
Enriching information extraction pipelines in clinical decision support systems

João Rafael Duarte de Almeida

Doctor of Philosophy (Ph.D.) Thesis 2023

Supervisors:

Alejandro Pazos Sierra

José Luís Oliveira

PhD Programme in Information and Communications Technology



UNIVERSIDADE DA CORUÑA

Dr. Alejandro Pazos Sierra, Profesor Catedrático del área de Ciencias de la Computación e inteligencia Artificial de la University of A Coruña, España.

Dr. Alejandro Pazos Sierra, Full professor at Computer Science and Information Technologies, University of A Coruña, Spain.

Dr. José Luís Oliveira, Profesor Catedrático del Departamento de Electrónica, Telecomunicaciones e Informática de la Universidad de Aveiro, Portugal.

Dr. José Luís Oliveira, Full professor at Department of Electronics, Telecommunications and Informatics, University of Aveiro, Portugal.

Autorizan:

Approve:

La presentación para su depósito de la tesis que dirige y que fue realizado por **João Rafael Duarte de Almeida** con número de identidad 14316971 con título “**Enriching information extraction pipelines in clinical decision support systems**”.

The presentation of the thesis “Enriching information extraction pipelines in clinical decision support systems”, written by João Rafael Duarte de Almeida with identity number 14316971.

Y para que así conste, firma esta autorización en A Coruña (España) y Aveiro (Portugal), en Enero del 2023.

And for the record, this certificated is issued in A Coruña (Spain) and Aveiro (Portugal), in January 2023.

Los directores de la tesis
Supervisors

Fdo. Dr. Alejandro Pazos Sierra

Fdo. Dr. José Luís Oliveira

A todos que me acompanharam nesta aventura.
A todos los que me acompañaron en esta aventura.
To everyone that stayed by my side on this journey.

Acknowledgements

This doctorate was a rewarding adventure, in which I had the opportunity to learn things and achieve results that are much beyond my initial expectations. Here, I want to express my gratitude to the good friends that helped me along this path.

In the first place, I would like to express my deepest gratitude to my advisors, Dr. Alejandro Pazos Sierra and Dr. José Luís Oliveira for giving me the opportunity of carrying out my research project. In that regard, a special thanks to Dr. José Luís Oliveira for his guidance and unconditional support in pursuing my research ideas. I also want to thank all my friends and family for supporting me during this journey. In particular, I would like to thank to:

- My closest family, my parents, my sister and my partner Sandra for their patience and love;
- Bastião for his friendship and welcoming me to BMD Software during my Erasmus;
- Eriksson for the long conversations and encouragement to be the best of myself;
- Cajús for being one of the best friends I always believed in me, when I thought of giving up;
- Jorge for his partnership and for being on this journey with me, doing his Ph.D.;
- Fran for all the support that he provided me during this doctorate, without him, nothing was possible;
- The *Velhotes*, as I like to call them, which includes Filipe, José and Manuel; Also I like to thank Manuel for the support in linguist issues;
- All contributors in the tools that I developed; A special thank to João, Leonardo and André for helping me in the creation of some of the developed tools in this work;
- The EHDEN project, especially to Peter Rijnbeek for insightful discussions and support.

Finally, I gratefully acknowledge the “FCT — Fundação para a Ciência e Tecnologia” for making possible this Ph.D. work, through the grant **SFRH/BD/147837/2019**.

” *Study while others are sleeping; work while others are loafing; prepare while others are playing; and dream while others are wishing.*

— **William Arthur Ward**
American motivational writer

Resumo

Os estudos sanitarios de múltiples centros son importantes para aumentar a repercusión dos resultados da investigación médica debido ao número de suxeitos que poden participar neles. Para simplificar a execución destes estudos, o proceso de intercambio de datos debería ser sinxelo, por exemplo, mediante o uso de bases de datos interoperables. Con todo, a consecución desta interoperabilidade segue sendo un tema de investigación en curso, sobre todo debido aos problemas de gobernanza e privacidade dos datos. Na primeira fase deste traballo, propoñemos varias metodoloxías para optimizar os procesos de estandarización das bases de datos sanitarias. Este traballo centrouse na estandarización de fontes de datos heteroxéneas nun esquema de datos estándar, concretamente o OMOP CDM, que foi desenvolvido e promovido pola comunidade OHDSI. Validamos a nosa proposta utilizando conxuntos de datos de pacientes con enfermidade de Alzheimer procedentes de distintas institucións. Na seguinte etapa, co obxectivo de enriquecer a información almacenada nas bases de datos de OMOP CDM, investigamos solucións para extraer conceptos clínicos de narrativas non estruturadas, utilizando técnicas de recuperación de información e de procesamento da linguaxe natural. A validación realizouse a través de conxuntos de datos proporcionados en desafíos científicos, concretamente no National NLP Clinical Challenges(n2c2). Na etapa final, propuxémonos simplificar a execución de protocolos de estudos provenientes de múltiples centros, propoñendo solucións novas para perfilar, publicar e facilitar o descubrimento de bases de datos. Algunhas das solucións desenvolvidas están a utilizarse actualmente en tres proxectos europeos destinados a crear redes federadas de bases de datos de saúde en toda Europa.

Palabras chave: Datos sanitarios, perfilado de bases de datos, integración de datos, minería de textos, OMOP CDM

Resumen

Los estudios sanitarios de múltiples centros son importantes para aumentar la repercusión de los resultados de la investigación médica debido al número de sujetos que pueden participar en ellos. Para simplificar la ejecución de estos estudios, el proceso de intercambio de datos debería ser sencillo, por ejemplo, mediante el uso de bases de datos interoperables. Sin embargo, la consecución de esta interoperabilidad sigue siendo un tema de investigación en curso, sobre todo debido a los problemas de gobernanza y privacidad de los datos. En la primera fase de este trabajo, proponemos varias metodologías para optimizar los procesos de estandarización de las bases de datos sanitarias. Este trabajo se centró en la estandarización de fuentes de datos heterogéneas en un esquema de datos estándar, concretamente el OMOP CDM, que ha sido desarrollado y promovido por la comunidad OHDSI. Validamos nuestra propuesta utilizando conjuntos de datos de pacientes con enfermedad de Alzheimer procedentes de distintas instituciones. En la siguiente etapa, con el objetivo de enriquecer la información almacenada en las bases de datos de OMOP CDM, hemos investigado soluciones para extraer conceptos clínicos de narrativas no estructuradas, utilizando técnicas de recuperación de información y de procesamiento del lenguaje natural. La validación se realizó a través de conjuntos de datos proporcionados en desafíos científicos, concretamente en el National NLP Clinical Challenges (n2c2). En la etapa final, nos propusimos simplificar la ejecución de protocolos de estudios provenientes de múltiples centros, proponiendo soluciones novedosas para perfilar, publicar y facilitar el descubrimiento de bases de datos. Algunas de las soluciones desarrolladas se están utilizando actualmente en tres proyectos europeos destinados a crear redes federadas de bases de datos de salud en toda Europa.

Palabras clave: Datos sanitarios, perfilado de bases de datos, integración de datos, minería de textos, OMOP CDM

Abstract

Multicentre health studies are important to increase the impact of medical research findings due to the number of subjects that they are able to engage. To simplify the execution of these studies, the data-sharing process should be effortless, for instance, through the use of interoperable databases. However, achieving this interoperability is still an ongoing research topic, namely due to data governance and privacy issues. In the first stage of this work, we propose several methodologies to optimise the harmonisation pipelines of health databases. This work was focused on harmonising heterogeneous data sources into a standard data schema, namely the OMOP CDM which has been developed and promoted by the OHDSI community. We validated our proposal using data sets of Alzheimer's disease patients from distinct institutions. In the following stage, aiming to enrich the information stored in OMOP CDM databases, we have investigated solutions to extract clinical concepts from unstructured narratives, using information retrieval and natural language processing techniques. The validation was performed through datasets provided in scientific challenges, namely in the National NLP Clinical Challenges (n2c2). In the final stage, we aimed to simplify the protocol execution of multicentre studies, by proposing novel solutions for profiling, publishing and facilitating the discovery of databases. Some of the developed solutions are currently being used in three European projects aiming to create federated networks of health databases across Europe.

Keywords: Health data, Database profiling, Data integration, Text mining, OMOP CDM

Table of Contents

1	Introduction	1
1.1	Motivation	2
1.2	Main objectives	3
1.3	Key contributions	4
1.4	Organization	6
2	General principals, hypotheses and means of verification	9
2.1	Fundamentals	10
2.1.1	Biomedical data	10
2.1.2	Data integration	15
2.1.3	Data analysis	22
2.2	Research questions	28
2.3	Hypotheses	29
2.4	Means of verification	31
3	Semi-automatic translation of data sources into a common schema	33
3.1	Contribution	34
3.2	Background	35
3.2.1	Most common ETL tools	36
3.2.2	Mapping concepts	39
3.3	Methodology for cohort harmonization	40
3.3.1	Overview	41
3.3.2	The cohort common data schema	42
3.3.3	OHDSI ETL tools	43
3.3.4	Collaborative ontology development	44
3.4	The cohort migrator toolkit	45
3.4.1	Data harmonisation	46
3.4.2	Customised operations	47
3.4.3	Data loading into OMOP CDM	48
3.4.4	Limitations	49
3.5	A collaborative web-based ETL tool	50
3.5.1	Software architecture	50
3.5.2	Main functionalities	51
3.5.3	Collaborative features	55

3.5.4	Usagi mapper connector	56
3.6	Results	57
3.6.1	Ontology for Alzheimer's disease cohorts	58
3.6.2	Cohort harmonisation	59
3.6.3	BIcenter-AD applied to Alzheimer's diseases datasets	61
3.7	Discussion	62
3.7.1	Data quality and analysis	63
3.7.2	Datasets interoperability	63
3.7.3	Data privacy	64
3.7.4	Multi-institutional environments	65
3.7.5	Impact of web ETL tools	66
3.8	Final considerations	67
4	From unstructured text to ontology-based registers	69
4.1	Contribution	70
4.2	Background	72
4.2.1	Retrieving patient information	72
4.2.2	Cross-language matching	73
4.2.3	Patient Relatives Extraction Approaches	74
4.3	Extract and harmonize drug mentions	75
4.3.1	Clinical Notes Analysis	75
4.3.2	Data Harmonisation	78
4.4	Multi-language Concept Normalisation	83
4.4.1	Supportive open-source tools	83
4.4.2	Multi-language mapper	84
4.5	Extraction of family history information	87
4.5.1	Dependency parsing rules	87
4.5.2	Phrase characteristics extraction	88
4.5.3	Rule-based engine	90
4.6	Results	92
4.6.1	Drug mentions extraction and harmonization	92
4.6.2	Multi-language cohort harmonisation	96
4.6.3	Patient family extraction	97
4.7	Discussion	99
4.7.1	Systems synopsis	101
4.7.2	Main limitations	103
4.8	Final considerations	106
5	Scalable database profiling for multicentre studies	107
5.1	Contribution	108
5.2	Background	109
5.2.1	Database profiling	110

5.2.2	Discovery of medical databases	112
5.2.3	Streamlining multicentre studies	113
5.3	Framework for profiling databases	115
5.3.1	Functional requirements	115
5.3.2	System overview	117
5.3.3	MONTRA Software Development Kit (SDK)	119
5.3.4	Data representation	120
5.3.5	Endpoints for interoperability	124
5.3.6	Access control mechanisms	125
5.4	Recommending health databases	127
5.4.1	Feature extraction	127
5.4.2	Collaborative filtering	128
5.4.3	Content-based retrieval	129
5.4.4	System overview	130
5.5	Explore distributed patient-level databases	131
5.5.1	Methodology overview	132
5.5.2	Functional requirements	132
5.5.3	Study Manager architecture	134
5.5.4	Features and user experience	135
5.6	Results	138
5.6.1	Portals for biomedical data sharing	138
5.6.2	Study Manager, a plugin for study orchestration	140
5.7	Discussion	141
5.7.1	Evolution from the first version of MONTRA	141
5.7.2	Compliance with FAIR principles	142
5.7.3	Recommender system's impact	143
5.7.4	Streamlining and orchestrating studies	144
5.8	Final considerations	145
6	Conclusions	147
6.1	Outcomes overview	148
6.2	Future work and limitations	148
6.3	Research directions	150
	References	153
	Appendices A Sinopsis in Spanish	171
A.1	Introducción	171
A.2	Traducción semiautomática de fuentes de datos a un esquema común	174
A.2.1	Metodología para la armonización de cohortes	174
A.2.2	El conjunto de herramientas del migrador de cohortes	175
A.2.3	BIcenter y BIcenter-AD	175

A.3	De texto no estruturado a registros basados en ontoloxías	176
A.3.1	Extraer y armonizar las menciones de medicamentos	177
A.3.2	Normalización de conceptos en varios idiomas	178
A.3.3	Extracción de información de historia familiar	178
A.4	Perfiles de bases de datos escalables para estudios multicéntricos . . .	179
A.4.1	Marco para crear perfiles de bases de datos	180
A.4.2	Recomendar bases de datos de salud	181
A.4.3	Explorar bases de datos distribuidas a nivel de paciente . . .	181
A.5	Conclusiones	182
Appendices B Sinopsis in Galician		185
B.1	Introdución	185
B.2	Tradución semiautomática de fontes de datos a un esquema común .	187
B.2.1	Metodoloxía para a harmonización de cohortes	188
B.2.2	O conxunto de ferramentas do migrador de cohortes	189
B.2.3	BIcenter e BIcenter-AD	189
B.3	De texto non estruturado a rexistros baseados en ontoloxías	190
B.3.1	Extraer e harmonizar as mencións de medicamentos	190
B.3.2	Normalización de conceptos en varios idiomas	191
B.3.3	Extracción de información de historia familiar	192
B.4	Perfís de bases de datos escalables para estudos multicéntricos	193
B.4.1	Marco para crear perfís de bases de datos	193
B.4.2	Recomendar bases de datos de saúde	194
B.4.3	Explorar bases de datos distribuídas a nivel de paciente . . .	195
B.5	Conclusións	196

List of Figures

2.1	Federated integration architecture.	19
2.2	OHDSI architecture using OMOP CDM.	20
2.3	The OMOP CDM schema	22
2.4	Overview of the study stages.	23
2.5	Methodology for performing distributed and moderated queries.	25
3.1	OMOP CDM tables adopted in cohort migration.	43
3.2	Migration workflow from CSV to OMOP CDM.	45
3.3	Key-value structure for cohort raw data.	46
3.4	Ad-hoc modules from the ETL workflow.	48
3.5	Blcenter architecture.	50
3.6	Example of ETL pipeline.	52
3.7	Example of Blcenter component.	52
3.8	Blcenter view for performance metrics.	53
3.9	Usagi mapper component in Blcenter.	57
3.10	Ontology node to define a standard concept.	59
3.11	Data transformation stages.	62
4.1	Workflow from text to a matrix structure	76
4.2	Example of clinical note being processed.	79
4.3	Overview of the data harmonisation pipeline.	80
4.4	OMOP CDM tables used for storing extracted drugs.	82
4.5	Ontology node with different translations.	85
4.6	Multi-language system operation nodes.	86
4.7	Example of dependency parsing and coreference resolution.	88
4.8	Example of an annotated narrative.	91
4.9	Family members detection workflow.	91
5.1	Representation of the fingerprint concept.	108
5.2	Three-layer MONTRA 2 architecture.	118
5.3	MONTRA 2 plugins life cycle.	120
5.4	Example of fingerprint.	121

