
<b>11 Conclusiones . . . . .</b>	<b>85</b>
<b>Bibliografía . . . . .</b>	<b>86</b>
<b>A Manual de ATLAS Broadsea . . . . .</b>	<b>90</b>
<b>B Glosario . . . . .</b>	<b>91</b>

# 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] . . . . .	30
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] . . . . .	32
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 . . . . .	36
6.2. Diagrama de actividad de RF-01:Añadir base de datos . . . . .	37
6.3. Diagrama de actividad de RF-02: Visualizar Reporte . . . . .	38
6.4. Diagrama de actividad de RF-03: Definir una cohorte . . . . .	40
6.5. Diagrama de actividad de RF-04: Definir un grupo de conceptos . . . . .	42
6.6. Diagrama de actividad de RF-05: Realizar Caracterización . . . . .	43
6.7. Diagrama de actividad de RF-06: Realizar caracterización . . . . .	45
6.8. Diagrama de actividad de RF-07: Realizar Predicción a nivel de Paciente	47
6.9. Diagrama de requisitos no funcionales . . . . .	48
7.1. Estructura del CDM v5.4. Extraída de la página de github del CDM [4]	53
7.2. Modelo Entidad-Relación del CDM v5.4. Extraída de la página de github del CDM [4] . . . . .	54
7.3. Captura de pantalla del menú principal de ATHENA . . . . .	56
7.4. Logo de ATLAS. Extraída del repositorio de github [5] . . . . .	57
7.5. Biblioteca de Métodos OHDSI. Extraída del Libro de OHDSI [3] . . . . .	58
7.6. Estructura de la WebAPI. Extraída de la wiki de github [6] . . . . .	59
7.7. Captura de pantalla del menú principal de ATLAS demo . . . . .	60
8.1. Esquema sencillo de Broadsea. Extraída de [7]. . . . .	65
8.2. Esquema de arquitectura <i>three-tier</i> en Docker. . . . .	66
8.3. Vista general de todos los componentes de Broadsea. Extraída de [7].	68
8.4. Captura de pantalla del menú principal de Broadsea . . . . .	68
8.5. Captura de pantalla del menú principal de ATLAS Broadsea . . . . .	70

## ÍNDICE DE FIGURAS

---

8.6. Captura de pantalla de pgAdmin de la estructura postgre del servidor de Broadsea . . . . .	71
9.1. Captura de pantalla de pgAdmin de la estructura del servidor del HUVR . . . . .	78
9.2. Captura de pantalla de pgAdmin del número de instancias de la tabla person . . . . .	79
9.3. Captura de pantalla de pgAdmin de la tabla source . . . . .	80
9.4. Captura de pantalla de pgAdmin de la tabla source_daimon . . . . .	80
9.5. Captura de pantalla de menú configuration de ATLAS Broadsea . . . . .	81
9.6. Captura de pantalla de la interfaz principal del menú Data Sources . . . . .	81
9.7. Captura de pantalla del reporte Dashboard generado en ATLAS Broadsea . . . . .	82

# 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 . . . . .	15
3.11. Coste real de personal del proyecto . . . . .	17
3.12. Posibles riesgos y planes de contingencia . . . . .	18
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 . . . . .	38
6.2. Caso de uso de RF-02:Visualizar Reporte . . . . .	39
6.3. Caso de uso de RF-03:Definir una cohorte . . . . .	41
6.4. Caso de uso de RF-04:Definir un grupo de conceptos . . . . .	42
6.5. Caso de uso de RF-05: Realizar caracterización . . . . .	44
6.6. Caso de uso de RF-06: Realizar Estimación a nivel de Población . . . . .	46
6.7. Caso de uso de RF-07: Realizar Predicción a nivel de Paciente . . . . .	48
6.8. RNF-01: Rendimiento . . . . .	49
6.9. RNF-02: Seguridad . . . . .	49
6.10. RNF-03: Usabilidad . . . . .	49
6.11. RNF-04: Portabilidad . . . . .	49
6.12. RNF-05: Interoperabilidad . . . . .	49
6.13. RNF-06: Mantenimiento . . . . .	50
 7.1. Dominios del CDM v5.4. Extraída del Libro de OHDSI [3] . . . . .	55
9.1. Recopilación de resultados del estudio del HUVR. Extraída de [8] . .	76

# **Índice de extractos de código**

---

9.1. Queries SQL para establecer la conexión con la base de datos del HUVR 80

# 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.

<b>Participante</b>	María del valle Alonso de Caso ortiz
<b>Institución</b>	Universidad de Sevilla
<b>Rol</b>	Jefe de Proyecto & Developer & Analista
<b>Información de contacto</b>	<a href="mailto:maraloort@alum.us.es">maraloort@alum.us.es</a>

**Tabla 3.1:** Descripción del primer participante del proyecto

<b>Participante</b>	Julián García García
<b>Institución</b>	Universidad de Sevilla
<b>Rol</b>	Tutor del TFG & Supervisor
<b>Información de contacto</b>	<a href="mailto:juliangg@us.es">juliangg@us.es</a>

**Tabla 3.2:** Descripción del segundo participante del proyecto

<b>Participante</b>	María José Escalona Cuaresma
<b>Institución</b>	Universidad de Sevilla
<b>Rol</b>	Tutor del TFG & Supervisor
<b>Información de contacto</b>	<a href="mailto:mjescalona@us.es">mjescalona@us.es</a>

**Tabla 3.3:** Descripción del tercer participante del proyecto

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

**Tabla 3.4:** Descripción del cuarto participante del proyecto

<b>Participante</b>	Carlos Luis Parra Calderón
<b>Institución</b>	Hospital Universitario Virgen del Rocío
<b>Rol</b>	Supervisor de prácticas en empresa
<b>Información de contacto</b>	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. 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 aunque todos los sprint guardan la misma duración (cuatro semanas), 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 sprint se estima una carga de trabajo menor (más ligada a la investigación) mientras que en los últimos sprint, 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.

Sprint 1			
Inicio:	Fin:	Esfuerzo estimado:	Esfuerzo real:
<b>Resumen</b>	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:
<b>Resumen</b>	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
<b>Resumen</b>			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
<b>Resumen</b>			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

Las horas totales de trabajo planificado fueron 300 horas, según los créditos asignados a la asignatura. No obstante, la dedicación real al proyecto ha sido superior a lo previsto, dedicándole **325 horas reales**, es decir, 25 horas extra suponiendo una **desviación aproximadamente del 8% sobre lo planificado**. Esta desviación no se considera un impedimento o anormalidad dentro de los riesgos previstos del proyecto, lo que no ha supuesto ningún impedimento mayor para la finalización del trabajo.

### 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, developer 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 developer 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 *per se* 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.
- **Developer.** 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
<b>Jefe de Proyecto</b>	39.16 €/h	105 h	4111.80 €
<b>Developer</b>	35.18 €/h	115 h	4045.70 €
<b>Analista</b>	33.12€/h	80 h	2649.60 €
<b>Coste total:</b>	<b>10807.10€</b>		

**Tabla 3.10:** Coste estimado de personal del proyecto

Por tanto, el **coste estimado de personal del proyecto es 10807.10€**.

## 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

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

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.**

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

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 11543.60 €.**

$$10807,10 \text{ €} + 729,50 \text{ €} = 11536,60 \text{ €} \quad (3.6)$$

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 13274,09 € de beneficio total del proyecto.

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 1154,36 €.

### Desviaciones

El proyecto ha sufrido una desviación del 8 % sobre el tiempo en horas planificado para su conclusión (véase 3.2 "Planificación temporal"). Esta desviación influye en los costes materiales, necesitándose recalcular algunos costes estimados, como el coste de personal y el de suministro eléctrico.

En cuanto al coste de personal, el coste real ascendería a 11685.90€.

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 €
<b>Coste total:</b>			<b>11685.90€</b>

**Tabla 3.11:** Coste real de personal del proyecto

En cuanto al coste de suministro de eléctrico, el coste ascendería de forma insignificante a 17,75 €, contribuyendo a coste material de 737.87 €.

$$0,182 \text{ €/kWh} \times 0,3 \text{ kWh} \times 325 \text{ h} = 17,75 \text{ €.} \quad (3.7)$$

Por tanto, sumando las desviaciones en costes de personal y material, **el coste total real del proyecto resultaría en 12423.77 €.**

$$11685,90\text{€} + 737,87\text{€} = 12423,77\text{€} \quad (3.8)$$

El coste final del proyecto supone una **desviación de 880.17 €** sobre el coste estimado del proyecto, que correspondía a 11543.60 €. 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 (1154,36). 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

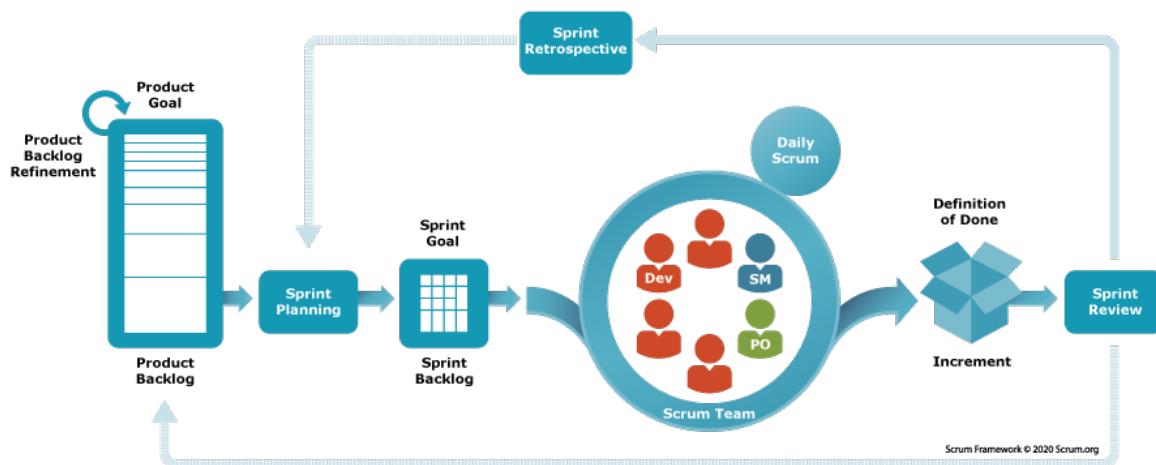
estas nuevas metodologías disruptivas. Los motivos y beneficios de ello son muy numerosos, las metodologías ágiles abogan por el cambio continuo y la adaptabilidad frente a la rigurosidad tradicional.