5.5	Example of catalogue view.	122
5.6	Overview of the Database-level Dashboard.	123
5.7	Overview of the Network Dashboard	124
5.8	Collaborative filtering matrix to correlate clicks.	131
5.9	Content-base matrix to compare data sources.	132
5.10	Methodology overview for managing the distributed queries.	133
5.11	Study Manager architecture.	134
5.12	Interaction diagram where data owners process a query.	136
5.13	Recommender system integrated in EMIF Catalogue	144

List of Tables

3.1	Criteria fulfillment by the candidates	38
3.2	Summary of attributes migrated cohorts.	60
4.1	Dataset statistics for the 2009 i2b2 medication extraction challenge. . .	93
4.2	Dataset statistics for the 2018 n2c2 medication extraction challenge. .	93
4.3	Evaluation results from the medication extraction component.	93
4.4	Results of the mapperd drug mentions.	95
4.5	Results of multi-language concept normaliser.	97
4.6	Dataset statistics for the n2c2/OHNLP track on family history extraction.	98
4.7	Results for n2c2 challenge, subtask 1.	99
4.8	Results for n2c2 challenge, subtask 2.	99
4.9	Error analysis of DrAC results.	103
4.10	Analyses of the most common false positives.	105

List of Abbreviations

AAI	Authentication and Authorization Infrastructure
ACHILLES	Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems
ACL	Access Control List
AD	Active Directory
ADE	Adverse Drug Events
AES	Advanced Encryption Standard
AI	Artificial Intelligence
AOD	Alcohol and Other Drug Thesaurus
API	Application Programming Interfaces
BBACL	BioBank Alzheimer Center Limburg
BI	Business Intelligence
BMC	Berlin Memory Clinic
BPM	Business Process Management
CDISC	Clinical Data Interchange Standards
CDM	Common Data Model
CIMI	Clinical Information Modeling Initiative
CRM	Customer Relationship Management
CRUD	Create, Read, Update, and Delete
CSF	Cerebrospinal Fluid
CSS	Cascading Style Sheets
CT	Computed Tomography
CUI	Concept Unique Identifier
DATS	Data Tag Suite
DBMS	Database Management System
DCAT	Data Catalog Vocabulary

DICOM	Digital Imaging and Communications in Medicine
DNA	Deoxyribonucleic Acid
DW	Data Warehousing
EHDEN	European Health Data Evidence Network
EHR	Electronic Health Record
EHR4CR	Electronic Health Records for Clinical Research
EIS	Enterprise Information Systems
EMIF	European Medical Information Framework
EMIF-AD	EMIF - Alzheimer's Disease
ETL	Extract, Transform and Load
FAIR	Findability, Accessibility, Interoperability, and Reusability
FIdM	Federated Identity Management
GDPR	General Data Protection Regulation
GUI	Graphical User Interface
HL7	Health Level 7
HL7-FHIR	HL7 - Fast Healthcare Interoperability Resources
HL7-CDA	HL7 - Clinical Document Architecture
HMORN	Health Maintenance Organization Research Network
HTML	HyperText Markup Language
i2b2	Integrating Biology and the Bedside
ICD	International Classification of Diseases
IDE	Integrated Development Environment
IdP	Identity Provider
IE	Information Entities
IT	Information Technology
JPA	Java Persistence API
JSON	JavaScript Object Notation
LDAP	Lightweight Directory Access Protocol
MeSH	Medical Subject Headings
MIMIC-III	Medical Information Mart for Intensive Care-III
MRI	Magnetic Resonance Imaging
MVC	Model-View-Controller
MWTR	Minimum Weighted Tree Reconstruction

n2c2	National NLP Clinical Challenges
NEN	Named Entity Normalisation
NER	Named Entity Recognition
NFT	Non-Fungible Token
NGS	Next-generation sequencing
NLP	Natural Language Processing
OAuth	Open Authentication
OHDSI	Observational Health Data Sciences and Informatics
OIDC	OpenID Connect
OLAP	Online Analytical Processing
OMOP	Observational Medical Outcomes Partnership
OMOP CDM	OMOP Common Data Model
PACS	Picture Archiving and Communication System
PDI	Pentaho Data Integration
PET	Positron Emission Tomography
PPDP	Privacy Preserving Data Publishing
PRRE	Private Remote Research Environment
RAD	Rapid Application Development
RBAC	Role-based Access Control
RDBMS	Relational Database Management System
RDF	Resource Description Framework
RP	Relying Parties
RSA	RivestShamirAdleman
SAML	Security Assertion Markup Language
SDK	Software Development Kit
SME	Small and Medium-sized Enterprise
SNOMED CT	Systematized Nomenclature of Medicine Clinical Terms
SOA	Service-Oriented Architecture
SQL	Structured Query Language
SSO	Single Sign-On
SVD	Singular Value Decomposition
T2DM	Type 2 Diabetes Mellitus
TLV	Tag-Length-Value
TOS	Talend Open Studio
UIMA	Unstructured Information Management Architecture

UMLS	Unified Medical Language System
URI	Uniform Resource Identifier
WHO	World Health Organization

Introduction

During the last decades, huge amounts of clinical data have been collected in Electronic Health Record (EHR) to support healthcare services. Besides this primary application, its secondary use at a large scale can help identify disease distribution among the population, understand the path of diseases' progression and comorbidities, or evaluate treatment efficacy. Due to the diversity of data structures and concepts between EHR vendors, the databases associated with these systems are not interoperable between them. A possible solution is the migration of the data into a common data schema. However, the process associated with the data conversion, as well as its later analysis, are still challenging tasks. This chapter outlines the motivation and objectives of this research, summarizes the main achievements, and describes the organization of the thesis.

The continuous demand for better health diagnostics and treatment has motivated many clinical research studies, such as observational studies and clinical trials. In clinical trials, patients are commonly divided into two or more groups (*e.g.* active and placebo), to study the effectiveness of the treatment for a particular clinical condition [1]. In this case, there is direct intervention with the patients, *e.g.*, administration of a drug or therapeutic procedures. However, this approach is not always the most appropriate, *e.g.* addressing research questions in plastic surgery through randomised controlled trials is often subject to ethical constraints [2]. In observational studies, researchers do not perform any active intervention with patients, and exposure occurs naturally or through other factors. Here, medical researchers limit themselves to documenting the relationship between the exposure and the outcome in the study [1].

Observational studies follow different strategies and are established by defining a set of inclusion and exclusion criteria for the subjects involved, as well as several features that are identified and observed over time [3]. Some initiatives reuse the data already collected in medical institutions to conduct observational studies. This practice saves time and enables pre-verification of the number of subjects before starting the analysis [4]. However, in specific diseases, it is necessary to collect information about the selected subjects. In these situations, the data are recorded based on the study guidelines and different solutions can be used to store the data, for instance, the institutional EHR system [5].

The dependence on technical teams is a major barrier when there is a need to extract data for analysis. Many other ethical and technical issues are raised when one wants to combine datasets from distinct organisations. This is the case of multicentre studies that aim to increase the population size, the power of the statistical evidence, and thereby the study's impact [6].

1.1 Motivation

While a medical study can be successfully conducted regardless of the data-collecting strategy, the number of subjects is still a big concern. One of the strategies to solve the problem of not having enough subjects to obtain impactful findings was the creation of multicentre studies [6, 4]. However, this type of study raise other challenges, namely the lack of interoperability between datasets. This problem is very common when the studies are not conducted following shared principles for data collection and storage. Examples of issues resulting from this lack of interoperability are associated with the data structure or the codification of medical concepts used to characterise the patients' conditions. Although a medical researcher can identify these similarities, for computational analysis, the data extraction and processing require distinct Extract, Transform and Load (ETL) processes specific for each institution [7].

Integrating multiple data sources is not just a technological problem. There are some ETL tools capable of performing this task using large amounts of data. The non-technical problem of this aggregation is the data domain, *i.e.*, identifying the concepts in the data and combine correctly the information associated with them that was extracted from multiple sources. Healthcare databases belong to one of the domains in which this is a concerning problem, due to the variety of concepts to represent similar procedures and medical terms. Solving these problems is helpful to optimize studies, but sharing patient-level data still raises some privacy issues, due to legal, ethical and regulatory requirements [8]. Patient data are very sensitive and disruption of this privacy can have dramatic consequences for individuals, healthcare providers and subgroups within society [9]. Besides, the legislation may be different in each country, which makes it difficult to define a protocol that fits all the institutions involved [10]. This is another challenge, which requires finding a solution that allows the analysis of multiple data sources, or parts of these, without exposing sensitive data.

The potential impact of multicentre studies has motivated researchers to seek more robust and reusable solutions to aggregate knowledge from distributed health datasets. Organisations and methodologies were established to explore clinical databases by reusing existent data [4]. One of these efforts aims to create a strategy to reuse EHR databases using a homogeneous schema, in order to facilitate the interoperability between databases. This integration is currently possible through the use of open-source frameworks that help support the whole process [7].

1.2 Main objectives

The main objective of this work is to investigate a new strategy to use medical data from distributed databases to conduct health studies. To achieve this objective, the present thesis seeks to answer the following research question:

Can medical researchers conduct distributed health studies, using multiple data sources from different institutions?

This research question can be addressed in multiple dimensions. Therefore, we answered this by dividing this work into four main tasks:

1. Investigate the state-of-the-art, namely: i) methodologies to migrate and harmonise medical records into a common data schema; ii) techniques to retrieve information from unstructured data; iii) solutions for profiling health data sources, including supportive tools for conducting medical studies;
2. Propose strategies for semi-automatically converting heterogeneous data sources into a common data schema;
3. Enhancing the information stored in these databases by retrieving medical concepts from unstructured text, using Natural Language Processing (NLP) techniques;
4. Providing a strategy for exploring health databases and facilitating health researchers to conduct distributed studies.

1.3 Key contributions

During this doctorate, several scientific contributions have been made, namely, 38 peer-reviewed indexed publications and 11 open-source tools. Of these publications, 13 were in journals, 23 in conferences, and 2 were book chapters. Besides these publications, as result of the work developed during this doctorate, I was appointed 2 times co-editor of proceedings at IEEE international conferences, with the role of Program Chair. This document is a compilation of some selected contributions, according to the following categories:

1. Scientific contributions

- Methodologies to harmonise cohort datasets in multicentre clinical research [7, 11];
- Collaborative web ETL solutions [12, 13];
- Methodologies to extract relevant concepts from clinical notes to enrich structured OMOP Common Data Model (OMOP CDM) databases [14, 15];
- Solutions to unify and extract family's health history information from clinical notes using rule-based techniques in NLP [16, 17];
- A flexible framework for health databases profiling (submitted);
- A modular task management system to support health research studies [18];
- A secure architecture for exploring patient-level databases from distributed institutions [19].

2. Open-source software/tools

- CMToolkit¹ is a python-based application designed to migrate and harmonize clinical cohorts from CSV format into the Observational Health Data Sciences and Informatics (OHDSI) OMOP CDM schema;

¹<https://bioinformatics-ua.github.io/CMToolkit/>

- TranSMART-Migrator² is a command-line application for migrating patients' information from the OMOP CDM database to TranSMART structure;
- BICenter³ is a web ETL tool using Pentaho Kettle as the DI execution engine;
- BICenter-AD⁴ is an evolution of BICenter focused in the context of Alzheimer's Diseases to support the collaboration between distinct entities in the definition and implementation of ETL pipelines;
- DrAC⁵ is software solution for extracting Patients' information from clinical notes and exporting it to an OMOP CDM database;
- PatientFM⁶ is an end-to-end system for extracting family history information from clinical notes;
- MONTRA2 Framework⁷ is a platform to profile and catalog biomedical database, aiming data sharing between medical researchers;
- EHDEN Network Dashboards⁸ is a tool used for profiling and comparing federated health databases for large-scale observational research;
- TASKA⁹ is a modular and easily extendable system for repeatable workflows used to simplify the coordination of teams while conducting medical studies.

3. Clinical application for the developed solutions

- One of the first clinical application made during this work was the upgrade of EMIF-Catalogue¹⁰, which started by adopting the new version of MONTRA Framework. These enhancements were reflected in an extension of

²<https://github.com/bioinformatics-ua/tranSMART-migrator>

³<https://bioinformatics-ua.github.io/BICenter/>

⁴<https://bioinformatics-ua.github.io/BICenter-AD/>

⁵<https://github.com/bioinformatics-ua/DrAC>

⁶<https://github.com/bioinformatics-ua/PatientFM>

⁷<https://github.com/bioinformatics-ua/montra2>

⁸<https://github.com/EHDEN/NetworkDashboards>

⁹<https://github.com/bioinformatics-ua/taska>

¹⁰<https://emif-catalogue.eu/>

the European Medical Information Framework (EMIF) project. In a branch of this extension, namely the EMIF - Alzheimer's Disease (EMIF-AD), it was created CMToolkit and TranSMART-Migrator to harmonise patients' data into standardised data structures. Both tools were validated using datasets of patients suffering from Alzheimer's Disease [7, 11].

- In close collaboration with European Health Data Evidence Network (EHDEN) partners, the EHDEN Portal¹¹ was launched. This platform is also supported by MONTRA Framework at its core. One of the plugins incorporated in this scope was the EHDEN Network Dashboards, to better represent the characteristics of the databases within this project.

1.4 Organization

This document is organized into more five chapters. Chapter 2 summarizes the current state-of-the-art of fundamental concepts that are the base of this work. It presents major biomedical data sources used by medical institutions, and some of the well-known initiatives that support health data integration and exploration. In this chapter, the general assumptions under which this work was conducted are also described, and an explicit list of hypotheses that have been verified along this work, including the means used for verification.

Chapter 3 describes the proposed methodologies to migrate and harmonise heterogeneous medical records into a standard data schema. These methodologies resulted into four open source solutions and four peer-reviewed publications.

Chapter 4 presents strategies for solving some of the existent gaps in ETL procedures regarding the harmonisation of clinical concepts extracted from clinical notes into a relational database. The work presented in mainly focused on four peer-reviewed publications, including three open source solutions.

Chapter 5 details the approach proposed to extract and expose metadata from health databases into a centralised platform. The work aims to streamline medical studies, supporting researchers with a set of tools and methodologies that simplify several steps from the study design to its results. This chapter is mainly based on four

¹¹<https://portal.ehden.eu/>

scientific publications, four open source tools and one master thesis that was made during this doctorate.

Finally, Chapter 6 presents the final remarks of this work, highlighting directions for future work.

General principals, hypotheses and means of verification

Multicentre studies can be conducted following different methodologies and focused on distinct data types. Understanding the most common biomedical data formats and the possibilities of reusing data already collected are essential to plan this work. Therefore, in this chapter, some fundamental concepts are described to contextualize and justify the decisions behind the produced outcomes. The chapter also specifies the research questions, hypotheses and the means of verification that we will propose.

The motivation of this work was to propose methodologies and tools to simplify the execution of multicentre studies in the medical field. Accomplishing this objective requires a fundamental analysis of several topics, namely related to the current biomedical data used when conducting medical studies. While understanding the most used data formats, we tried to provide strategies capable of reusing data already collected from other ends, aiming to optimise the study execution. This enhancement of information may help researchers to understand the study feasibility at early stages, by identifying if the proposed studies would have enough samples (or patient information) to produce impactful findings.

When discussing multicenter studies, data integration and analysis are two essential topics. An overview of those is also covered in this chapter when describing the distinct biomedical data sources, and initiatives to integrate and explore the information from heterogeneous data sources.

The chapter also introduces the research questions behind this work, followed by the hypotheses that we will try to verify. It finalizes with a brief description of the means used to validate the solution that will be proposed in this document.

2.1 Fundamentals

One of the most important parts of a medical study is the patients' information correlated to the study scope. The patient's characteristics are crucial to the success of medical studies since it is estimated that up to 50 % of trials are not completed due to insufficient enrollment [20]. When collecting these characteristics, the procedure is achieved to a specific goal, resulting in distinct data types. For instance, distinct medical tests can be performed on the subjects, such as blood analysis, medical imaging, and electrocardiogram, in order to identify any health issue. This information contributes to the medical history of each person, which can be a valuable insight for future diagnostics or prognostics [21]. However, the digital data formats for those records can be different.

2.1.1 Biomedical data

The secondary use of medical data to conduct research studies has become a common practice. These studies have provided complementary support to generate new insights and knowledge, namely in pragmatic trials using records collected from routine clinical care visits, comparative effectiveness studies or patient-centred outcomes research [22]. These data were not primarily generated to support research or secondary analysis. This concept refers to the use of data for purposes other than those that it was originally collected [23]. However, over the last few years, the clinical research community recognized that recruiting patients to record their medical characteristics over time is challenging. Although this practice is required in some types of studies, medical research is currently not limited to them. Therefore, this section provides a brief overview of the most relevant data types in the medical domain.

Electronic health records

An Electronic Health Record (EHR) is a digital version of the data collected about a patient. Hospitals usually have a EHR system to make the information available in real-time to the health professionals of the institution [24]. Besides patient data, there is an amount of additional information regarding patients' medical conditions that is also stored in such systems. EHR aims to simplify the data management and data exchange within the institution, between different services, resulting in higher quality and safer care for patients.

Some of the data stored in these systems follows a tabular structure, following the principles of relational databases [25]. Although there are some efforts in having interoperable databases supporting the EHR systems, each vendor has its own data schema. Over the last years, several EHR standards were developed, namely the CEN ISO 13606 [26], OpenEHR [27], OMOP CDM from OHDSI [4, 28] and Health Level 7 (HL7) standards (Clinical Information Modeling Initiative (CIMI), HL7 - Clinical Document Architecture (HL7-CDA) HL7 - Fast Healthcare Interoperability Resources (HL7-FHIR)) [25]. The HL7 standards were proposed to simplify the transfer of clinical and administrative data between software applications used by different healthcare providers. They define guidelines and methodologies to support the communication between various healthcare systems [29].

Utilising the data from the EHR system to answer healthcare questions differs from the traditional approach based on collecting data after defining a question [24]. The tabular data can help medical researchers conducting different types of studies, namely by identifying patient populations with specific healthcare interventions and outcomes, *e.g.*, related to drug exposure, procedures, and conditions, among others. The parameters available to characterised these patient populations are various, including demographic information, healthcare delivery, utilization and cost, morbidities, treatments and sequence of treatment, and disease natural history.

Although EHR has been used for many years, as well as the idea of secondary use of this data, the process of reutilising its data still raises some challenges. These challenges include limitations of processing ability [30, 31], interoperability [32, 33], inability to extract the required information [34, 35], and security and privacy concerns [36].

EHR can be a formal strategy for federating all biomedical data into a single integrated view. Some information may not be possible to represent in this format, such as medical images or omics data. However, this tabular format already contains valuable information to be used for conducting medical studies.

Clinical notes

EHR systems can have repositories of non-tabular patients' information, *e.g.*, clinical notes. The text data contained in these notes is typically subdivided into main categories, depending on whether they are structured or not. The structured notes, as the name indicates, integrate some structured format, for instance, a form. Examples of this data are the diagnosis forms or the laboratory analysis results. Alternatively,

unstructured notes refer to notes that contain free-text, for instance, some of the physician's notes transcripts [37].

Free text notes are characterised by their vast variability, especially due to their heterogeneity. EHR contains agglomerates of different types of medical narratives (for progress, admission, operative, primary care and discharge, among others), with different dimensions (from very short to very long) [38]. Structuring the information available in medical narratives is challenging due to this reason, but also because those are often ungrammatical. They contain short telegraphic sentences, plenty of misspellings, and are filled with abbreviations. In some cases, these abbreviations refer to local dialectal shorthand expressions, which may overload the use of acronyms, *i.e.*, the same group of letters with different meanings [37].

One strategy to introduce some kind of structure in these notes is making use of pseudo-templates or integrating tabular data into the narratives, for instance, the laboratory results. This pseudo-structure is not generalised, and nothing ensures that this is used in all narratives present in the system. Besides, the adoption can vary between physicians, services, or institutions [37]. A different attempt to increase the readability of free-text notes was through the adoption of standard lexicons to encode the information present in the narratives [39]. The lexicons are explained in more detail in Section 2.1.2.

Nonetheless, since clinical text poses great interest, some approaches have been developed for extracting relevant information. Even though this process has historically consisted of having clinical experts manually review clinical notes, a process that cannot scale with the growing rate of generation of medical data [40], much research has been made during the past years in domains such as clinical NLP to create systems capable of automatically annotating and summarising important text content in clinical notes [41].

The use of unprocessed narratives for conducting multicentre studies raises several challenges, namely regarding patients' privacy and data interoperability. Mapping the medical concepts to their standard definition solves the latter issue, but raises other challenges since the mapping task can be extremely complex and time-consuming. Moreover, it is acknowledged that the challenging nature of the free text can make it difficult to develop automatic information extraction solutions for clinical text [42].

Medical imaging data

The development of medical imaging transformed medicine and made it a valuable source of data for prognosis and diagnosis. It entails obtaining visual information from the patient's body with less invasive techniques than those used before the development of this technology [43].

Picture Archiving and Communication System (PACS) defines a set of systems, that includes software, hardware and communication networks, for the acquisition, distribution, storage and analysis of digital images in order to allow connectivity, compatibility and workflow optimizations between different medical imaging equipment [44]. The proliferation of PACS was possible mostly due to the development of Digital Imaging and Communications in Medicine (DICOM), the standard for the handling of medical imaging data.

The DICOM standard supports not only the pixel data that defines the medical images but also a wide range of metadata information related to all the stakeholders involved in the clinical practice, such as the patient, procedure, equipment, staff-related data or structured report. Data relative to these stakeholders is conveyed by DICOM data elements which compose DICOM objects or files.

DICOM data elements are encoded using a Tag-Length-Value (TLV) structure. The tag field identifies the data element and includes two subfields: i) the group identifier; and ii) the element identifier within the group, both encoded using 16-bit unsigned values. DICOM data elements are grouped by their relation with real-world entities, *i.e.*, the Information Entities (IE) that represents, for instance, the patient, the study and the series. These elements hold the information related to the patient that is encompassed in the patient group. Apart from the tag, DICOM data elements include also the fields length (in bytes) and value (that holds the actual element's data).

DICOM object is an umbrella term to describe a DICOM file, which could be images, and structure reports, among others. The information enclosed in DICOM objects is very heterogeneous. There are data elements for representing names, measures, and dates, among others. Therefore, in order to express all these data types, the encoding of the value field changes according to the element's type.

Although medical images contain reliable information that can be used for specific medical studies, we did not contemplate this type of data in this work. However,

we recognise that this data can enrich the clinical information retrieved from EHR datasources [45].

Omics

Omics encompasses a large number of biology areas of study that aim the analysis of the complete genetic or molecular profiles of humans or other organisms. This research includes the study of genomics, proteomics and metabolomics. Next-generation sequencing (NGS) has become an essential technology in genetic and genomic analysis with a substantial impact in the fields of biomedicine and anthropology. The advantages of NGS over traditional methods include its multiplex capability and analytical resolution, making it a time and cost-efficient approach for fast clinical and forensic screening [46]. This technology prompted a new step in clinical research, in which it is possible to scan the whole genome of individual DNA samples at an acceptable cost and time [47, 48].

Biobanking currently represents a new research field that involves international infrastructures and government agencies requiring the creation of policies to provide ethical and legal guidelines for public health [49]. The need for high-quality and clinically annotated biospecimens for personalised medicine and forensic applications is raising new research challenges [50, 51]. However, other major problems have followed this growth, namely the evolution of biobanking in a decentralized way, with heterogeneous procedures for data collection and storage, as well as different legal policies for data access [52].

One of the key challenges is to find the right balance between preserving the privacy of the subjects in the study and the data availability for sharing the results through global research networks [48]. Although genomics datasets are not linked to medical records, which preserves subject identity [53], some authors tried to reverse the process only using the DNA present in the datasets.

Privacy issues are one of the main obstructions in health research, including in the area of genomics [54, 55]. Answers to biomedical questions may currently be hidden in private data repositories that are not explored due to the lack of methodologies to analyse this data [56]. The problem can be addressed at different levels, from biomedical data discovery to multi-repository analysis, *i.e.*, there are gaps in the way biobanks are exposed to the research community and the methodologies currently available are not designed to simplify the exploration of multiple and private repositories [57].

2.1.2 Data integration

In the medical domain, studies can be successfully conducted regardless of the data-collecting strategy, however, the number of subjects is still a big concern. In some diseases, there are not enough subjects for a study with impactful findings. The idea of multicentre studies emerged from this need, aiming to increase the number of subjects, the power of the statistical evidence, and thereby the study's impact [6, 4]. One of the issues of this strategy is the lack of interoperability between data sources, in particular when the studies do not follow the same principles for data collection and storage. For instance, the same procedure can have different designations depending on the institutions that collected the data.

Relational databases and NoSQL databases

Despite the fact that many solutions do exist nowadays, a common database can be in two different types: Relational Database Management System (RDBMS) and NoSQL [58]. Choosing the right Database Management System (DBMS) for each use case is important to optimize and ensure the longevity of developed applications. For instance, a RDBMS database is a better choice when scaling up the system vertically, whereas NoSQL benefits solutions that aim to be horizontally scalable. The structure used to store data in these databases is also different. RDBMS are more likely to have a less-dynamic data schema compared to the NoSQL engine, and sometimes this flexibility is actually not desirable [59].

The differences between both are not limited to data structure and scalability. More differences between the two could be discussed, however, to better understand the current techniques applied to anonymise data and how these can be integrated with our problem, we first need to establish which type of DBMS fits better in our scope. The variety of NoSQL paradigms would require an extensive review of privacy breaching techniques that may end out of the scope of this work. For instance, in document-oriented databases it is difficult to sanitize introduced data, therefore malicious queries can be introduced to manipulate the backend of NoSQL databases by adding, modifying or deleting data [60, 61]. This is an example of many that currently exist for the different NoSQL paradigms.

In health databases, the different data formats use specific systems for data storage. However, we aim to focus on the most common type of database used to support health researchers when using the data for conducting studies. RDBMS are the most

common database type used in EHR systems, although we recognize that these systems are not limited to relational databases. For instance, specific features may need to integrate additional components for storing data in a non-structured format, like repositories for medical imaging [62] or clinical notes [63], for caching data [64], among others.

Since the data integration strategies are directly influenced by the data format, in this work, we limited this topic to relational data, *i.e.*, data stored in relational databases. Although we restricted this discussion to one data type, there are still different strategies to solve the integration issue [65]. Independently of the strategy chosen, in the end, there is an interoperable layer that enables the data exchange between the different institutions involved in the study. This can be the adoption of a common data model or mapping the concepts to an ontology shared between peers. These strategies can be used to harmonise the data in distributed facilities, but in some situations in which the data are anonymised, researchers want to export the data and aggregate them before conducting the analysis. In recent years, several strategies have been investigated for performing clinical studies using heterogeneous data from multiple institutions. As a result of these efforts, several organisations, projects and tools were created.

Initiatives, projects and organisations

Informatics for Integrating Biology and the Bedside (i2b2) [66] was one of the first projects aiming to create tools to support clinical researchers in integrating patient data. One of its outcomes was a web application capable of performing cohort estimations and determining study feasibility using anonymised EHR data [67]. A common issue in this approach is the need to have the data centralised and accessible to platform users. However, the centralisation of health data from distinct institutions is complex due to legal, ethical and regulatory policies.

The Electronic Health Records for Clinical Research (EHR4CR) was a European project that aimed to improve the design of patient-centric trials [68]. Therefore, during this project, a platform was developed to support researchers in clinical trials' feasibility assessment and patient recruitment by accessing the existent EHR systems. The platform could perform queries in real-time using multiple clinical data warehouses across Europe containing anonymised patient data. The researcher obtained as output the aggregated results. Although the architecture provides a good solution to access multiple datasets, its success depends completely on health institutions joining the network.

The Health Maintenance Organization Research Network (HMORN) was another project focusing on creating a large-scale distributed network of health data [69]. PopMedNet¹ is an open-source application resulting from this project, designed to simplify the operations over distributed health data queries. However, like the previously described initiatives, it focuses on creating strategies to access the data, but not on developing a standard strategy to harmonise and anonymise patient information.

The OHDSI² had a similar goal. This international organisation aims to develop methodologies to support large-scale observational studies in health care data. This organisation was initiated as an outcome of the Observational Medical Outcomes Partnership (OMOP) project, to continue the research started on performing observational studies worldwide. Currently, this organisation supports an ecosystem with several open-source solutions to perform medical product safety surveillance using observational databases [4]. An example of such solutions is ATLAS³, a web-based platform to design cohorts and make a population-level analysis of observational data.

The EMIF⁴ project was inspired by the core principles of OHDSI and aimed to enhance access to patient-level data from distinct health institutions across Europe, including the possibility of conducting multicentre studies on different diseases [65]. A branch of this project focused on discovering and validating new biomarkers to diagnose Alzheimer's disease, in the pre-dementia stage, led to a track denominated EMIF-AD [70]. In this track, data from patients suffering from this disease were collected in multiple institutions, although in its first version, the data were collected for heterogeneous data schemas. In the final stages, some strategies were studied aiming to harmonise the data into a common data model. These strategies have adopted the OMOP CDM to store the patient data collected during these follow-up visits [7].

Interoperability between data sources

The interoperability between databases simplifies the distribution of queries between different institutions. This can be expressed in two types: i) syntactic interoperability, enables different applications to cooperate and communicate to exchange data; and ii) semantic interoperability, the ability of the system to share meaningful data [71, 72]. In networks of homogenous databases, interoperability can be easily accomplished

¹<http://www.popmednet.org/>

²<http://www.ohdsi.org/>

³<http://www.ohdsi.org/web/atlas/>

⁴<http://www.emif.eu>

when the same data schema is used by each database and data concepts follow a standard vocabulary. This would be the ideal situation, since data retrieval could be performed by querying the databases using the same Structured Query Language (SQL) query. Although this is the simplest scenario for sharing queries between databases, it is not always a possible solution.