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].

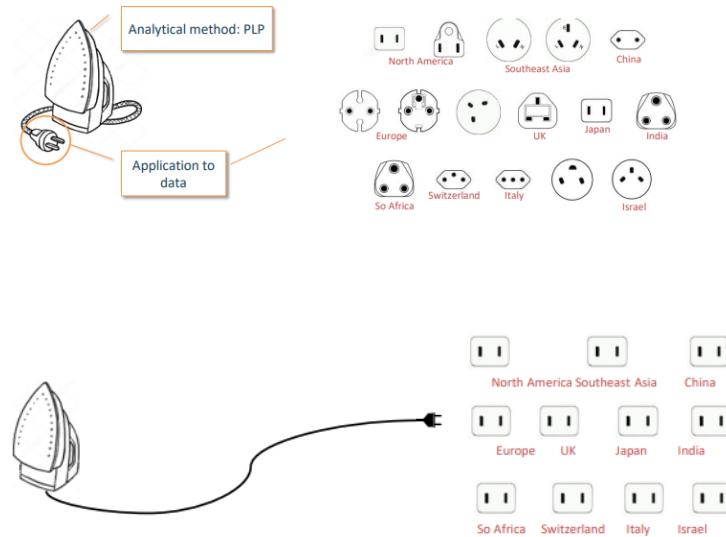
Otras vías de divulgación son publicaciones científicas [33], tutoriales para principiantes, grabaciones de las reuniones semanales de la comunidad o las conferencias anuales a través de su canal de youtube [34]; canales de mensajería abierta como discord [35] o MS Teams [36], cientos de repositorios 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

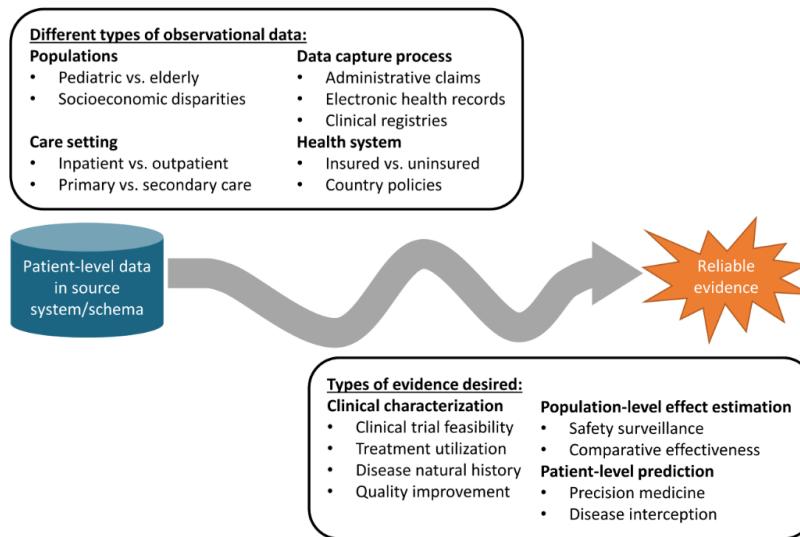
demostraron la viabilidad de establecer un CDM que además reuniese diferentes vocabularios estandarizados, reuniendo en un mismo estándar diversos tipos de datos de diferentes entornos de atención y representados por diferentes vocabularios de origen. Esta característica facilitó la colaboración y aumentó 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"](#).

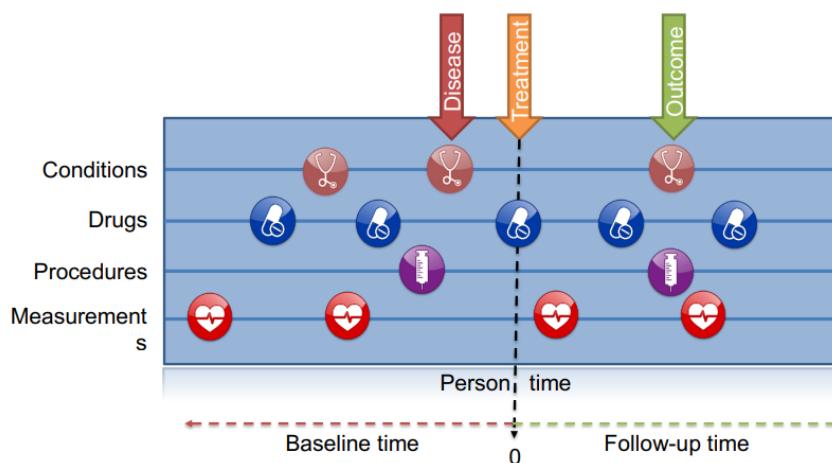
Con ello la organización define un marco para llevar a cabo **estudios observacionales o fenotípicos** sobre datos. Un estudio observacional es una investigación que observa y recopila información sobre individuos o fenómenos sin intervenir en ellos. En el caso del estudio sobre datos, en vez de realizar 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 estudio de cohortes en la base de datos. 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 la cohorte.**

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 periodo de observación (*End of the observation period*, en verde).

Dentro del periodo 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).

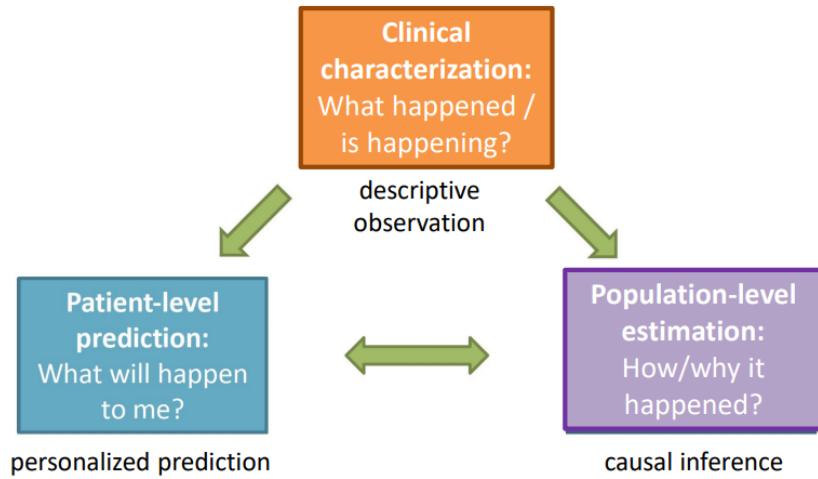
- **Evento de entrada.** Define el evento que cualifica al paciente para entrar a la cohorte. El conjunto de pacientes que satisfacen el evento de entrada conforman la cohorte inicial.
- **Evento de salida.** Define el evento de salida de la cohorte, cuando el paciente ya no es elegible para formar parte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

#### 5.4.2. Casos de uso para la investigación

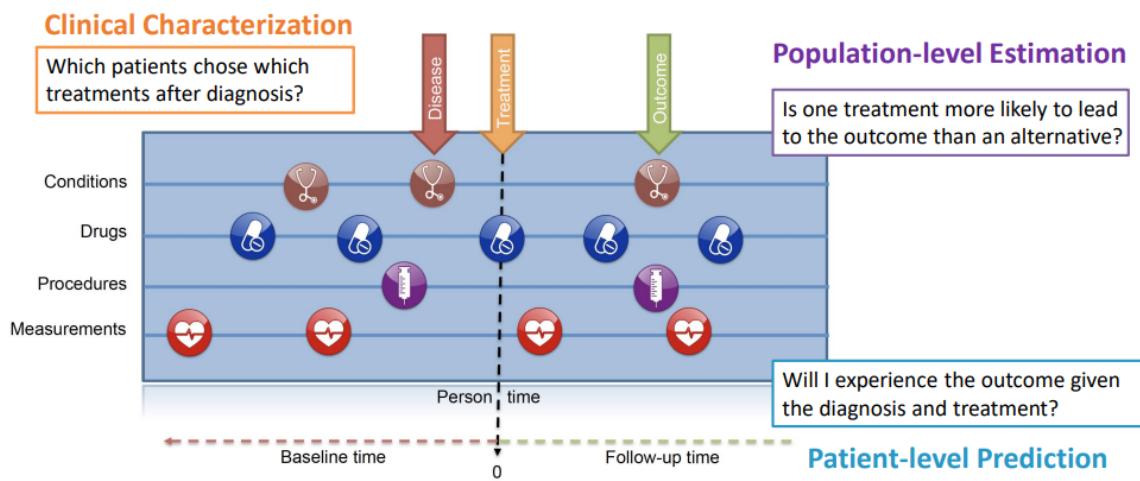
Con el fin de estandarizar y proveer un marco metodológico en el camino hacia la generación de evidencia, OHDSI define tres casos de usos que establecen los

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

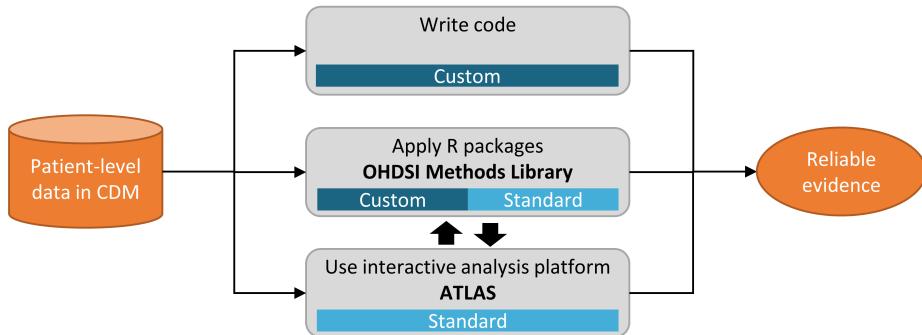
La predicción a nivel de paciente busca, en base a los datos obtenidos de los conjuntos de pacientes en la base de datos, realizar predicciones concretas para individuos concreto.

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 customizada, 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 dificulten 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.