A common scenario is the use of ad hoc strategies to retrieve data from heterogeneous databases. In health databases this is a time consuming issue, since researchers depend on the availability of the technical teams from each institution to retrieve the subset of information desired for conducting a study [73]. Therefore, in these scenarios there are some data management challenges on data placement, integration, and querying. The first challenge is focused on establishing which are the best data schemas and RDBMS engines. This cannot be generalized and requires a deep understanding of the data and query processing capabilities of the underlying storage. Although structured, for heterogeneous databases, this may require different processing capabilities during the ETL procedures [74].

The second challenge is related to data linkage when defining the query. The RDBMS has a defined data schema and in its design different components need to be explored to retrieve the desired results. Therefore, integrating the data from heterogeneous data sources may require additional processing. For instance, in the health care scenario it is necessary to have an additional understanding of the semantic information that can be related to specific ontologies, vocabularies and dictionaries [75]. The latter challenge is related to query placement in the different engines. Although in RDBMS databases queries are performed using SQL, this needs to be well-defined between all entities involved. Both the storing engine and data schema directly affect query construction when there are no interfaces to uniformise data stored in heterogeneous data sources, like a query builder [76].

Some strategies addressed these challenges by creating the concept of federated queries. This concept is focused on data integration models that combine the data into a logical structure, by providing a uniform view without moving the data [77]. Uniform views can be achieved with wrappers designed for each heterogeneous data source, as shown in Figure 2.1. In this case, each database would have an ad hoc wrapper, prepared to deal with its structure.

A different strategy to address this challenge requires the use of a homogeneous data schema, that would contain a replica of the information stored in the source

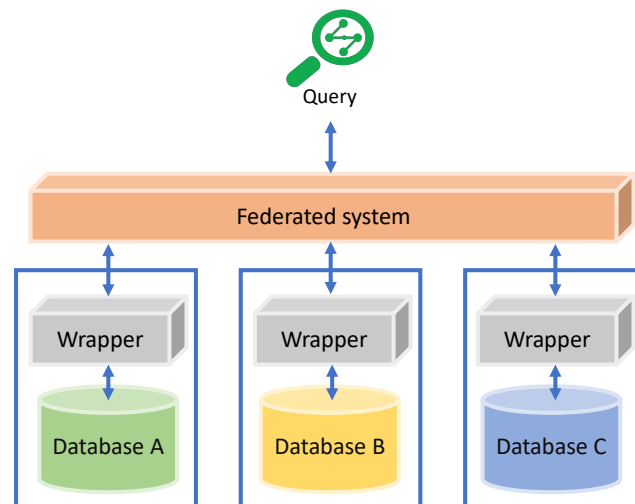


Fig. 2.1.: Federated integration architecture for providing uniform views over heterogeneous data sources.

database. This data schema should be interoperable with the network of databases and requires a ETL procedure to transform the original data into this new format. Application of this strategy is the use of health databases to conduct studies on the OHDSI⁵ community. OHDSI is an international organisation that aims to develop methodologies to support large-scale observational studies in health care data [4]. In these methodologies, they convert EHR data into a Common Data Model (CDM), and the research community can query this new format using a web-based query builder. Interoperability is ensured because they use their own standard data schema in the community. Therefore, new members of the community should migrate their data to this format and harmonise the medical concepts into standard vocabularies available for health concepts. The strategy is represented in Figure 2.2, where databases are converted to the OMOP CDM format so they can be then analysed using the analytic tools available on the OHDSI community.

Standard lexicons

In the literature, lexicons are mentioned as nomenclatures, vocabularies, ontologies and thesaurus. These are standards created aiming to encode textual information into a single definition. For clinical text, different lexicons have been proposed, from the most commonly used there is SNOMED CT [78], International Classification of Diseases (ICD) [79], Medical Subject Headings (MeSH) [80] and Unified Medical Language System (UMLS) [81].

⁵<https://www.ohdsi.org/>

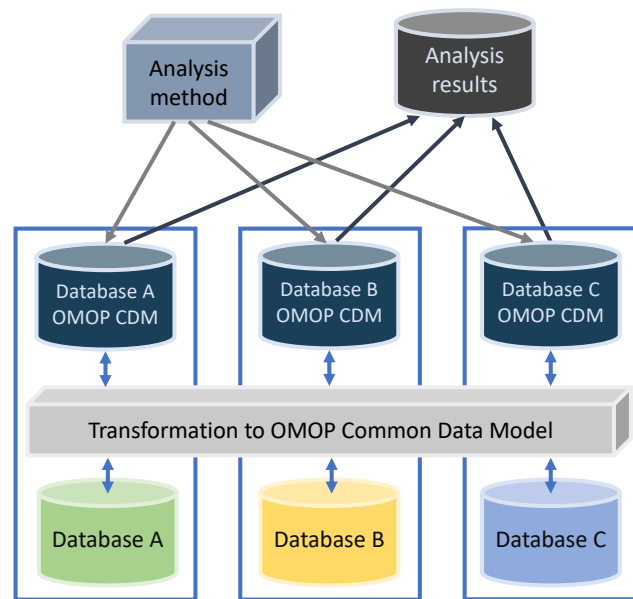


Fig. 2.2.: OHDSI methodology where the Data Owner convert the EHR data into a CDM database. The research community can query this new format using the same analysis methods.

SNOMED CT is a multilingual vocabulary of clinical terminologies managed by the International Health Terminology Standards Development Organization. It is an ontology-based lexicon targeted at clinical data in the EHR, that comprehends a large number of unique concepts. These concepts can map diseases, procedures, and medical findings, among other medical terms. The extensive range of unique concepts in this lexicon allows for more specific diagnosis when compared with other lexicons, like the ICD.

ICD is a hierarchical lexicon that was developed by the World Health Organization (WHO). It was created for the classification of vital statistics and it is in its 11th revision. This lexicon can be used for causes of death, cancer registration, dermatology, patient safety, primary care, pain documentation, allergology, reimbursement, clinical documentation, and data dictionaries for WHO guidelines in narratives. ICD was designed for classification, which is a key aspect explored for medical billing purposes.

MeSH is a curated medical vocabulary used by the U.S. National Library of Medicine, one of the world's largest biomedical libraries. This vocabulary is organised hierarchically and it provides terminology for cataloguing and indexing biomedical information targeted to the life sciences field.

UMLS is umbrella vocabulary that incorporates different lexicons, including the previously described and others. The main goal of this vocabulary is to facilitate the development of interoperable biomedical information systems and services. This can be achieved by providing mappings between different lexicons, expanding the list of existent codes, by considering synonyms that exist in other lexicons.

Although there are already a vast number of standard lexicons, in some situations, these are not used by physicians because the task of finding the correct code can be time-consuming, and the available concepts might not correctly characterise the situation under scrutiny. In these situations, physicians choose to use free-text, since it is easier to deal with uncertainties when writing the narratives [39].

Detailing the OMOP CDM schema

The major outcome of the OMOP project was the definition and dissemination of a CDM, which is a database schema to standardise the content of healthcare databases [82]. The original focus of this model was drug safety surveillance, but it was extended to many other use cases, such as quality of care, health economics and comparative effectiveness [83, 82]. The model accommodates standard definitions for patients' clinical data, allowing the use of federated queries across databases, enabling multiple and distributed analyses. Although this model is currently used for the data in EHR databases, we believe that its potential is not limited to this domain. Another outcome of OHDSI was the ETL procedures and tools defined in this context. These tools were specifically designed for EHR data, but they could be adapted for the proposed scenario.

The OMOP CDM data schema is divided into six groups of tables. This division is only to simplify the organisation of the data schema. From a computational point of view, the OMOP CDM is a single database structure. The complete OMOP CDM data schema is detailed at [28]. Figure 2.3 represents the tables in each of the following six standardised groups:

- Clinical data: contains the tables used for storing data directly related to the patient. Observations, medical procedures, measurements, drug prescriptions, among others are stored in tables associated with this group.
- Health system: a group of three tables with the information about the health institution, namely the healthcare providers, which are the individuals providing hands-on healthcare to patients, the institution location and other details.

- Health economics: group of two tables for capturing details regarding costs and health plan benefits.
- Derived elements: tables that store information about the clinical events of a patient that were obtained from other tables of the OMOP CDM.
- Vocabularies: structure of several tables designed to store in an interoperable format all standard vocabularies (lexicons) used in the OHDSI ecosystem. For instance, RxNorm, SNOMED, ICD10, among others.
- Metadata: tables with metadata about the current OMOP CDM version.

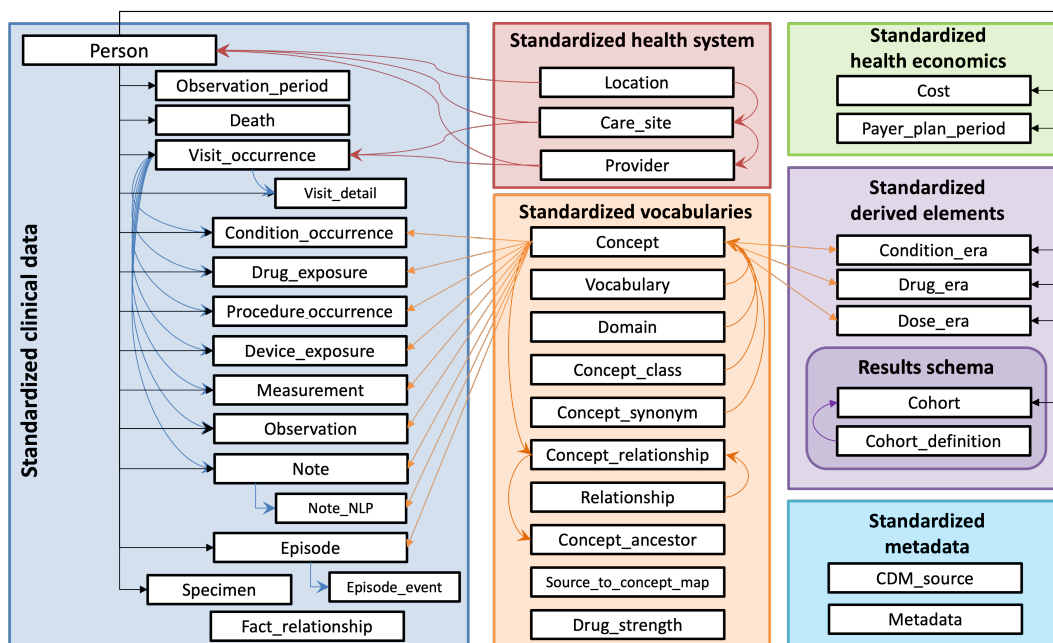


Fig. 2.3.: Diagram of OMOP CDM schema version 5.4. More information about these tables is available at [28].

2.1.3 Data analysis

Observational clinical studies typically are guided by protocols with several steps [84]. Based on the current strategies adopted to analyse EHR databases, the definition of a research study can be divided into seven main stages, as illustrated in Figure 2.4.

The first step is to translate the research interests into a precise question that can be addressed using data already collected in the past. For instance, a clinical diabetes researcher wants to investigate the quality of care that is delivered to patients with Type 2 Diabetes Mellitus (T2DM). This objective can be broken down into much

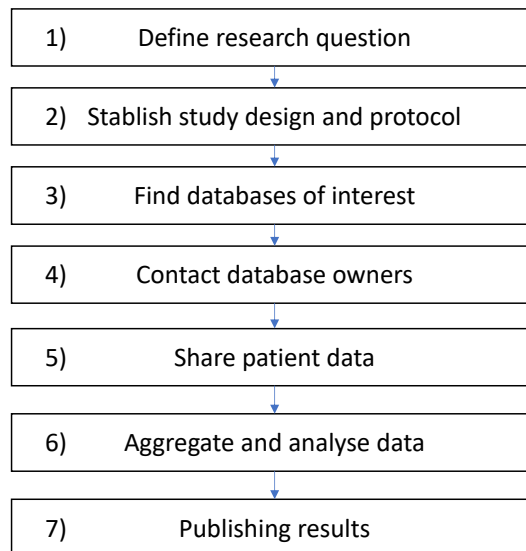


Fig. 2.4.: Overview of the study stages. From the idea to publishing the results, a study can be divided into seven stages.

more specific questions that may fall into a more precise category of studies. In a characterisation study, a researcher question can be defined as “do prescribe practices conform to what is currently recommended for those with mild T2DM versus those with severe T2DM in a given healthcare environment?” [28].

The second stage establishes the study design and protocol. In this stage, the inclusion and exclusion criteria are defined and described the expected study outcomes. The third stage aims to find databases of interest to support the study. It is crucial to understand the study’s feasibility. Sometimes these three stages are not complete in one round, requiring several interactions over them, until the researchers can find a study protocol that meets the initial objectives, ensuring enough data samples to possibly produce impactful findings.

The fourth stage is defined as contacting the database owners, recruiting their participation in the study, and providing access or information about their data sources. The fifth stage is based on information sharing and data aggregation, which we discuss in more detail in Section 2.1.3. The last stage is focused on publishing the results as a scientific contribution.

Data availability

Regardless of database interoperability, data privacy concerns are influenced due to data availability. For instance, a private and protected dataset may not need to

be anonymised. Health institutions have operational databases containing sensitive patient information that is not harmonised, such as the EHR systems. However, these databases are not exposed to the public and should not be accessible without permission. On the other hand, there are datasets with sensitive attributes that may require to be public, so these should be anonymised. In between these two sharing policies, there are other levels of data availability. Thus, we identified the most common strategies that are currently used for sharing data in different contexts.

The most commonly used approach to access data is through the Role-based Access Control (RBAC) mechanism. RBAC is used to segregate users into roles, and each role has a set of permissions for operating the resources, namely Create, Read, Update, and Delete (CRUD) operations [85]. This method is widely used in different domains, for instance, the previous example of accessing non-anonymised patient information is protected with RBAC mechanism, even when these databases are isolated in private environments. With this mechanism, medical staff can have different levels of access compared to the administrative staff.

A different level of access is found when the data is publicly available without constraints. Data released with this level of availability is usually anonymised or does not reveal sensitive information. For instance, the yellow pages are telephone directories where people's names, locations and phone numbers are discriminated. A similar situation occurs with voter lists that may contain a more detailed location of the subjects. In an isolated use, these datasets do not violate people's privacy. However, these can be used to cross information and break some levels of anonymity.

An intermediate level of access is when the data is made public with constraints. In these cases, subjects need to sign a declaration of honour commitment ensuring the adequate use of the data. This strategy is commonly used for studies and research purposes where the data owner wants to make the data available for a community, sharing sensitive information without revealing the owners' identity. However, this declaration only allows the researcher to use the data for research purposes, without trying to revert the anonymisation procedure.

Another strategy created for more specific scenarios involves the collaboration of three entities, namely the researchers, a study manager and the data owners. In this strategy, the researchers never access the data but may obtain responses to their answers, and the data owners never show the dataset. This methodology was implemented for conducting medical studies using distributed databases [65]. Figure 2.5 shows

all of the stages of this methodology. In the initial step, the researcher creates the study request in a platform implemented to moderate these procedures. This platform manages requests that are then analysed by the study manager. This entity defines the SQL query and coordinates its dissemination with the data owner. This dissemination can be orchestrated with the support of a work management system [18]. The fourth and fifth tasks are performed by each data owner involved in the study which would respond with the query results, an aggregation of these results or not respond due to the levels of data sensitivity. In any case, the data owner has full control of the data. In the sixth step, the study manager gathers the results and sends them to the researcher. A similar methodology was also implemented for accessing private biobanks [57].

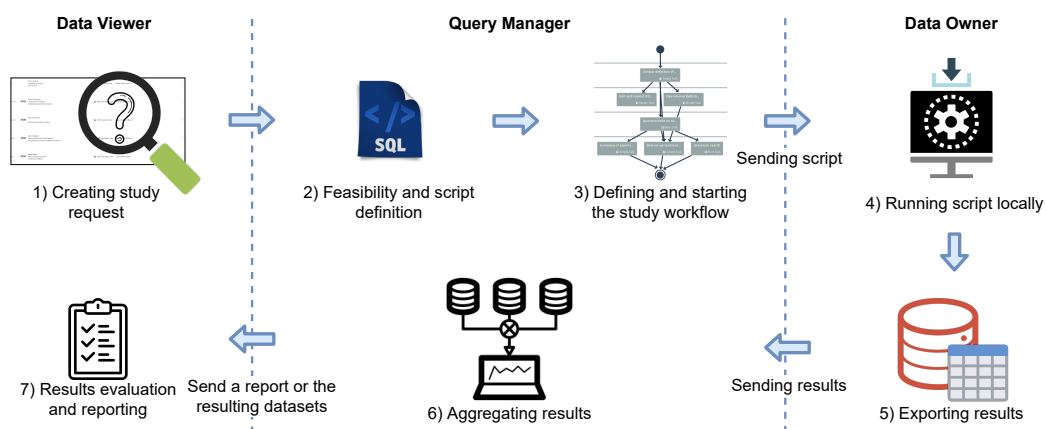


Fig. 2.5.: Methodology for performing distributed and moderated queries to sensitive datasets. In the first step, the Data Viewer creates the study request. This request is analysed by the Query Manager, which defines the SQL query and coordinates its dissemination between the Data Owner. The fourth and fifth steps are performed by each Data Owner locally. In the sixth step, the Query Manager gathers the results, and send them to the Data Viewer.

The strategy used to share the data and make it accessible influences the procedure for anonymising the data. Although it would be desirable to have all the datasets publically available, this may not be possible in the health domain.

Streamlining studies

The process of recruiting data owners, distributing the research questions and study protocol between them, and gathering the data results can be complex. In this section, we describe some of the current strategies applied in the healthcare domain. We prioritize strategies that are already applied to the OHDSI community, without discarding others that can provide some contributions to this work.

Fajarda *et al.* [86] proposed a semi-automatic methodology to perform distributed queries over EHR databases that are part of the OHDSI network. In this work, the authors used a database catalogue where: i) the metadata about each database is exposed; and ii) researchers can identify databases of interest and perform a research question. However, this methodology relies on two more actors, one for managing the study after receiving the research question, and the data owner, who should execute the query manually against the database.

Still in the OHDSI community, another proposal was made which streamlines the process but maintains the same actors in the pipeline. This process is supported by the ARACHNE tool, which is a platform designed to automate the process of conducting network studies for this community. ARACHNE supports the OHDSI standards and establishes a research procedure to conduct observational studies across multiple organizations [84]. In both approaches, the databases need to be compliant with the OHDSI principles, including having the data migrated to the OMOP CDM format. This format is currently widely adopted by several institutions worldwide.

Topaloglu *et al.* [20] proposed TriNetX, a network of healthcare organisations to optimise clinical trials. In this ecosystem, the authors proposed a mechanism for querying the databases, including some security aspects. However, this requires the adoption of their principles, and these are not widely spread as OHDSI.

All of the proposed solutions try to query distributed health databases. However, the efforts to anonymise data are mainly focused on removing the identifiers, but none of these works have performed a deep analysis to identify privacy metrics of the data that is released to the researcher. Although these solutions address the problem of querying distributed databases, they present some limitations, namely regarding data sharing and privacy preservation. This work proposes to evolve these solutions, by facilitating data access and respecting confidence levels of privacy.

Privacy-preserving at data publishing

Apart from the data domain, the scientific community is constantly studying new strategies and models to work with subjects' data without violating their privacy. Although the greater risk of violating subjects' privacy is centralized in data publishing, data policies such as privacy, data security and intellectual property should be addressed in all phases of dealing with the data (*e.g.*, data collection, processing, anonymisation, sharing and analysis) mainly when dealing with person-specific data [87].

The dataset schema may reveal information about a group of individuals, namely the relation between the sensitive attributes. For instance, a released dataset with data collected for a health study on a specific disease can contain specific attributes from this domain, showing that every individual in that dataset may suffer from the given disease. Although this information does not violate the privacy of a specific individual, it reveals that in a specific region, exists at least that given number of subjects suffering from that disease.

The decentralization of processing methods would enable the usage of distributed databases without significantly affecting the subjects' privacy, which could benefit the development of Artificial Intelligence (AI) strategies [88]. However, this has some limitations and it does not help when the goal is to obtain data to be analysed in local settings. Therefore, considering the possible application of data sharing, the trade-off between privacy and utility remains a challenge in Privacy Preserving Data Publishing (PPDP).

The future of computing paradigms should benefit from big data analysis. In the health care data domain, this is already a reality, where studies have shown the benefits of using distributed data collected in distinct organizations, which improved the health research findings [65, 7, 4]. Despite these advantages, there are numerous barriers to widespread the adoption of these approaches, such as security concerns, implementation issues, privacy concerns, and technological fragmentation. Among these, subjects' privacy has significantly hampered the development of new strategies for data analysis, mainly in the health domain. More recently, the community has invested some effort in creating strategies to overcome this issue and to be used for privacy protection in future computing paradigms, namely through homomorphic encryption, federated queries, data harmonisation and anonymisation, among others [89, 90, 91].

Achieving effective privacy should be done by focusing on exploring the intrinsic characteristics of the subjects' data for the target data domain [92, 93]. To attain better privacy protection and data utility, the use of ad-hoc approaches derivated from the standard techniques has become more emergent than ever [94]. In the health care domain, the community has made some efforts to find solutions capable of analysing distributed databases without violating the subjects' privacy, through the creation of ecosystems to explore the data in privately protected environments, and only releasing aggregations of the processed data [73].

2.2 Research questions

The research objectives of this work can be paraphrased into several questions focused in addressing specific problems. We recognize that multicentre medical studies may raise additional issues that will not be considered in this work, mainly due to the data types used on these. For instance, research studies based on biomarkers that are correlated with DNA information, or studies primarily focused on using medical images. Therefore, to achieve a scenario capable of supporting distributed health studies using multiple data sources from distinct institutions, we limited the scope to EHR databases. This objective will be accomplished by answering the following research questions:

1. *How to execute a database query over a network of heterogeneous health databases?* The lack of interoperability is the main problem in this scenario. An ecosystem with heterogeneous databases may not share the same data schema, which invalidates the query-sharing. Besides this problem, the healthcare domain contains huge amounts of medical concepts, that may differ between institutions, at the national or international level. A methodology to harmonise such databases is required, which converts them into an interoperable format. This process may have different stages and components, and some of them will be automatized aiming to reduce the cost and time of executing this procedure. This may result in a new software solution.
2. *How to enrich patients' health history using information available in clinical notes?* The free format of this type of information raises several challenges in terms of named-entity recognition and normalization. Currently, the research lines focused on NLP, may lead to new solutions that can enrich this field. In this work, we aim to investigate a solution capable of improving the quality of information stored in the relational databases. This solution will be a software solution capable of processing clinical free text and storing this information in a relational database, following the harmonisation principles defined in the ETL procedures for EHR databases.
3. *How to select the most adequated health databases for a specific research study?* This question can be addressed from different perspectives. However, by correlating it with the previous statements, we identified that the main problems medical researchers meet are: i) the discovery of databases of interest; and

ii) the access to those databases without violating privacy policies and ethical regulations. The solution to these problems is too complex to be solved by a software application alone. To answer this question, it is necessary to create an ecosystem of tools and methodologies, in which data owners can feel confident in sharing characteristics about their databases, while researchers can have enough information to select the databases that fit better their needs. Therefore, the solution proposed for this problem will be a portal that integrates: i) a web catalogue of database characteristics; ii) tools to visualize and compare these characteristics; and iii) tools for orchestrating distributed studies.

The software solutions proposed to answer these questions will be developed aiming to be used by non-technical users. Although during their development, complementary tools may arise that does not completely match with this requirement.

2.3 Hypotheses

The results of the research made to answer the presented questions is better detailed in the following chapters, which are based on the following hypotheses:

1. *How to execute a database query over a network of heterogeneous health databases?*
 - Current initiatives have already shown that using a common data schema to store data extracted from EHR databases allows the reproducibility of the same cohort using data sources from different institutions. This is crucial to support multicentre studies. We expect that optimizing the ETL pipelines can reduce the harmonisation costs, resulting in more institutions adopting this strategy.
 - Among the many strategies possible to have a network of interoperable databases, the adoption of OMOP CDM in this work may lead to more impactful results. Besides, OHDSI already defined some ETL guidelines and supportive tools, however, there are still a lot of possibilities to contribute to the automatization of pieces of this pipeline.
2. *How to enrich patients' health history using information available in clinical notes?*

- Clinical notes are known for being used by physicians to document complete descriptions of patients' medical status. Although some of the information in these notes is redundant with the EHR system, some annotations about the patient history are only kept in this format [40]. We presuppose that the information stored in these notes may enrich the data stored in the relational databases. We assumed that having this information in a structured schema would bring more value than being only in free text format.
- The OMOP CDM schema already contemplates the possibility of gathering clinical notes from the institutional repository and storing their information into tables specifically created for this end. We hypothesize that using OMOP CDM would increase the impact of this task namely due to the efforts made by the community in this context and for being the data schema used to store the data present in the EHR from the previous task.

3. *How to select the most adequate health databases for a specific research study?*

- Several initiatives have identified the need of simplifying the discovery of biomedical data sets. This is especially relevant to select the most appropriate databases when conducting a study. However, we identified some issues in the characterisation of such databases. We assumed that optimising the process of profiling such databases and exposing them to a web catalogue would provide more valuable information to be exposed.
- When dealing with medical data, privacy issues arise due to the data being sensitive. Database profiling can interfere with this topic if some principles are not followed. Considering these issues would increase the confidence of data owners when sharing the database characteristics. We deduced that only exposing aggregates of information would help researchers to find suitable databases for their studies while data owners do not violate any ethical policy.
- The process of conducting the study from planning to obtaining the aggregated results is time-consuming and can demand high costs for the research institution. Therefore, to enhance this process, we hypothesise that providing tools and methodologies to support researchers would be helpful and create value for the community.

2.4 Means of verification

The validation of the proposed solutions will be made independently. Each hypothesis will have specific means of verification. Since this work will be conducted in close collaboration with European research projects, we aim to use real cases to validate each tool. Therefore, we expect to use patient data to test and validate the ETL tool developed to answer the first research question. To validate this approach, we aim to use data sources that belong to one of the projects we collaborate on, namely EMIF, EHDEN, or EMIF-AD projects.

We expect to validate the NLP methods proposed to answer question number two, using public datasets provided in scientific challenges, namely organized by National NLP Clinical Challenges (n2c2). With these, we can test and validate tools using datasets annotated by medical specialists.

Finally, we aim to validate the solutions proposed for research question number three within EMIF and EHDEN projects. We expect to profile and expose health databases from the data partners involved in these consortia.

Semi-automatic translation of data sources into a common schema

Many clinical trials and scientific studies have been conducted aiming for better understanding of specific medical conditions. These studies are often based on a small number of participants due to the difficulty in finding people with similar medical characteristics and available to participate in the studies. This is particularly critical in rare diseases, where the reduced number of subjects hinders reliable findings. To generate more substantial clinical evidence by increasing the power of the analyses, researchers have started to perform data harmonisation and multiple cohort analyses. However, the analysis of heterogeneous data sources implies dealing with different data structures, terminologies, concepts, languages and, most importantly, the knowledge behind the data. In this chapter, we present some methodologies created in the context of this work to migrate and harmonise heterogeneous datasets into a standard data schema.

An unprecedented amount of data is being generated in many economic sectors, *e.g.*, the advent of industry 4.0 and e-health paradigms, raising new challenges regarding data collection [95]. The potential value of all these data also led to an increasing interest in big data solutions and data-driven decision-making tools [96]. Despite many efforts in this area already, especially in deep learning algorithms, the process of converting different data sources into a heterogeneous and interoperable repository still has some issues related to data complexity, scalability, timeliness, and privacy policies [97].

Big Data is usually defined as the daily basis production of large volumes of either structured or unstructured data. These volumes of data are noisy and contain a significant amount of invalid or corrupted records that should be discarded [98]. These cleaning operations are usually performed following ad-hoc approaches, which are difficult to generalise. Another challenge is the amount of unstructured data that despite containing valuable information needs to be structured to enable the usage of analytical methods [99]. The major challenge in this topic can be data integration when considering multiple heterogeneous data sources. Overcoming these challenges can lead to improvements in several business domains [98, 99].

A strategy adopted by some organisations to analyse internal data or external data sources is based on the use of Business Intelligence (BI) tools [100]. BI is a domain that incorporates applications and methodologies aiming to collect, prepare and explore data from diverse sources of information. These tools enable the analytical exploration of data, which can result in reports and dashboards for data visualisation [101]. This concept is focused on access to and exploration of heterogeneous data sources in order to have more information about the business and to make better informed decisions [102].

This lack of flexibility in users' collaboration during the design and definition of the ETL pipelines is a problem for some application domains. For instance, in the medical scenario, when clinical data needs to be harmonized into a common data schema, this requires collaboration between the technical teams and the medical researchers [73]. This collaboration is required in different stages: i) design; ii) implementation; and iii) validation. In each stage, there are some challenges that we addressed in this work.

3.1 Contribution

In this chapter, we describe strategies and solutions to support the translation of data sources into a common data model, mainly in medical scenarios, by proposing:

- A methodology to harmonise disease-specific cohorts, by storing data in a standard common model, and mapping clinical concepts to a normalised representation. The data schema is being used to harmonise EHR databases in observational studies at a world-wide scale, enabling the leveraging of previous knowledge and open source tools to perform multi-centric and disease-specific studies [73]. The methodology was implemented in Python language and is available, under the MIT license, at <https://bioinformatics-ua.github.io/CMToolkit/>;
- A solution for migrating patient's information from OMOP CDM databases to tranSMART structure. This solution was implemented in Python language and is available, under the MIT license, at <https://github.com/bioinformatics-ua/tranSMART-migrator>;
- A collaborative web-based ETL application that allows users to design, share and execute ETL pipelines, across multiple centres. The system is supported

by a user-friendly interface in which non-technical users can build the ETL pipelines without the need to grasp the ETL details, and most importantly, without having direct access to the data. This tool is available, at <https://bioinformatics-ua.github.io/BICenter/>;

- An evolution of BICenter focused in the context of Alzheimer's Diseases to support the collaboration between distinct entities in the definition and implementation of ETL pipelines. These pipelines are constructed using drag-and-drop features and intuitive forms to customise the ETL steps. This tool is an open-source project and is accessible at <https://bioinformatics-ua.github.io/BICenter-AD/>.