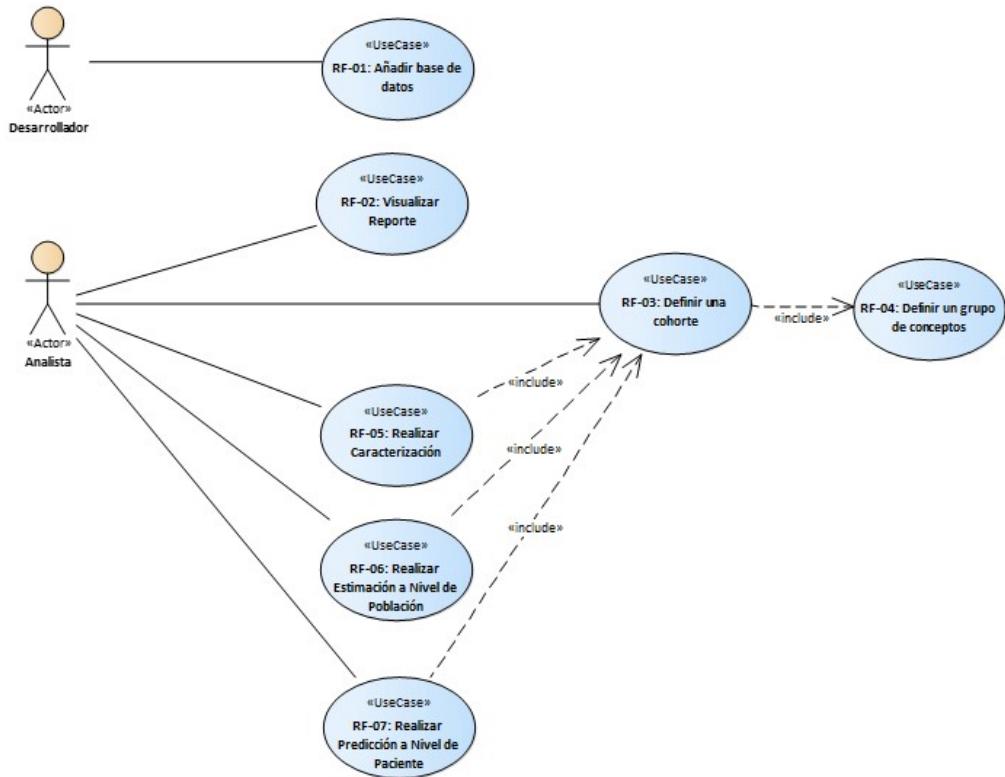
Para el sistema de ATLAS Broadsea se han definido seis requisitos funcionales y dos actores o usuarios del sistema: el desarrollador y el analista de datos. Los requisitos funcionales hacen referencia a las tareas que puede realizar el usuario a la hora de conducir un análisis utilizando el sistema. 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"

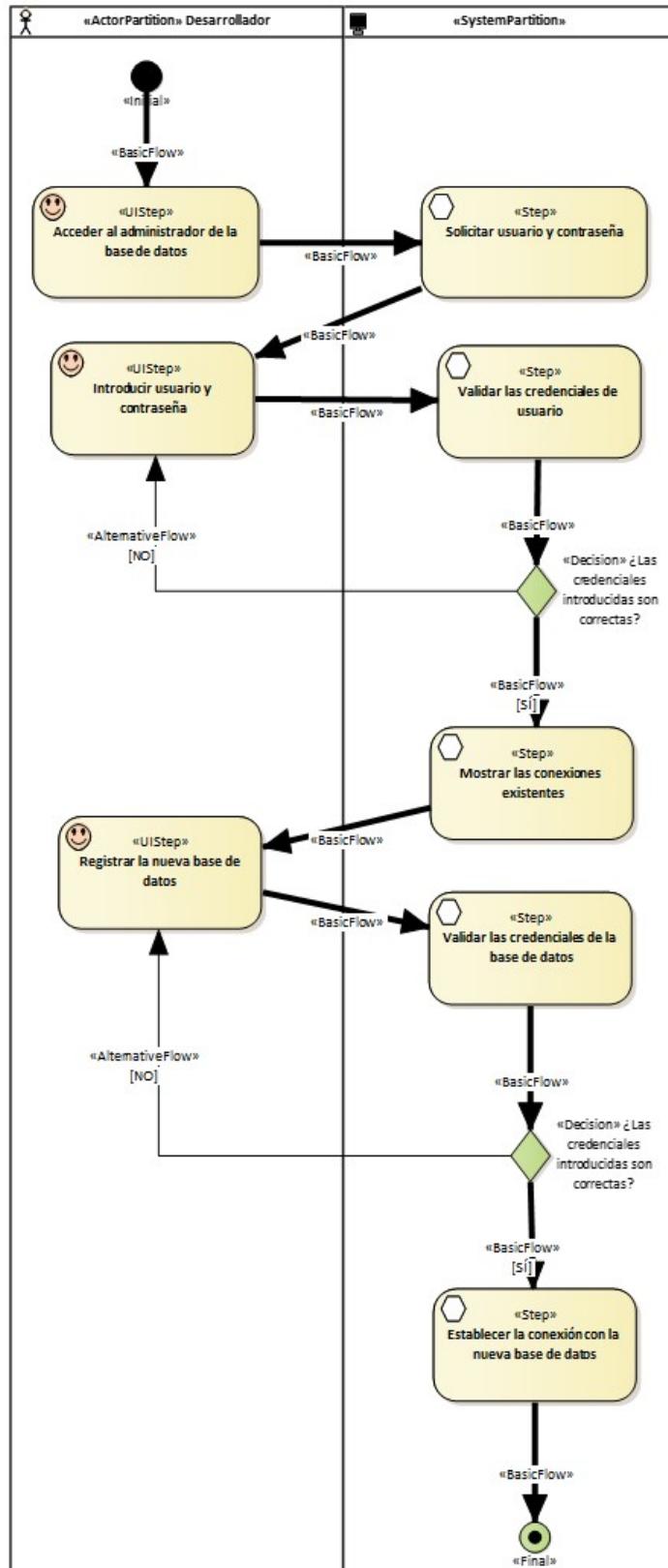


Figura 6.2: Diagrama de actividad de RF-01:Añadir base de datos

«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	<b>Pre-condición:</b> La base de datos debe estar estandarizada al CDM de OMOP. <b>Post-condición:</b> 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