Therefore, this chapter is mainly based on the following publications:

- **João Rafael Almeida**, Luís Bastião Silva, Isabelle Bos, Pieter Jelle Visser and José Luís Oliveira, *A methodology for cohort harmonisation in multicentre clinical research*, Informatics in Medicine Unlocked, 2021, DOI: 10.1016/j.imu.2021.100760;
- **João Rafael Almeida**, Leonardo Coelho and José Luís Oliveira, *BICenter: A collaborative web ETL solution based on a reflective software approach*, SoftwareX, 2021, DOI: 10.1016/j.softx.2021.100892;
- **João Rafael Almeida**, Luís Bastião Silva, Alejandro Pazos and José Luís Oliveira, *Combining heterogeneous patient-level data into tranSMART to support multicentre studies*, in proceedings of the IEEE 35th International Symposium on Computer-Based Medical Systems, 2022, DOI: 10.1109/CBMS55023.2022.00018;
- **João Rafael Almeida**, Alejandro Pazos and José Luís Oliveira, *BICenter-AD: Harmonising Alzheimer's Disease Cohorts using a Common ETL Tool*, Informatics in Medicine Unlocked, 2022, DOI: 10.1016/j.imu.2022.101133.

3.2 Background

BI is a concept used as a hypernym that covers the domains of Data Warehousing (DW) [102], which is the consolidation of data from heterogeneous sources and it can define the foundations of BI methodologies. Most large and medium-sized organizations are currently adopting DW systems to support their BI tools [103].

The core of BI is based on two components that directly support decision-making, namely the Online Analytical Processing (OLAP) and Enterprise Information Systems (EIS). In some cases, these two components should be able to provide a minimal solution of a BI application. However, to comprise various concepts and applications, ad-hoc reporting or text-mining components can be added. Besides these, more advanced features can be included, such as an analytical Customer Relationship Management (CRM) [104].

The process of integrating and transforming the data into a data warehouse is time-consuming and requires human validation, independently of the data domains [103, 105, 106]. This integration process, usually denominated as ETL, is a workflow aiming to collect the raw data and process them through three distinct stages: i) Extraction, where the data are accessed from heterogeneous sources; ii) Transformation, which manipulates and converts the loaded data into the desired form; and iii) Loading, to store the resulting data into the target database.

These operations are processed at the database level and they can be coded using a programming language. With the growing complexity of these procedures and the need to involve multiple entities in the design of these workflows, more user-friendly approaches were created [107]. Some of these approaches are focused on documenting the process, while others were specially designed to have a Graphical User Interface (GUI) to specify the ETL workflows [108].

3.2.1 Most common ETL tools

There are a large number of ETL tools available in the market created for different purposes. Some vendors have open-source solutions while others only have commercial options available. Majchrzak *et al.* [104] made a evaluation of open-source ETL tools based on their efficiency. The indicators selected by the authors were derived from ISO 9126 norm [109] and specific literature with a focus on measuring ETL performance [110]. Following the selection criteria described in this study, only two tools were considered. Since this study was conducted in 2011, new tools were released. Thus, we decided to include these in the study, adopting the same criteria to compare them.

This analysis aims to reveal which tools are compliant with a set of criteria and that can be used to support the ETL workflows behind the methodologies we proposed

during this work. Table 3.1 summarizes the classification of selected open-source tools. The criteria used to classify these were the following:

- **Connectors:** The tool contains interfaces to connect to the most relevant systems, including operating systems and DBMS. These interfaces are typically found as data sources that need to be supported.
- **Documentation:** The tool has available reliable documentation. This should include the information necessary to contribute to the tool development, by extending specific features, or simply using it as an end user.
- **Graphical editor:** The GUI for modulating components is important to modulate the ETL processes. Classifying the existence of this feature is important when non-technical people are required to design ETL processes.
- **Integration:** It evaluates the capability of integrating the ETL processes into the existing systems.
- **Support:** It relays on the support provided by the vendors or community, when available.
- **Third-party:** Besides the basic ETL features already available in the tool, it classifies the possibility of integrating third-party libraries.
- **Updated:** It considers the current development status of the tool, namely if it is currently used or if active developers are contributing to maintain and improve the tool.

Some of the tools were discarded because although they were announced as ETL tools, they do not support data transformations. An example of these tools is Airbyte, which is a robust open-source tool for data integration, but it does not support data transformations. Another issue is the timeline for when some of these tools achieve a mature and stable stage. For instance, Apache NiFi is currently a promising ETL tool, but when this work started, it was in its early releases, which discouraged us adopting it.

Tab. 3.1.: Criteria fulfillment by the candidates

	Connectors	Documentation	Graphical editor	Integration	Support	Third-party	Transformations	Updated
Airbyte ¹	✓	✓		✓	✓	✓		✓
Apache NiFi ²	✓	✓	✓	✓	✓	✓	✓	✓
Aptar Data Integration ³	✓		✓		✓	✓	✓	
CloverETL ⁴	✓	✓		✓		✓	✓	
Jitterbit Integration Environment ⁵	✓		✓	✓	✓	✓		✓
KETL ⁶	✓			✓		✓	✓	
Pentaho Data Integration (PDI) ⁷	✓	✓	✓	✓	✓	✓	✓	✓
Scriptella ⁸	✓	✓		✓		✓		
Singer ⁹	✓	✓		✓	✓	✓	✓	✓
Talend Open Studio (TOS) ¹⁰	✓	✓	✓	✓	✓	✓	✓	✓

Metkewar *et al.* [111] conducted a comprehensive survey of ETL tools aiming to identify the strengths and weaknesses of the most used ETL tools at that moment. More recently, Gina *et al.* [108], published a literature review of critical factors that drive the selection of BI tools. When considering open-source tools, Talend Open Studio (TOS) and Pentaho Data Integration (PDI) are by far the most relevant options currently available on the market. However, when considering commercial solutions, Informatica PowerCenter and IBM Infosphere Data Stage are the most popular.

PDI is an open-source BI application that provides a wide range of features to support ETL workflows. This tool is also known as Kettle. It provides a graphical editor (designated as Spoon), where users can build data integration procedures. The procedure, also known as transformations, can be run by Kettle using different interfaces, namely: i) command-line utility (Pan or Kitchen); ii) remote servers (Carte); or iii) directly from the Integrated Development Environment (IDE) (Spoon).

¹<https://airbyte.com/>

²<https://nifi.apache.org>

³<http://www.apatar.com/>

⁴<http://www.cloveretl.com/products/community-edition>

⁵<https://www.jitterbit.com/platform/>

⁶<https://sourceforge.net/projects/ketl/>

⁷<https://www.pentaho.com/>

⁸<https://scriptella.org>

⁹<https://www.singer.io>

¹⁰<https://www.talend.com/products/talend-open-studio/>

PDI follows a meta-driven approach, which exploits the use of data dictionaries to automate the ETL management and accelerate the development of new ETL workflows.

TOS is another open-source ETL tool with the support of data integration. It uses a different approach compared with PDI. Rather than using meta-data driven, it relays on code-drive approaches. This tool also has a user-friendly GUI for user interaction, similar to Spoon. The property responsible for generating code supports Java or Perl programming languages as output, which can be then executed on a server.

Both PDI and TOS have strong community support, as well as they represent the most deployed open-source ETL solutions. Both tools are very reliable and real-world enterprises have used them to support practical implementations. Although they are very similar, TOS is more focused on data quality and management, while PDI seems to be more focused on BI.

Based on all these studies, we identified that the most relevant, open-source and complete tools aiming to simplify the design and creation of ETL processes are TOS and PDI [111, 104, 107]. Biswas *et al.* [107] studied alternative ETL approaches, namely focused in custom-coded solutions without GUI. These authors present a comparative evaluation of these code-based solutions. Although some code-based ETL applications may present in general a lower implementation cost and effort to maintain, specific domains may demand the support of non-technical members to design and implement the ETL pipelines. Therefore, in such cases, it is essential to have a ETL solutions with GUI to specify the ETL workflows, inclusively in the health domain.

3.2.2 Mapping concepts

One of the critical issues in the ETL processes when dealing with medical data is in the transformation stage. The ETL tools previously described may provide some support when transforming the data from the original data schema into the final outcome. However, due to the high number of medical concepts that need to be mapped to their standard definition, these tools may require additional features to simplify these mappings. The target mappings are standard medical lexicons, as it is explained in more detail in Section 2.1.2.

An approach often used to tackle this problem focuses on using ontologies to represent semantically the data from different systems [112]. The adoption of ontologies can optimise some tasks, but the code mapping is not dependent on the ontology. Instead, it may require software to support the manual mapping, or use automated algorithms to perform this task [113]. The works present regarding concept annotation or code mappings are often based on NLP approaches, which increments an unnecessary step in the workflow. Therefore, there are few tools for concept mapping, that could support medical researchers to correctly annotate concepts into their standard definitions.

AutoMap is a tool that aims to automatically map medical codes across different EHR systems. It constructs target embeddings unsupervised based on the target EHR data, mapping them against the source embeddings. This feature allows the quick deployment of the pre-trained deep learning model in the target system, without manual code mapping [114]. Although this promising system, it was not available at the beginning of this work for being too recent, as well as, it performs the mappings without the validation step.

Usagi¹¹ is a tool to support the manual process of mapping concepts to their standard definition. It can automatically suggest mappings based on the textual similarity of code descriptions. If the source codes are only available in a non-English language, the user needs to translate them using an external resource, *e.g.* using tools as Google Translate. Additionally, Usagi contains searching features, to help users to find the appropriate target concepts when those are not correctly suggested. The user can indicate which mappings are verified and manually approved, so these can be used in the ETL pipeline [28].

3.3 Methodology for cohort harmonization

The proposed methodology reuses as many open-source tools and methodologies as possible, avoiding the development of new ones with similar goals. Therefore, we adopted some of the OHDSI tools and principles in some components of our methodology. Regarding the data schema to store the migrated cohort, we used part of the OMOP CDM without any changes in its structure, as keeping the data schema as it is may increase the interoperability between the databases created from cohorts and the EHR databases, if necessary. This interoperability is ensured because the

¹¹<https://github.com/OHDSI/Usagi>

data schema was not adapted for the cohort scenario. Instead, we tried to fit the information into the existing tables. Therefore, it is possible to use the same analytical tools used for exploring the EHR data migrated to OMOP CDM.

We also used some of the ETL supportive tools from OHDSI, which we adapted for the cohort mapping scenario. Although this was a good starting point, we felt the need to use a collaborative platform to manage those mappings through a semantic ontology. This ontology characterises all the elements involved in the vocabularies with extra information and organises them by their relationship with each other.

3.3.1 Overview

The proposed methodology is based on ETL principles. Therefore, in the extraction stage, the selected source data are read by pulling them from one or several data sources. The main goal of this stage is to get the data from the source systems without interfering with their usual performance. In health databases, this is a sensitive task because the EHR cannot be overloaded due to the data extraction procedure. However, in clinical studies, the amount of data is not sufficient to crash the systems during this stage. Furthermore, clinical studies were exported to a tabular format, which do not require direct interaction with the system used to collect the patient data.

The transformation stage is the most complex component in this pipeline. This stage requires the mapping of the source database into the target schema, as well as the harmonisation of the content. For a data source, this procedure requires a full mapping, which is time-consuming and requires specialised entities to validate the mappings. Content harmonisation could have custom operations over the data based on the source of the data. In clinical databases, there is a wide variability of clinical concepts that need to be harmonised using standard vocabularies. Although we were able to automatise parts of this stage, we still require manual validation by a specialised health professional to ensure that all mapped data are correct.

Finally, the loading stage inserts the processed data into the target database, which can be then accessed using analytical tools. Clinical databases are populated with pseudo-anonymous data, allowing clinical studies to be conducted without violating patients' privacy rights. Additionally, when the data are migrated to a standard data schema, the original data end up being validated, and inconsistencies can be found in the source database. This is possible due to the quality mechanisms that were created

in the pipeline, which are responsible for checking whether loaded data respect the rule attributes for each standard concept.

3.3.2 The cohort common data schema

One of the key points in cohort harmonisation is the use of a common data schema for data storage, such as the OMOP CDM. This standard data schema, which is continuously being improved by the OHDSI community and serves as the base for the observational databases in this community, has an excessive number of tables for the identified problem of cohort harmonisation, mainly because this model was designed to extract data from EHR systems. However, clinical studies focused on a disease only need a small part of this schema to store information.

The complete OMOP CDM data schema is detailed at [28]. This data schema was optimised for observational research purposes and the tables and the field of each table were defined by the OHDSI community. However, our approach relies on the set of OMOP CDM tables presented in Figure 3.1, without changing their relations and structure.

The “Person” table stores the patient’s personal information, *i.e.* gender, date of birth, race and ethnicity. The “Observation” table maintains all the observational data collected during the study. Each entry in this table contains: i) a numerical entry for patient identification, which is only used in this database; ii) the standard code for the observation concept, *i.e.* specific exam conducted during the patient’s visit; iii) the standard code for the observation type concept, which characterises the measure/exam done on the patient when it can be represented using a standard code; iv) the date and value of the observation. This value can be characterised by its type, *i.e.* it can be numeric, text or a code. The “Observation Period” table contains the time interval when each patient was under observation.

The OMOP CDM has a set of tables belonging to the “Standardized Health System Data” group. Therefore, we also used the “Care Site” and “Location” tables to store information about the institution where the clinical study was made. Additionally, we used all the tables from the “Standardised Vocabularies” group to store the standard concepts’ dictionaries. This data schema is created and the database is loaded in the third stage of the workflow, namely the loading stage.

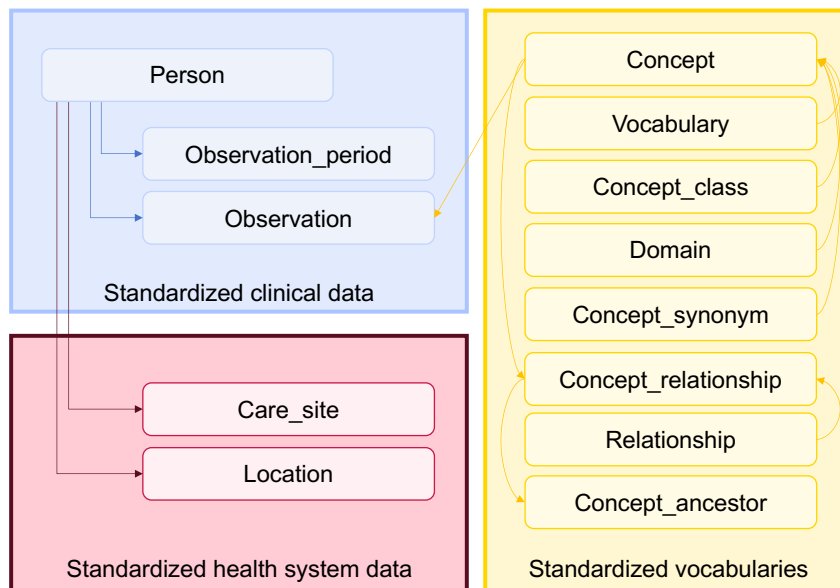


Fig. 3.1.: Tables used from OMOP CDM schema in the proposed methodology. The complete data schema was presented on Section 2.1.2.