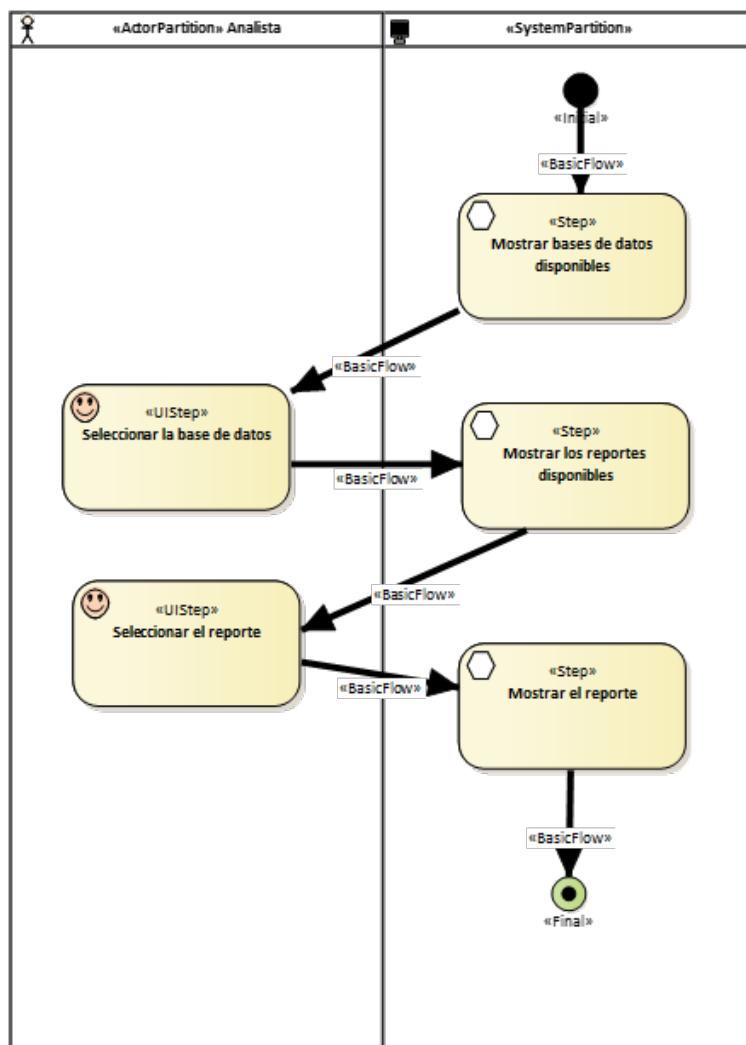


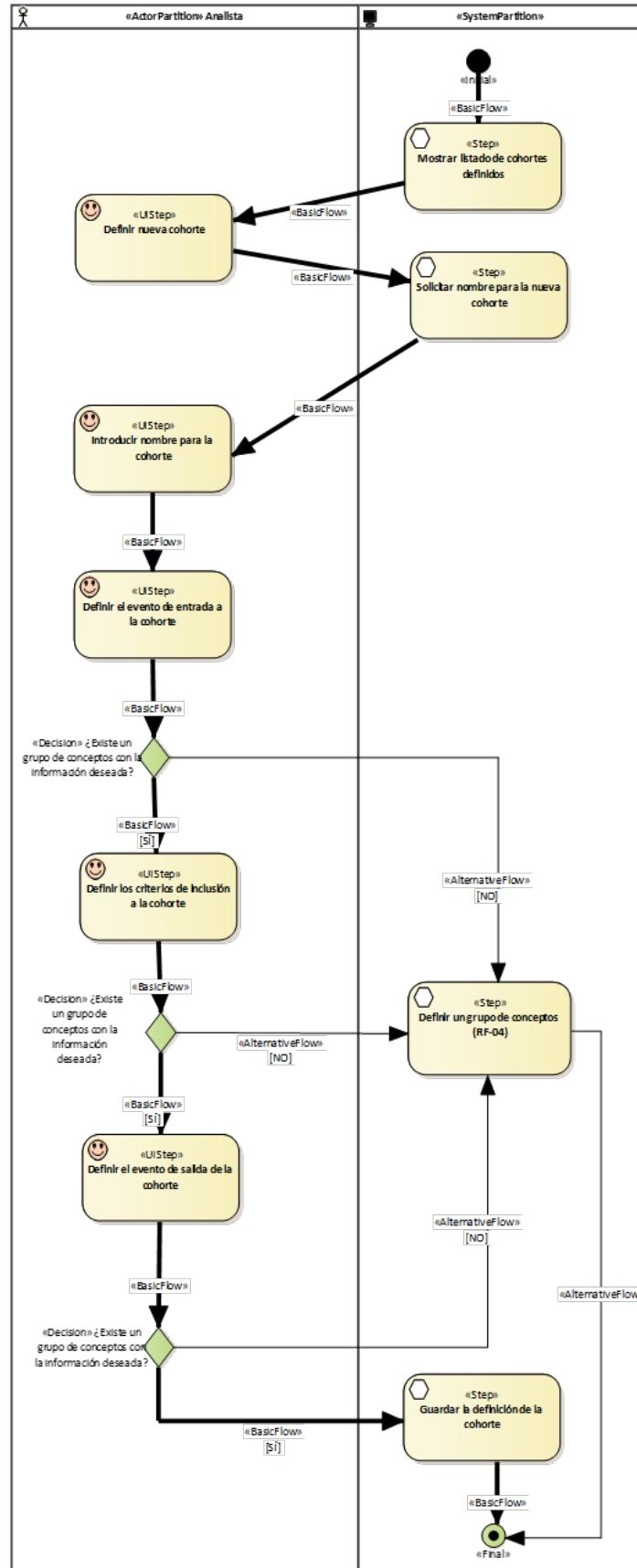
Figura 6.3: Diagrama de actividad de RF-02: Visualizar Reporte

## 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.		
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	<b>Pre-condición:</b> La base de datos debe estar estandarizada al CDM de OMOP. <b>Post-condición:</b> La base de datos debe quedar registrada en el sistema.		
Estado	Implemented		

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



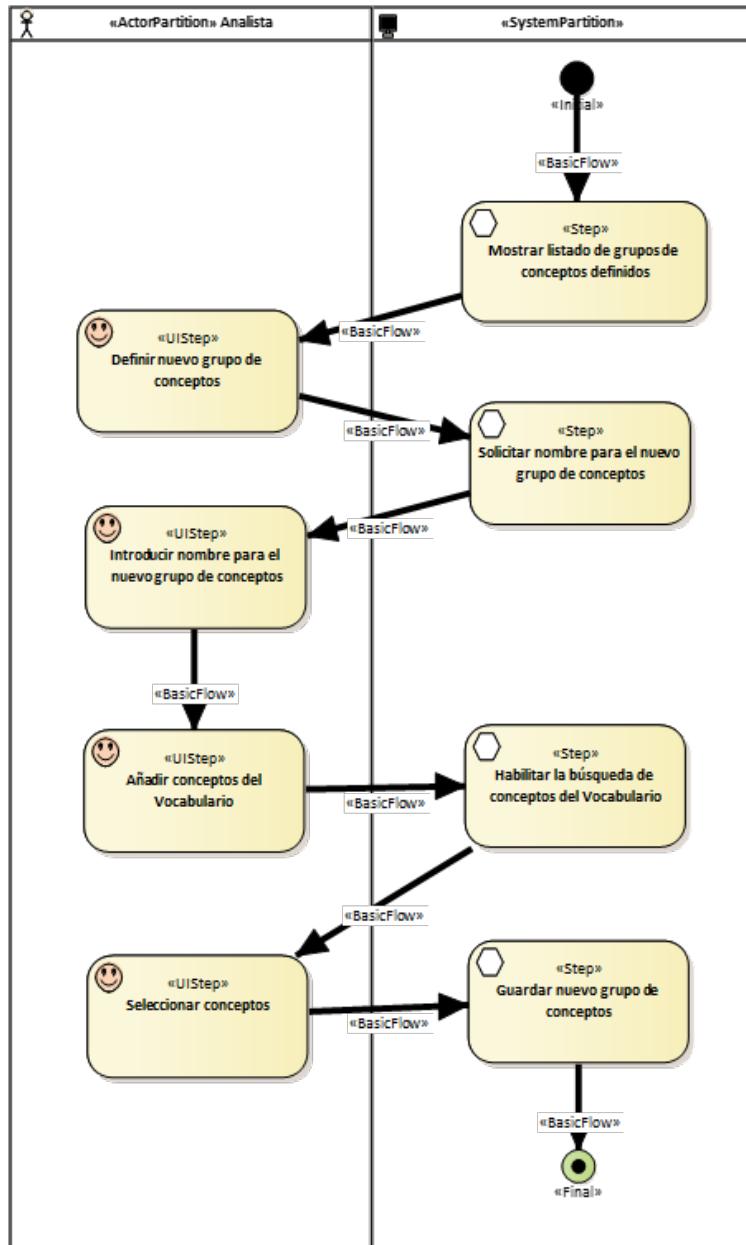
**Figura 6.4:** Diagrama de actividad de RF-03: Definir una cohorte

## CAPÍTULO 6. DOCUMENTO DE REQUISITOS

---

«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	<b>Pre-condición:</b> Puede haber definido un grupo de conceptos que describa las características de la cohorte <b>Post-condición:</b> 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 RF-03: Definir una cohorte	Estereotipo «include»	Destino 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	<b>Pre-condición:</b> No procede <b>Post-condición:</b> 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

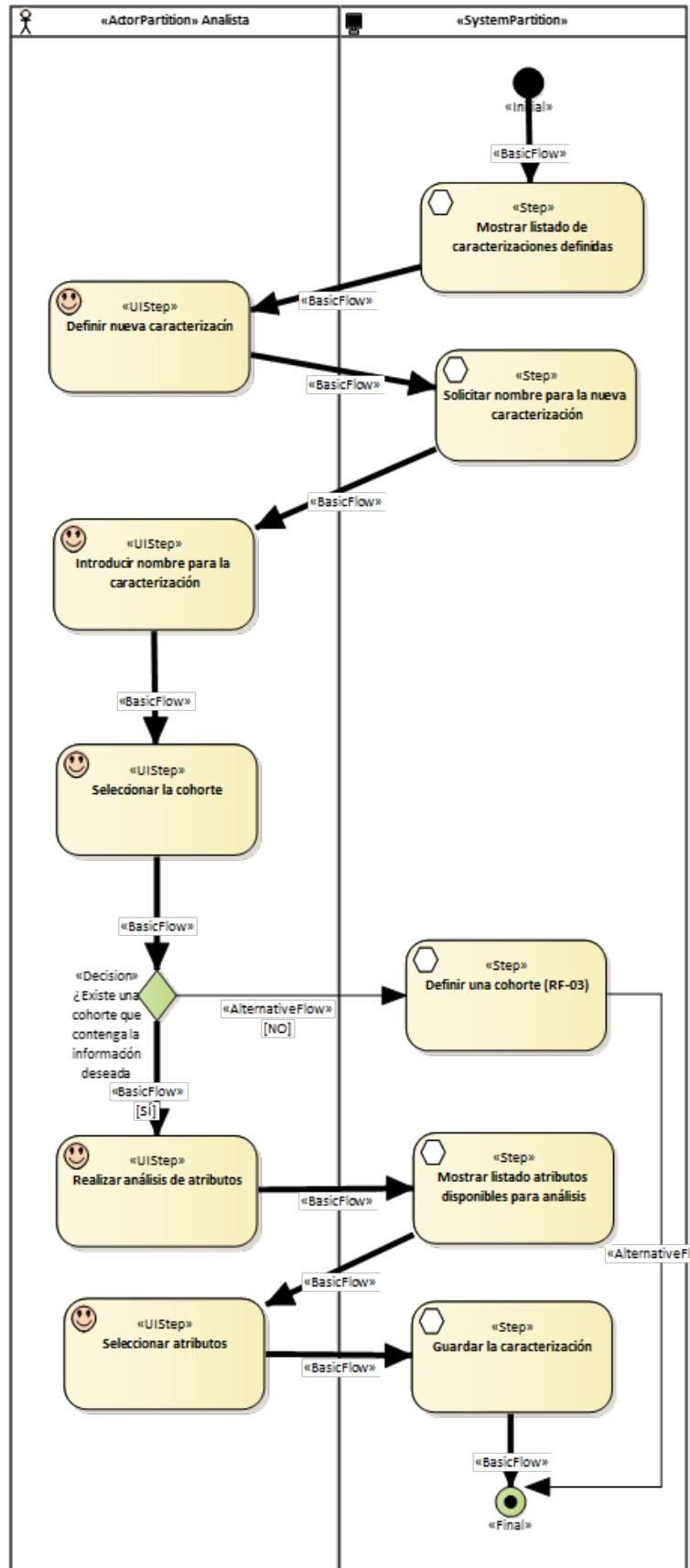


Figura 6.6: Diagrama de actividad de RF-05: Realizar Caracterización

## 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		
	<b>Pre-condición:</b> Puede haber una cohorte registrada que describa las características poblaciones del estudio <b>Post-condición:</b> La caracterización debe quedar registrada en el sistema		
Estado	Implemented		

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

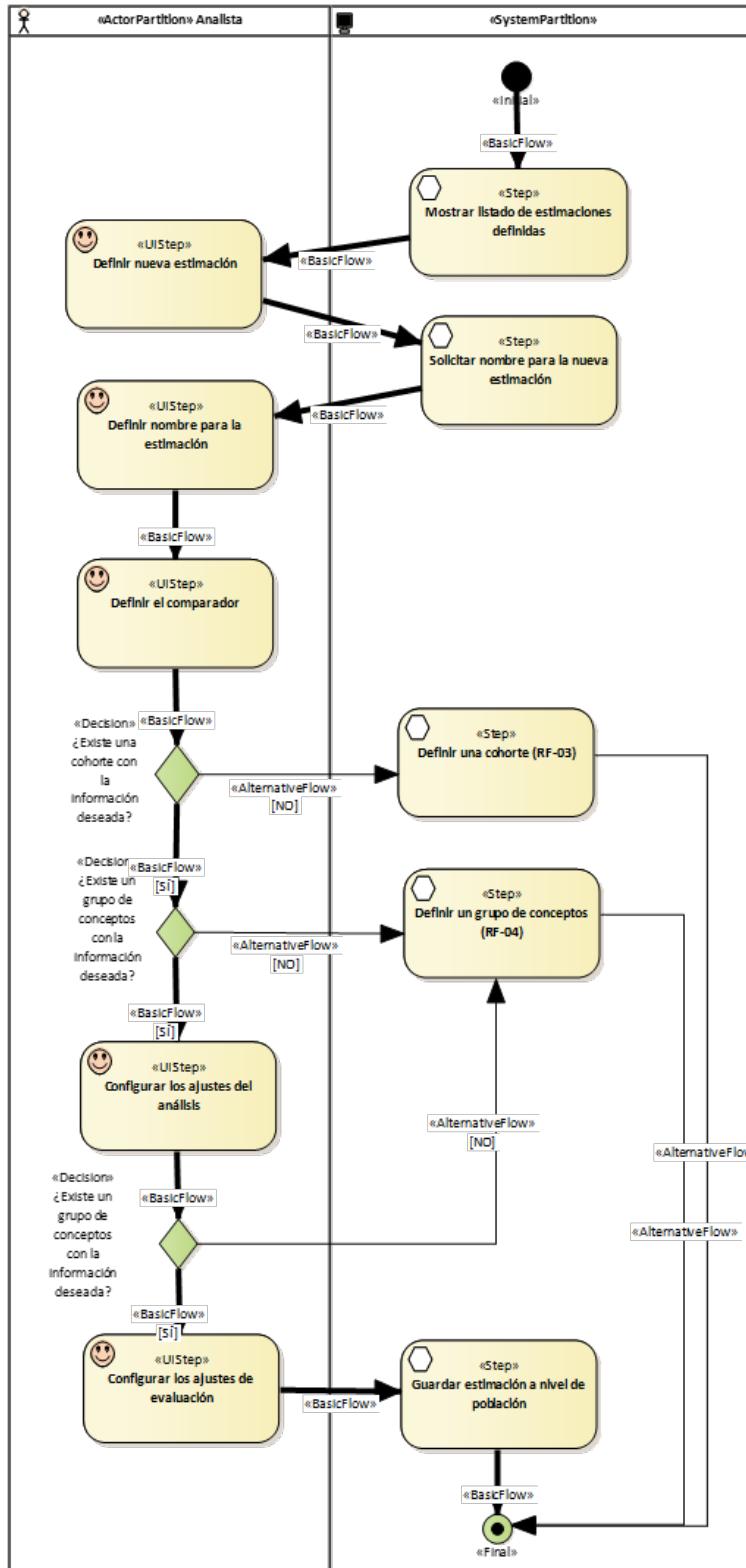
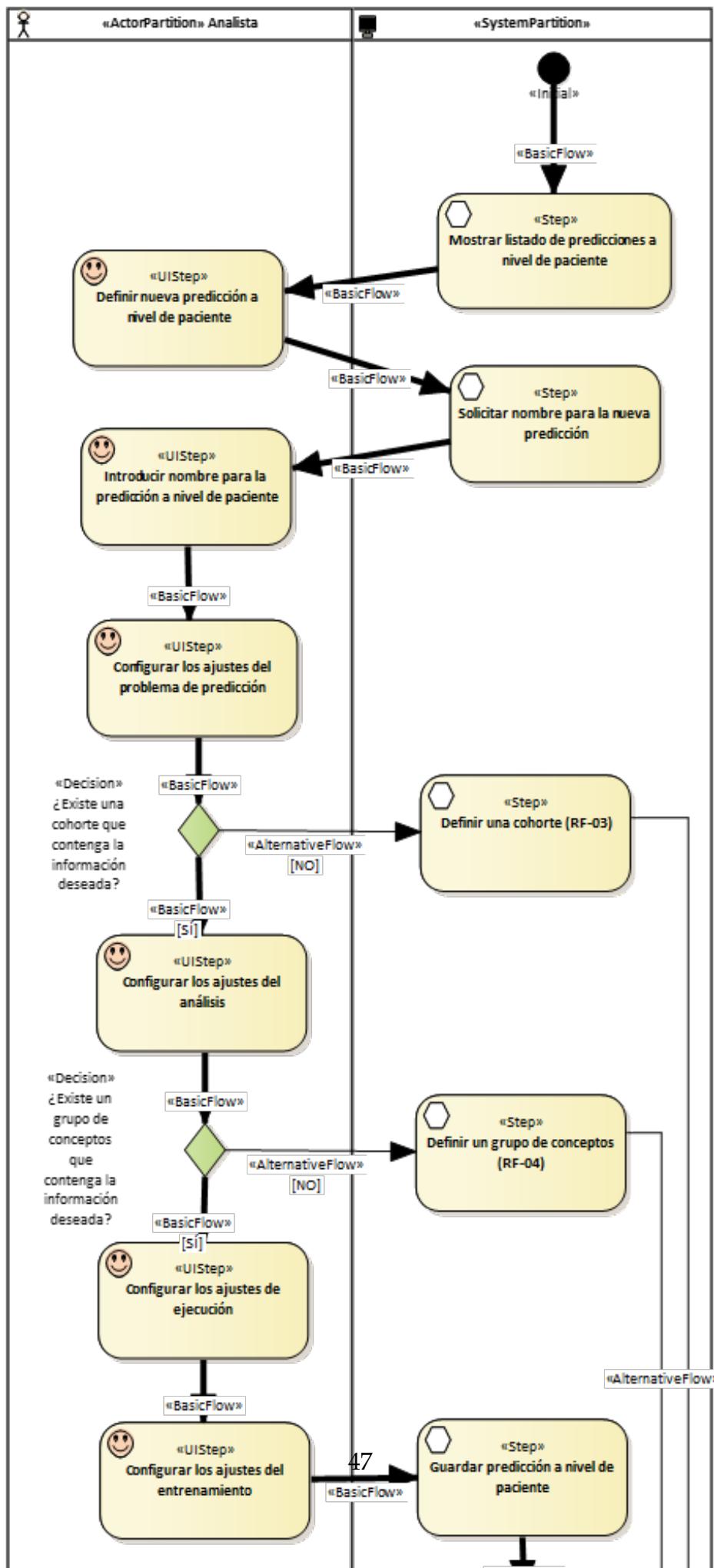


Figura 6.7: Diagrama de actividad de RF-06: Realizar caracterización

«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		
	<b>Pre-condición:</b> Puede haber una cohorte registrada que describa las características poblaciones del estudio <b>Post-condición:</b> 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
Analista		«use»	RF-07: Realizar Predicción a Nivel de Paciente
Descripción	Realizar un estudio de Predicción a Nivel de Paciente sobre los efectos que experimentará un individuo concreto		
Pre-condición y Post-condición	<b>Pre-condición:</b> Puede haber una cohorte registrada que describa las características poblaciones del estudio <b>Post-condición:</b> 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 de 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.

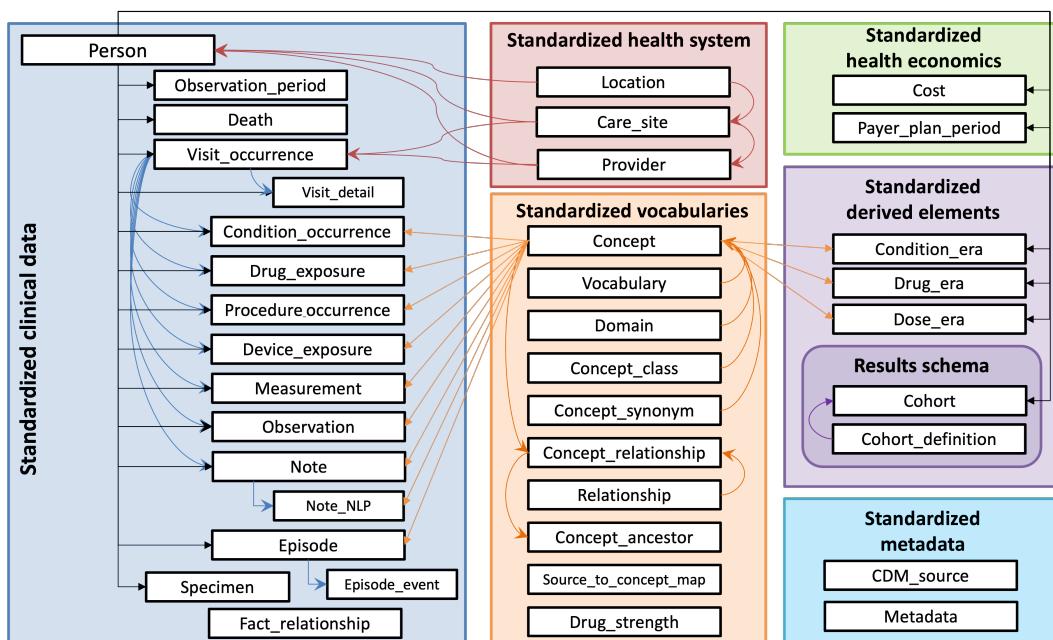
## Modelo de Datos Lógico

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]