3.3.3 OHDSI ETL tools

OHDSI provides a ETL toolkit that was an excellent starting point to harmonise the data recorded in the clinical studies, mainly because they were developed to handle clinical data independently of the data format. In the proposed methodology, we used some features of White Rabbit and Usagi for the extraction, harmonisation and mapping of the patients' clinical data. White Rabbit scans the data source and creates a structured report with all the information about the database content. Usagi is a complementary tool that receives some of the information available in this report to map the concepts with their standard definition.

Cohorts follow a spreadsheet structure because this was the export format usually used in the institutional systems, or in some cases the way that the data were recorded. Using White Rabbit in the extraction stage of the methodology workflow, we can have an overview of these datasets, namely the different records made in the clinical study and some statistical representation of their content. The report produced by this tool helps identify some anomalies in the data in a first glance. This report is also used as input in some components during the different steps of our methodology.

Our adaptation of the Usagi tool plays two different roles in our proposal. One is concept mapping, which is similar to the original goal of this tool. In this way, we can map the study columns and observations into the standard vocabularies. The other

role is to map the cohort structure into the OMOP CDM data schema. This tool is a core component of the transformation stage of our workflow.

3.3.4 Collaborative ontology development

The mapping of concepts to their standard definition simplifies its name recognition by medical experts and also the identification the same concept in different cohort studies. This procedure refines the data existent in the dataset by discarding unmap-ped concepts, but the raw cohort data contains more patient information that is not directly present. Depending on the clinical study scenario, *i.e.* the diseases or health effects in the study, the observations can have additional meanings. Traditional ETL typically extracts the data, converts them into the source target schema and loads the data in a new data schema. This is very efficient when applied to data with a static and well-organised structure [115].

In the proposed scenario, we have additional information that needs to be annotated during the transformation. In a very simple example using two common measure-ments such as weight and height, we can calculate the patient's body mass index. As a result, when this value is above 30, the patient is classified as obese which means that the patient has a cardiovascular risk factor that can be classified as a comorbidity [116]. This example, only based on the patient's height and weight, shows how much information is in the raw data that can improve the efficiency in the patient selection stage in clinical studies. There is more information that could be extracted in this way, but the teams responsible for designing ETL mappings are not able to infer this.

We rely on WebProtégé [117] to build and keep updated the ontology applied in our harmonisation workflow and to add this semantic information to the dataset. This web platform facilitates collaboration between the clinical experts involved in the project, leading to the definition of a disease-specific ontology in the Alzheimer's Disease domain. In the end, we were able to obtain a structured semantic ontology containing properties to infer knowledge by correlating fields during the migration, and as an additional feature, other properties to validate the input information in each concept. This ontology is used in the transformation stage of the ETL workflow.

3.4 The cohort migrator toolkit

The proposed methodology was implemented in Python using the adaptations of the previously described tools, and is publicly available, under the MIT license, at <https://bioinformatics-ua.github.io/CMToolkit/>. This methodology includes the stages of the ETL operations, *i.e.* the workflow from the cohort's raw data into the OMOP CDM database is divided into three stages, as presented in Figure 3.2. In this implementation, we split these stages to enable their execution in an isolated manner.

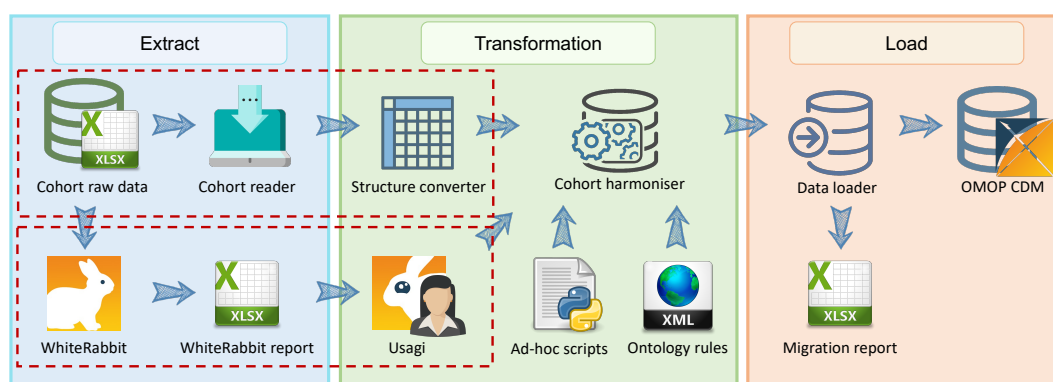


Fig. 3.2.: The migration workflow from raw data to the OMOP CDM structure, using the proposed methodology combined with the ETL OHDSI tools. This workflow is divided into three main stages, having two processes running in parallel (marked in a red dashed line). The first stage extracts cohort information and loads it into the system. The transformation stage performs all the defined operations over the raw data using the mappings mixed with the ontology rules. Finally, the loading stage inserts the data in the database, producing a migration report which indicates all the problems with the original raw data.

The WhiteRabbit, represented in the extraction stage, provides fingerprinting of the cohort structure. Concurrently, the cohort reader loads the data into a pre-transformed format. These two outputs are used in the transformation stage, following a parallel flow. Usagi reads the WhiteRabbit outputs and generates the mappings to be used by the Cohort Harmonizer. This main block centralizes a set of operations to generate an output file that can be exported to CSV files or a database using the OMOP CDM loader in the loading stage.

The implementation of some components raised some challenges due to the data sensibility of the proposed use case. When dealing with medical data, it requires deep knowledge of the data source, in order to perform the harmonisation correctly. Another challenging task was the custom operations in each cohort's raw data, *i.e.*, when the data were collected, it did not follow any standard strategy. This lack of

interoperability when recording the data complicated the implementation of the migration workflow. The data loading into the OMOP CDM and quality assurance of the created database was another task that raised some challenges.

3.4.1 Data harmonisation

Data harmonisation is the most important task in this workflow and consequently, this stage was split into several steps, which work in parallel. As shown in Figure 3.2, there is a pre-processing component to access cohort data stored in the temporary structure and to reorganise that information based on the patient follow-ups. This component creates a key-value structure where each measurement is represented with all the patient and time information. This structure includes the patient's measurements, the date of the exam and the follow-up visit number, which also represents the length of time from the first visit.

The key-value would contain as key: i) the patient identifier; ii) an attribute such as the visit date; and iii) the exam or cohort attribute. The value would be the entry for that attribute and in the next interaction, the standard concept codes for this entry and the attribute. Figure 3.3 illustrates an example of the cohort raw data (first table) and its structure in the format processed during the workflow (table below).

Patient ID	Visit date	...	Anmial Fluency 1 min	...
10424	15-01-2013	...	16	...
10424	24-02-2013	...	20	...
...

Patient ID	Visit date	Original exam	Value	Harmonised Exam	Harmonised Value
10424	15-01-2013	Anmial Fluency 1 min	16		
10424	24-02-2013	Anmial Fluency 1 min	20		
...		

Fig. 3.3.: Example of cohort raw data (first table) and its structure in the format processed during all the workflow. The blue box represents the key of the key-value structure, and the green box represents the fields that would receive the harmonised concept codes.

The blue box (on the left) contains the three fields that define the key of the key-value structure. The green box (on the right) shows two fields that receive the concept codes of the harmonised values. There are situations in which the harmonised value is empty, such as the presented example. However, the harmonised exam needs to be filled, otherwise, that entry would be discarded during the loading stage.

The cohort owners describe the harmonisation and mapping of concepts using our adapted version of Usagi. The goal of this adaptation was not to improve the metrics obtained from this tool, but instead, to reduce the complexity when dealing with multi-language cohorts, which demanded a significant effort in translating and mapping the concepts manually [118]. The outputs obtained from this procedure are essential to know the cohort variables that are important for migrating, what the standard concepts are for each one and the mapping of the measurements.

In the cohort harmoniser component, the system uses a new structure and adds new attributes. The structure with key-value measurements and the information needed for their characterisation now has more fields identifying the concept type, as well as the standard code for the variable mappings, whereas for the measurement it is possible to have numeric values, strings or concepts.

As mentioned in section 3.3.4, there is knowledge stored in raw data that is not directly represented. During harmonisation, the proposed system reads the cohort's ontology to check and calculate these new variables following the predefined rules. For instance, the same exam in two distinct cohorts can have a different abnormal range of values depending on the technology used to perform the exam, which was easily calculated by specifying in the ontology the normal range of values. This information combined with another patient condition led to a new entry in the database regarding a comorbidity that was not previously defined in the raw data.

3.4.2 Customised operations

The harmoniser component is capable of processing almost all the cohort migration. However, some scenarios are cohort-specific, requiring extra attention. In these cases, we need to develop custom methods, *e.g.* using Python, which will then be called by the harmoniser and process the data prior to the usual migration.

An example of the use of those methods is when there are variables such as “0” and “1” which should represent “no” and “yes”, respectively, but in a specific cohort the “0” can represent the absence of response and the real values for “no” and “yes” are “1” and “2”. Although this example could be solved in Usagi mapping, it can also be solved in this stage of the workflow. These methods are particularly useful to deal with errors in the cohort data. For example, when the cohort originally stores the patient height in centimetres, but some measurements were recorded in other units,

a custom method can easily solve the problem without changing the data source. In the end, this situation is reported to the data owners, so that they can fix the data inconsistency.

Another example regards variables that are split into columns or when two variables are in the same column. For both situations, the best solution is to pre-process the data with a custom method that will reorganise it without performing any mapping. In this way, the system will run as it was foreseen in normal execution.

These operations need to be implemented in Python, as modules that the harmoniser will load when executed. Figure 3.4 shows a diagram that represents the interaction with these modules in the transforming stage. The ad-hoc modules are represented in green, and these are loaded in the Cohort Harmoniser through a connector. Therefore, the person responsible for handling these special fields only needs to create a module to transform the data mapped to specific standard codes. This module is then injected during the pipeline in the harmoniser.

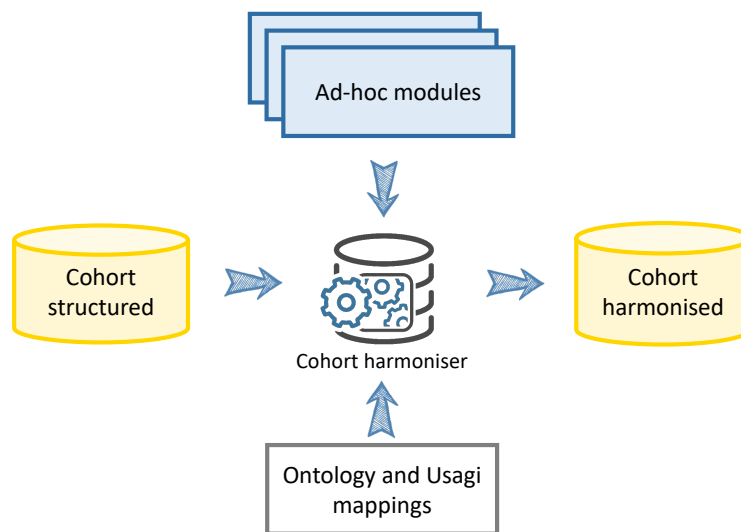


Fig. 3.4.: Fragment of the ETL workflow focused on the ad-hoc modules (represented in blue). The Cohort Harmoniser is responsible for orchestration of the transforming operations. The datasets in yellow represent the cohort raw data after being transformed to a processing structure (on the left) and the cohort in the same format with the medical concepts mapped to their standard definition (on the right).

3.4.3 Data loading into OMOP CDM

In the final stage of the ETL workflow, we can load the data into the OMOP CDM schema. The system can connect to a new database and perform this loading automa-

tically or return a set of CSV files with the data harmonised and structured. When this methodology is adapted for a new cohort, for further data updates, the pipeline does not need to be changed, and it is ready to append the new data or clean and write a new database. Then, the data can be analysed and validated.

At this stage, the system also produces a migration report, which is an execution log with all the errors and warnings that occurred during the procedure. This report helps in validating the migration and identify data inconsistencies. For instance, when there are measurements with values out side the defined range of values in the ontology or when these values are not of the same types as specified, this will appear as a warning. Additionally, this report shows incorrect dates and missing records, with the latter being detected based on the mappings done by the annotator. If a variable is mapped, this report will contain a warning for each patient with a missing measurement in that variable.

3.4.4 Limitations

The methodology was developed to generate OMOP CDM databases using cohort raw data. However, changing the output data schema to be completely different from the OMOP CDM may require a restructuring of the loading stage of the proposed pipeline. Small adjustments in this structure are possible with minor effects on the developed system. When we developed the workflow, we kept in mind possible adjustments in the OMOP CDM, because OHDSI is an active community that has improved the OMOP CDM aiming to expand to other medical domains.

The methodology was implemented and validated using Alzheimer's disease cohorts. We do not consider this methodology limited to this domain. However, applying this migration workflow using cohorts from other diseases may require some adjustments, namely in defining an ontology for this new domain. The methodology is focused on the ETL procedure, which contemplates dataset harmonisation at different levels, and it adopts well-established tools designed to perform EHR observational studies in cohort datasets. Although these cohorts are more disease-specific, the aggregation of results from different institutions has revealed impactful findings [119, 120].

3.5 A collaborative web-based ETL tool

BICenter¹² is a web-based ETL tool that covers some limitations and problems currently found in building and managing ETL tasks in multi-institution environments. This tool simplifies the description of ETL workflows and helps users without technical expertise to understand such workflows through an intuitive GUI. BICenter replicates the Kettle features in an HTML5 browser and simplifies some of the procedures in Kettle that may require deep technical knowledge of this tool.

3.5.1 Software architecture

The system follows a client-server model, with an architecture that considers four different tiers (Figure 3.5). The client-side was developed aiming to provide a responsive web application that allows building ETL pipelines and configuration of each pipeline step. The drawing features rely on mxGraph components, which communicate with a service on the application side that converts the stored information of the ETL processes into mxGraphModels. To have cross-device support, the platform uses AdminLTE, a fully responsive template based on the Bootstrap Framework that dynamically adjusts the visual components in order to fit in different screen resolutions.

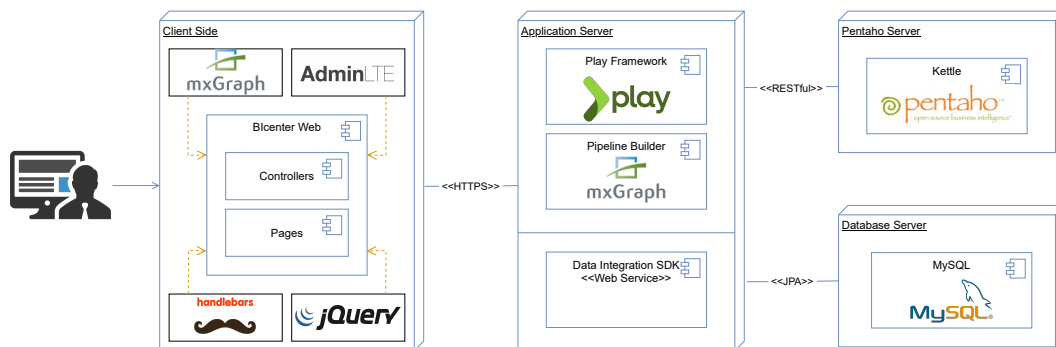


Fig. 3.5.: BICenter architecture with 4 main components: client side; application server; database server; and Kettle server.