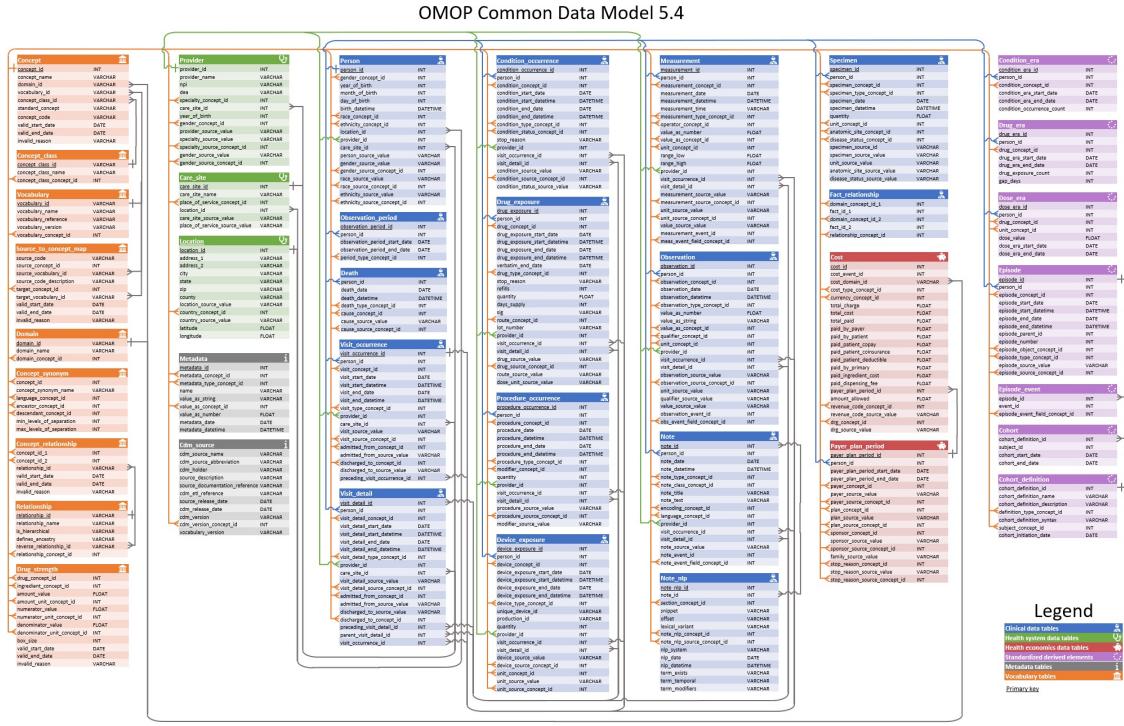


Figura 7.2: Modelo Entidad-Relación del CDM v5.4. Extraída de la página de github del CDM [4]

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?”](#)).

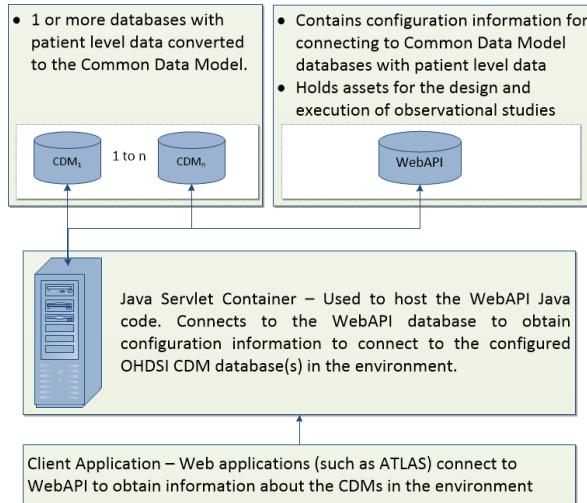
- III. **Reciclabilidad.** Por último, otro beneficio es que gracias a estas características ATLAS permite diseñar estructuras para el estudio de los datos que puedan utilizarse en diferentes bases de datos distintas. Volviendo al ejemplo de la plancha en [5.2 “¿Qué es OHDSI?”](#), esto quiere decir 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.

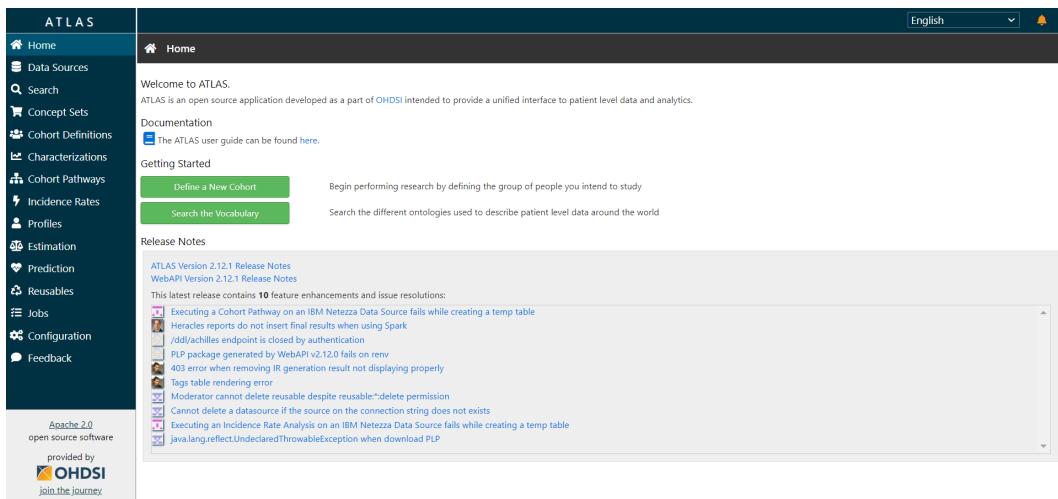
Sin embargo, es importante recalcar que tanto ATLAS como la mayoría de las herramientas de OHDSI no consiste en un archivo ejecutable aislado sino en una aplicación contenida y dependiente de un ecosistema completo basado en web. La dependencia principal y red que sostiene a ATLAS es la **WebAPI**.



**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 estrictamente 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**.



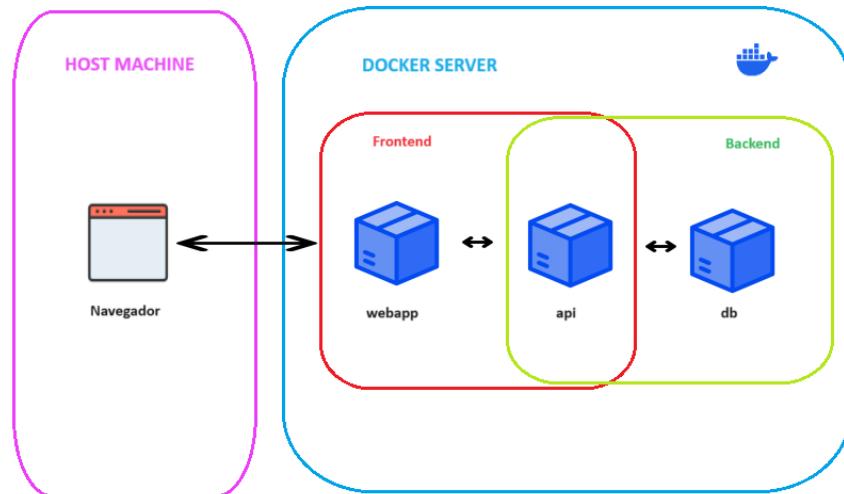
**Figura 8.1:** Esquema sencillo de Broadsea. Extraída de [7].

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.

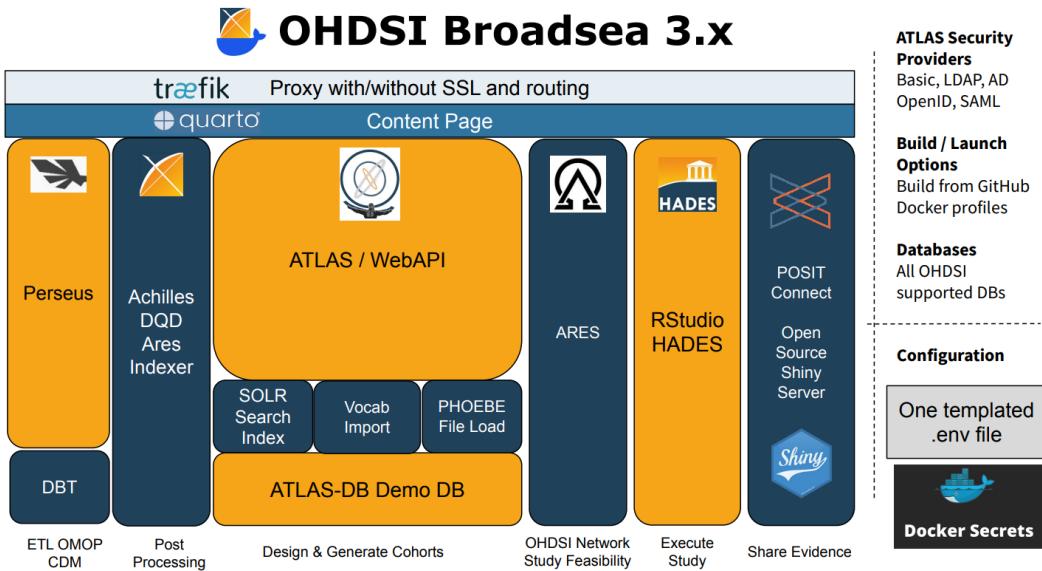


Figura 8.3: Vista general de todos los componentes de Broadsea. Extraída de [7].

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.

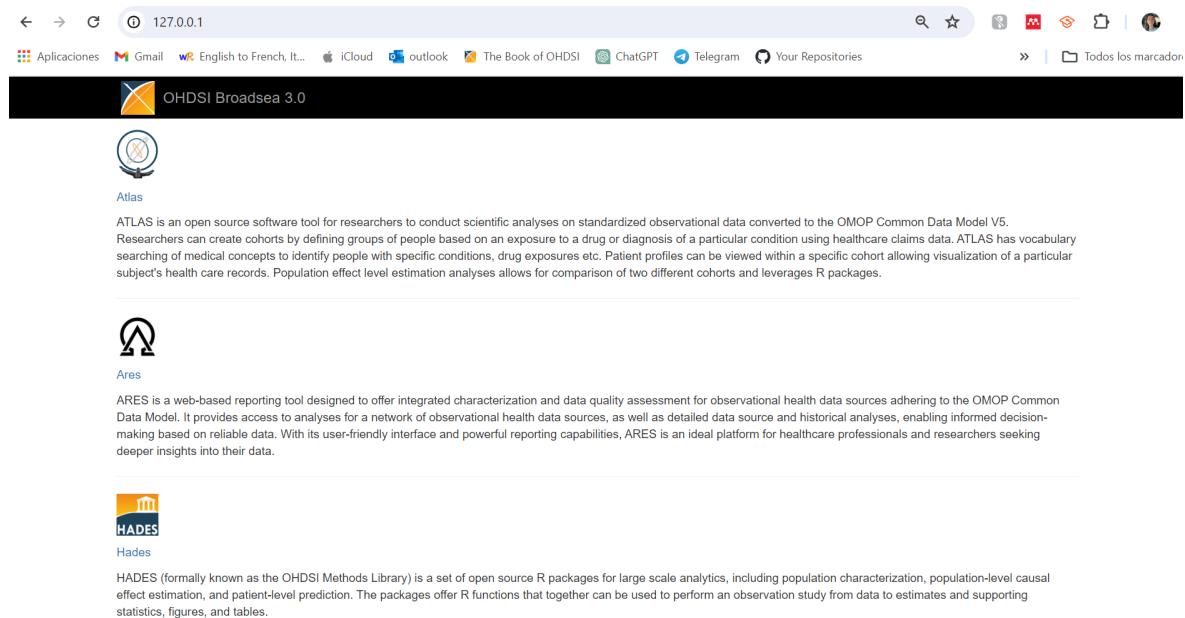


Figura 8.4: Captura de pantalla del menú principal de Broadsea

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”).

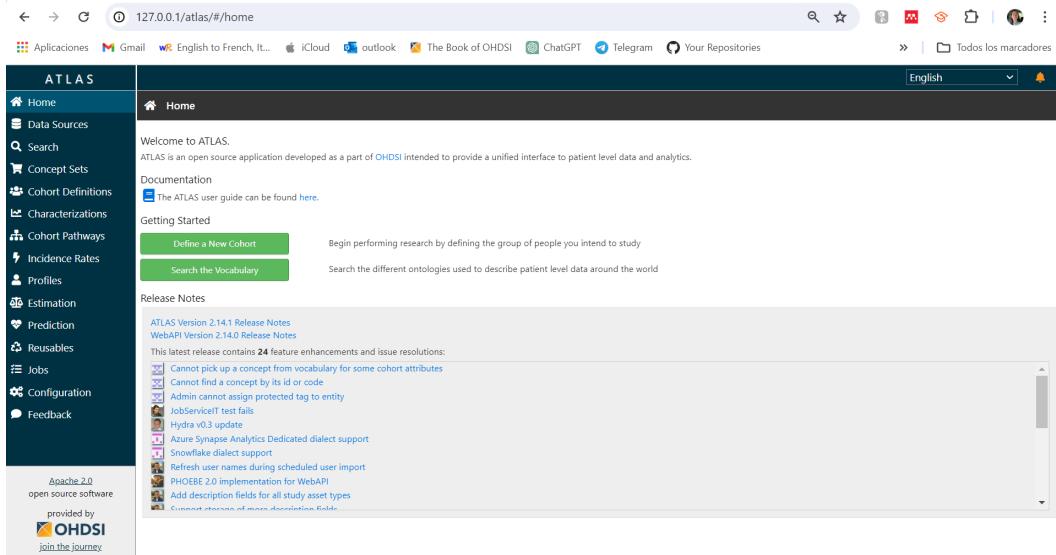
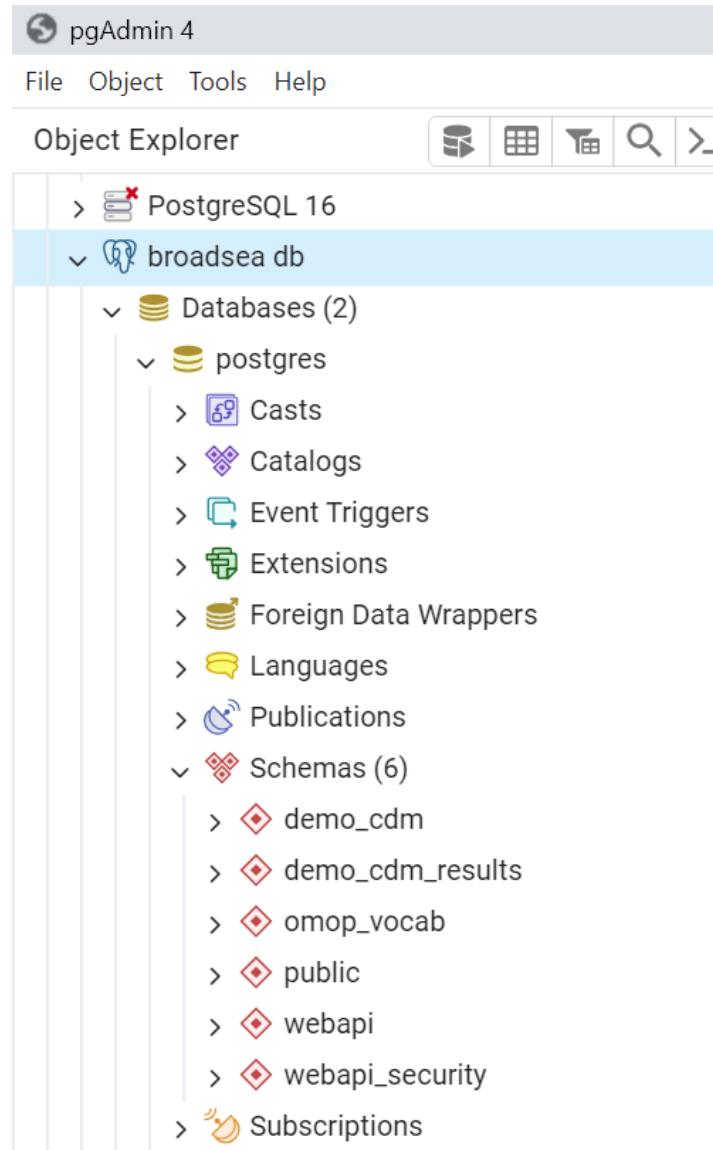


Figura 8.5: Captura de pantalla del menú principal de ATLAS Broadsea

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 accesibles *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

- RWHD
- 875 pacientes del registro s31 y observación del huVR??

## 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	$\chi^2$ + 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 finalmente 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**, en este caso mediante el Modelo de Datos Común de OMOP y la herramienta de análisis de datos ATLAS. Es decir, no se pretende meramente reproducir el estudio haciendo uso del ecosistema de OHDSI sino que se destaca que el fin último del proceso es estandarizarlo, adaptarlo al marco de investigación OHDSI para que cualquier nodo de la organización pudiera procesarlo, analizarlo y reproducirlo fácilmente.

**El estudio realizado por el HUVR en el marco de OHDSI corresponde al caso de uso de investigación 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 un paciente en base a ciertas características de experimentar un efecto concreto.

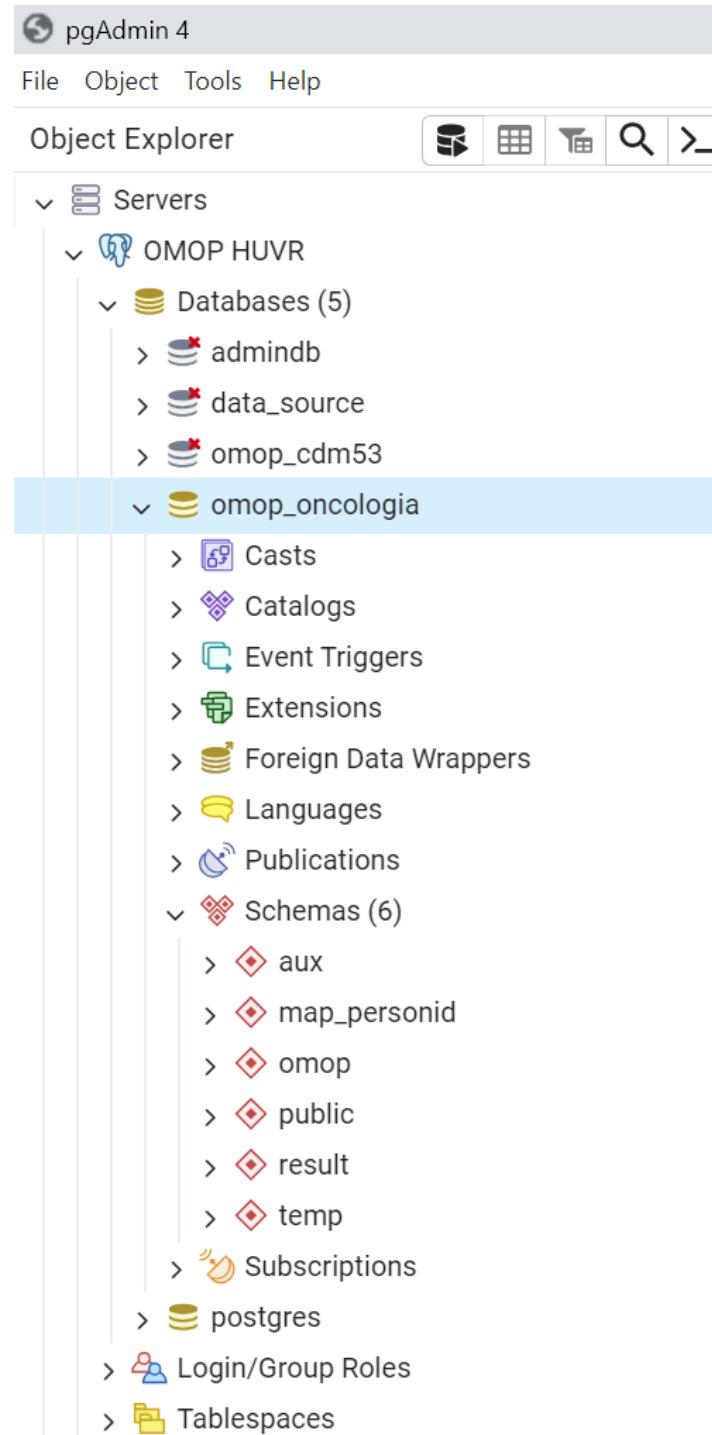
No obstante, para poder aplicar los otros dos casos de uso estudiados (Caracterización y Estimación a Nivel de Población) en el análisis con ATLAS, se ha adaptado o reinterpretado el estudio cuando fuere necesario para que tengan sentido.

Por ello el fin del caso práctico no es una reproducción fiel del estudio con ATLAS sino permitir la reinterpretación del estudio que permita estandarizarlo a la metodología OHDSI.

### **9.3.1. Datos**

En cuanto a los datos empleados, tras la aprobación del comité de ética para el uso secundario de los datos, el HUVR ha facilitado para la realización del proyecto el mismo dataset utilizado en el estudio original.

No obstante, la estructura de la base de datos del estudio original no correspondía con el Modelo de Datos de OMOP por lo que una tarea crucial para su utilización ha sido el proceso de omopizado de la base de datos y la comprobación de calidad del proceso de ETL, 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".



**Figura 9.1:** Captura de pantalla de pgAdmin de la estructura del servidor del HUVR

Una vez que la base de datos se preparó al CDM se me proporcionó el acceso al servidor del hospital donde se aloja la base de datos, tal y como se muestra en la Figura 9.1 "Captura de pantalla de pgAdmin de la estructura del servidor del HUVR".

La base de datos utilizada es `omop_oncologia`. Esta base de datos presenta seis esquemas de los cuales `omop`, `result` y `temp` son relevantes para ATLAS.

- **omop.** Este esquema almacena todas las tablas con la toda la información omopizada de los pacientes y eventos clínicos de la base de datos.
- **result.** Este esquema almacena los resultados de ejecutar ACHILLES (véase [7.3 "Herramientas de OHDSI"](#)) sobre la base de datos. Es crucial para la correcta integración de la base de datos con ATLAS.
- **temp.** Este esquema almacena información temporal durante las ejecuciones de ACHILLES.

La base de datos contiene 1332 instancias de pacientes registradas, como se muestra en la Figura "Captura de pantalla de pgAdmin del número de instancias de la tabla person".

>	person	7	7	8532	1965	10	31	1965-10-31 00:00:00
>	procedure_occurrence	8	8	8507	1950	9	24	1950-09-24 00:00:00
>	provider	9	9	8532	1949	4	10	1949-04-10 00:00:00
>	relationship	10	10	8507	1939	11	7	1939-11-07 00:00:00
>	source_to_concept_map	11	11	8507	1951	1	19	1951-01-19 00:00:00
>	specimen	12	12	8507	104	4	29	1044-04-29 00:00:00
>	visit_detail							
Total rows: 1000 of 1332 Query complete 00:00:05.823								
Ln 1, Col 1								

**Figura 9.2:** Captura de pantalla de pgAdmin del número de instancias de la tabla person

### 9.3.2. Metodología

En cuanto a la metodología del estudio, se realiza el análisis de datos a través de la herramienta ATLAS del sistema Broadsea, instalada localmente en el ordenador de la alumna (proceso descrito en Anexo [A "Manual de instalación, despliegue y configuración de ATLAS Broadsea"](#)).

Las tareas que se realizan a continuación corresponden con los requisitos funcionales del sistema, descritos en [6 "Documento de Requisitos"](#).

#### 9.3.2.1 Conexión con la base de datos del HUVR

El proceso de integración de una base de datos externa en la WebAPI de Broadsea se describe con gran detalle en el Anexo [A "Manual de instalación, despliegue y configuración de ATLAS Broadsea"](#). No obstante, en esta subsección se presenta de forma sencilla el proceso de conexión con la base de datos del HUVR.

Para establecer la conexión con una base de datos externa se debe registrar la base de datos en el esquema de la WebAPI de Broadsea, concretamente en las tablas source y source\_daimon. Para ello se ejecutan las siguientes *queries*:

Estas queries también se encuentran en el archivo del repositorio de github Thesis-ATLAS-OHDSI/environment/thesis/sql/queries\_insert\_huvr\_source.sql.

En las Figuras [9.3 "Captura de pantalla de pgAdmin de la tabla source"](#) y [9.4 "Captura de pantalla de pgAdmin de la tabla source\\_daimon"](#) se muestran los resultados de la ejecución correcta de las queries.

```

1 -- los datos del string jdbc de la base de datos no se muestran al
   completo por confidencialidad
2
3 INSERT INTO webapi.source (source_id, source_name, source_key,
   source_connection, source_dialect)
4 SELECT 3, 'S31 Registry HUVR', 'S31HUVR', 'jdbc:postgresql://servidor
   :5432/omop_oncologia?user=usuario&password=contraseña', 'postgresql';
5
6 INSERT INTO webapi.source_daimon (source_daimon_id, source_id,
   daimon_type, table_qualifier, priority)
7 SELECT nextval('webapi.source_daimon_sequence'), 3, 0, 'omop', 1
8 FROM webapi.source
9 WHERE source_key = 'S31HUVR';
10
11 INSERT INTO webapi.source_daimon (source_daimon_id, source_id,
   daimon_type, table_qualifier, priority)
12 SELECT nextval('webapi.source_daimon_sequence'), 3, 2, 'result', 0
13 FROM webapi.source
14 WHERE source_key = 'S31HUVR';

```

**Extracto de código 9.1:** Queries SQL para establecer la conexión con la base de datos del HUVR

	source_id	source_name	source_key	source_connection
1	1	OHDSI Eunomia Demo Database	EUNOMIA	jdbc:postgresql://broadsea-atlasdb:5432/postgres?user=postgres&password=mypass
2	3	S31 Registry HUVR	S31HUVR	jdbc:postgresql://[REDACTED]:5432/omop_oncologia?user=[REDACTED]&password=[REDACTED]

**Figura 9.3:** Captura de pantalla de pgAdmin de la tabla source

	source_daimon_id	source_id	daimon_type	table_qualifier	priority
1	1	1	0	demo_cdm	0
2	2	1	1	omop_vocab	10
3	3	1	2	demo_cdm_results	0
4	7	3	0	omop	1
5	9	3	2	result	0
6	10	3	5	temp	0

**Figura 9.4:** Captura de pantalla de pgAdmin de la tabla source\_daimon

Se puede comprobar que la integración de la base de datos en Broadsea se ha

realizado correctamente a través del menú configuration de ATLAS, donde se muestran los detalles de las bases de datos integradas con la herramienta. En la Figura 9.5 "Captura de pantalla de menú configuration de ATLAS Broadsea" se observa que hay dos bases de datos: la base de datos de Eunomia que viene preinstalada con Broadsea y la base de datos recién instalada.

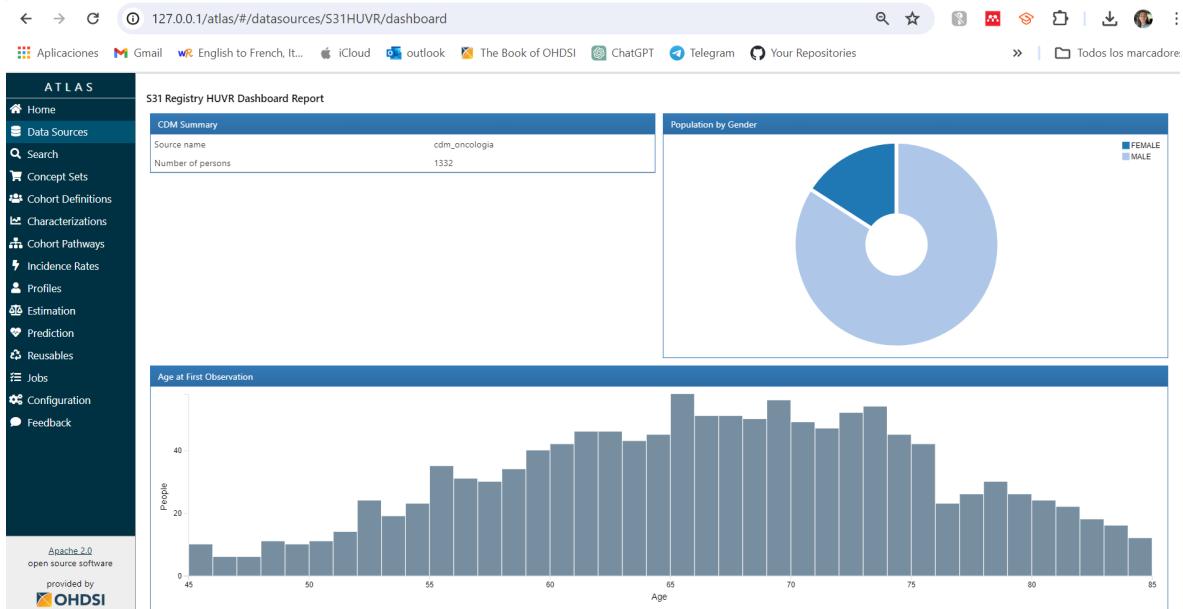
**Figura 9.5:** Captura de pantalla de menú configuration de ATLAS Broadsea

### 9.3.2.2 Reporte de la base de datos

La obtención de un reporte de la base de datos utilizando ATLAS es un proceso muy sencillo. Accediendo al menú lateral Data Sources se abre una interfaz que permite seleccionar la base de datos de la que se quiere obtener el reporte y el tipo de reporte que se quiere obtener.

**Figura 9.6:** Captura de pantalla de la interfaz principal del menú Data Sources

A continuación se generan algunos reportes para realizar un análisis exploratorio de la base de datos. Todas las imágenes que se obtienen en cada reporte se encuentran en el repositorio de github Thesis-ATLAS-OHDSI/environment/thesis/atlas/data sources/.



**Figura 9.7:** Captura de pantalla del reporte Dashboard generado en ATLAS Broadsea

### 9.3.2.3 Definición de grupos de conceptos

### 9.3.2.4 Definición de cohortes

Necesario para utilizar ATLAS

Cohorte = Personas con cancer de pulmon

### 9.3.2.5 Caracterización

Para conocer los pacientes que tenemos en el cohorte

### 9.3.2.6 Estimación a nivel de población

Hacer un estudio sobre los efectos adversos que sufrirá la población del cohorte

### 9.3.2.7 Predicción a nivel de Paciente

Para un paciente

### **9.3.3. Resultados**

## **9.4. Discusión de resultados**

Comparación de los resultados obtenidos en el estudio del HUVR y el estudio realizado con ATLAS

## **9.5. Conclusiones**

En este capítulo se concluye que...

# **10. Resultados**

---

## **10.1. Resultados**

- Muchos problemas

## **10.2. Trazabilidad de objetivos**

Trazabilidad de objetivos con resultados

## **10.3. Lecciones aprendidas**

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

...

# **11. Conclusiones**

---

# 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](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/](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/documentation/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.