The application server, developed using the Play Framework, controls all application functionalities, maintains the system business logic, and provides a service layer to support the client side. This component communicates with the database server using Java Persistence API (JPA). The MySQL database stores all the system information, including the ETL pipelines and their status, namely execution history, performance metrics and possible issues. To extract this information and to have

¹²We would like to thank Leonardo Coelho for the initial developments of BICenter.

the system communicate with the Kettle instances, which run autonomously, a Data Integration Software Development Kit (SDK) was developed. This SDK contains the methods required to build and execute Kettle's ETL processes.

The Data Integration SDK can represent any ETL process using six classes, similar to Kettle [121]. These classes are the following:

- **TransMeta**: A class that defines the information about the ETL process and offers methods to save and load these processes to and from XML. It also defines the methods to alter an ETL process, by adding and removing databases, steps, and hops, among other components.
- **Trans**: Represents the information and operations associated with the concept of an ETL process. This class can load, initialise, run, and monitor the execution of the ETL process.
- **DatabaseMeta**: Defines the database-specific parameters for a certain database type.
- **StepMeta**: Is the class that defines the information about a process of an ETL Step.
- **TransHopMeta**: Defines a link between two Steps in an ETL process.
- **BaseStep**: Represents the information and operations associated with the process of an ETL Step. This class contains methods for initialisation, row processing, and step clean-up.

3.5.2 Main functionalities

Blcenter contains the usual features available on the most popular solutions designed for ETL pipelines, mainly due to being constructed on top of a Kettle instance. However, this system was developed aiming to fill some of the existent gaps in these tools. In this section, we highlight some of the major characteristics of Blcenter.

ETL task editor

A common but simple use case to demonstrate the operation of an ETL solution is the processing of data retrieved from a weather station. Therefore, we used a rainfall dataset to demonstrate the basic ETL features in this tool. The ETL pipeline implemented on Bcenter is represented in Figure 3.6, and the goal of this pipeline is to count the number of rainy days by month and year.

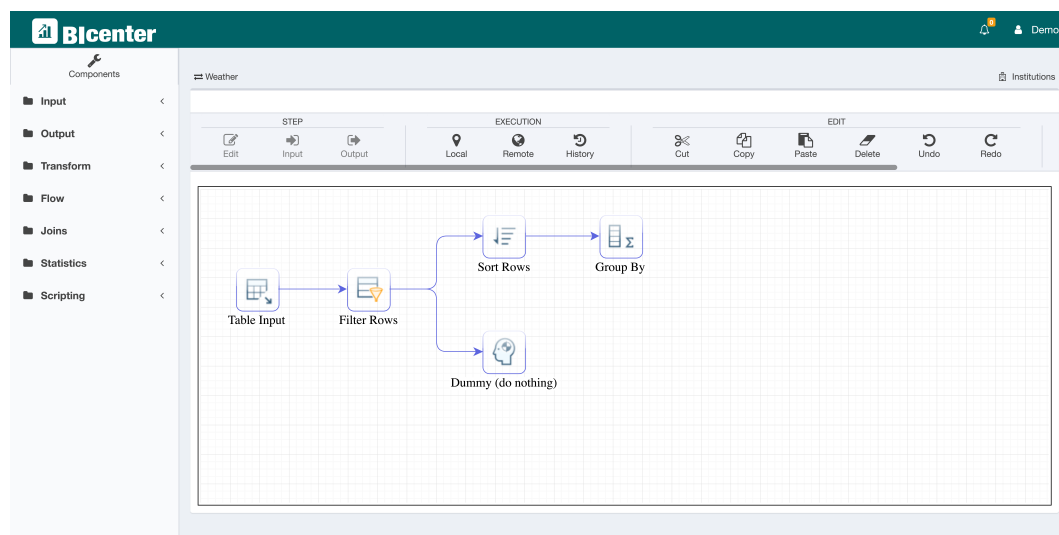


Fig. 3.6.: ETL pipeline implemented in Bcenter to process a dataset collected from a synthetic weather station.

The ETL Steps presented in this pipeline were configured in the web interface. Figure 3.7 represents the interface to define the ETL Step to filter the data by rows. This component allows the definition of conditions to be applied in the filter and returns the information that fits in the condition as true to the step designed “Sort By Date”. The data that does not fit in this condition is returned to step “Sunny Days”.

Filter Rows

Step Name:

Send 'true' data to step:

Send 'false' data to step:

Condition:

OR AND OR NOT AND NOT XOR

+ Add rule + Add group

Value > 0

Delete

Reset

Positive

Fig. 3.7.: Bcenter interface for the filter component.

After defining the ETL pipeline, it is possible to run it and also analyse the execution history. This history provides information about each execution state and the current status of each pipeline's steps. In addition, it is possible to check the performance metrics of the pipeline, detailed by each component. In Figure 3.8, these metrics are shown for the pipeline defined in Figure 3.6.

stepName	nRecords	read	write	enter	output	update	refuse	error	state	time	speed	printOut
Count Rainy Days	0	4858	525	0	0	0	0	0	Finished	28.8s	169	-
Filter Rainy Days	0	16993	16993	0	0	0	0	0	Finished	13.0s	1.308	-
Meteorology	0	0	16993	16993	0	0	0	0	Finished	12.4s	1.366	-
Sort By Date	0	4858	4858	0	0	0	0	0	Finished	22.9s	212	-
Sunny Days	0	12135	12135	0	0	0	0	0	Finished	13.0s	934	-

Fig. 3.8.: BICenter interface to display the performance metrics after running the weather station ETL pipeline.

Pipeline execution

BICenter can execute the ETL pipelines in local databases or in private and remote servers, without the need for a new local installation. The connection details for these databases or for the remote servers are associated with each institution, and the access to these connections is controlled. When connected directly to a local database, it is necessary to define a data source in the system through a form that generates the connection link between BICenter and the database. The ETL pipelines are then executed using the databases defined.

The private and remote servers have different behaviours. These servers aim to ensure data protection and isolation when dealing with sensitive data. Therefore, it was developed using Carte, which is a lightweight HTTP server available on Kettle that allows remote and parallel execution of ETL tasks. BICenter can perform authenticated requests to the servers that are running Carte. These requests contain the definition of the ETL tasks to be executed.

Carte also contains clustering features, enabling a single transformation to be divided and executed in parallel by multiple machines that are running a Carte server. BICenter contains mechanisms to simplify the process of sending commands to control the deployment, management and monitoring of transformations on the Carte slave server.

ETL tasks extensibility

Although Kettle already contains a set of ETL operations, there are always specific scenarios that may require the implementation of a new component. BICenter is deployed by default containing the most common ETL components available in Kettle. However, the system was developed with the objective of supporting the addition of new tasks without requiring the development of additional code, *i.e.*, if a new task is developed on the Kettle instance, BICenter can recognise it through its definition. These definitions are maintained in a JavaScript Object Notation (JSON) file which is automatically processed during the application start-up. For instance, in Code snippet 3.1, we show the JSON configuration to add the SortRows task on a BICenter instance.

```
{
  "name": "SortRows",
  "label": "Sort Rows",
  "componentProperties": [{
    "label": "Step Name",
    "shortName": "setStepName",
    "type": "input"
  }],
  {
    "label": "Fields",
    "type": "table",
    "componentMetadatas": [{
      "label": "Field Name",
      "method": "setFieldName",
      "type": "select",
      "source": "inputFields"
    }],
    {
      "label": "Ascending",
      "method": "setAscending"
    },
    {
      "label": "Case Sensitive Compare?",
      "method": "setCaseSensitive"
    }
  ]
}]
}
```

Code snippet 3.1: JSON configuration to specify the SortRows component.

The definition of a new ETL task using this format requires the component properties and metadata. This setup procedure should be made by an entity with solid knowledge about the Kettle task. However, after defining the new component in the system, this is available to be used by non-technical users in the web interface.

Multi-institutional access control

Users can have different roles in the application and can also belong to different institutions. Therefore, RBAC mechanisms were incorporated, in which each role maintains a set of permissions. These permissions consist of an association of an operation to a resource. The authentication entity is a facade to a given user group, that can use Lightweight Directory Access Protocol (LDAP) or Active Directory (AD) services. When a user accesses the platform, the underlying user group is determined by trying to authenticate each configured user group. If authentication succeeds, the user can be instantiated in the database. Depending on the group to which users belong, they may acquire the corresponding roles and institution access. The mechanisms to access and manage the ETL tasks and institutions can be characterized into four distinct types of users:

- Administrator: entity responsible for moderating the platform. This role contains permissions to create and delete institutions, and manage all the features associated with an institution.
- Resource manager: entity capable of managing private data sources and execution servers. This role has permission to create and delete private data sources and execution servers, within specific institutions.
- Task manager: this entity can build and execute ETL tasks, and is capable of accessing the ETL Task Editor to create and configure ETL tasks, within specific institutions.
- Data analyst: this is the most limited role, and can inspect task execution history, namely the resulting data, execution logs and performance metrics.

3.5.3 Collaborative features

Since the ETL pipelines for this use case may require the intervention of cohort owners, a tool with collaborative features may simplify their implementation. Although PDI

is considered one of the most relevant and complete tools aiming to simplify the design and creation of ETL processes [111, 107], it lacks collaborative features. However, BICenter covers some limitations and problems currently found in building and managing ETL tasks in multi-institution environments [12].

One of the main features of BICenter is the Visual ETL editor. This editor is illustrated in Figure 3.6, where an ETL pipeline with four simple steps was defined. This tool can fill some of the existent gaps in the ETL tools, namely related to collaborative environments. With BICenter, cohort owners can participate actively in implementing the ETL pipelines, which may simplify the ETL design, implementation and validation.

This application organises users by project or institution and for each cohort, there is a set of users with permission to work in the ETL tasks. These tasks can be executed using a local database, or private and remote servers. In our case, we used the local database during the development of the ETL tasks, to then apply the same pipeline using the remote servers. These servers are based on Carte, which is a lightweight HTTP server available on Kettle that allows remote and parallel execution of ETL tasks. This approach aims to ensure data protection and isolation when dealing with sensitive patient data.

3.5.4 Usagi mapper connector

Although BICenter already includes a set of ETL operations, some flows can be optimized, namely by creating a new step. A component capable of applying the transformation defined on the Usagi tool directly in the data would reduce a set of operations in the diagram to a single step. This transformation would be able to identify the source concepts in the data and change them for the standard codes. Furthermore, this component would reduce the complexity of the ETL diagrams considerably and the cohort owners would only need to update the file with the mappings in each update.

Figure 3.9 illustrates the interface of the Usagi Mapper in BICenter. This interface aims to be intuitive for the cohort owners, and the fields in this form can be easily understood by non-technical people. The input “Fields to use” identifies the column in the source data which would be applied to this transformation. The data in this column are matched with the mappings in the Usagi output, that is specified in the system using the option selected in the “Input Column” field. The new values for this

transformation are defined in the same output but in a different column. This column is defined in the “Output Column” field. These options are compliant with the Usagi file structure.

The screenshot shows the BICenter web interface. On the left is a sidebar with a 'Components' menu containing: Input, Output, Transform, Flow, Joins, Statistics, and Scripting. The main panel is titled 'Berlim cohort > observation_concept_id > Edit' and contains a 'Usagi Mapper' configuration form. The form has the following fields: 'Step Name' with the value 'observation_concept_id'; 'Fields to use' with a dropdown menu showing 'Variable'; 'USAGI File Name' with a file explorer icon and the filename 'conceptMapping.csv'; 'Input Column' with a dropdown menu showing 'sourceCode'; and 'Output Column' with a dropdown menu showing 'targetStandardConcept'. At the bottom of the form are 'Cancel' and 'Submit' buttons.

Fig. 3.9.: Configuration view for the Usagi Mapper component. The first field represents the step name in the ETL task. The second field selects the field in which the transformation would be applied. The remaining fields are for uploading the Usagi export file, and to select the input and output column.

The complexity of updating the ETL mappings in BICenter is reduced to the operation of uploading a new file. This simple task does not require programmatic knowledge, and it can be easily executed by the non-technical users collaborating in the cohort harmonisation. In the case of cohorts with non-English medical attributes, we can use an adaption of the Usagi tool prepared for multi-language [15]. This solves an important issue since it is common to have the original data in a non-English form.

3.6 Results

The proposed methodology enabled the creation of a research ecosystem using multiple cohort data of patients suffering from Alzheimer’s disease. However, during the development of this work, in collaboration with the cohort owners, medical researchers and technical teams, an ontology was created to be used as a base for migrating other Alzheimer’s disease cohorts.

3.6.1 Ontology for Alzheimer's disease cohorts

The ontology was built using the Clinical Data Interchange Standards (CDISC)¹³ as a guideline, in which we integrated the knowledge of clinical experts and from previous harmonisation efforts related to Alzheimer's disease [122, 123]. In the ontology, we added the same concepts of the standard vocabularies, reducing the vocabulary size considerably, and simplifying the mapping task. This has two main benefits: it provides an elegant structure to manage the rules to apply to the concepts during the migration process, and it decreases the number of concepts in the Usagi dictionary, which increases the tool's performance.

The ontology created follows a hierarchical structure, subdivided into 12 domains:

- Clinical information: Contains sub-domains that describe some clinical information, namely related to alcohol use, smoking, vital signs, comorbidities, clinical visits and follow-ups, and medication use.
- Cognitive screening tests: Contains the concepts for cognitive screening tests, namely cognitive estimation, memory alteration, montreal cognitive assessment and mini-mental state tests.
- Demographics: It is a small domain for characterising patients at the demographical level.
- Harmonized biomarker values: It is a node for storing meta-information about the possible values of the harmonised biomarkers.
- Imaging: Contains the standard concepts to map information of Computed Tomography (CT), Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) exams.
- Laboratory test results: Includes the concepts related to Blood and Cerebrospinal Fluid (CSF) protocols.
- Lifestyle factors: Contains the concepts to map the patient's information about nutrition, physical activity and sleep.

¹³<https://www.cdisc.org>

- Neuropsychological examination: It is a node with several layers related to neuropsychological exams, namely visuoconstruction, language, memory, intelligence and attention.
- Pharmacogenetics findings: It is a specific class, that is mostly related to the apolipoprotein E gene present in the patients.
- Rating scales: Defines the rating scales for the different institutions, which is used as a control value, when available.
- Subject characteristics: This node contains information about the patient's lifestyle and education.
- Study information: Contains the cohort raw data metadata.

Each of these domains, includes several sub-domains with more detailed information, for instance, the concept type, range-of-values, a brief description and some additional information relevant to the migration workflow. Figure 3.10 shows an ontology entry, which represents how a particular concept is defined in this ontology.

Annotations	
rdfs:label	Amyloid Beta 1-40 (pg/mL)
Abbreviation	AMyLB40
Abbreviation	AMyLB40
DefinedinCDISC	Yes
Range_or_Possible_Value	>0
rdfs:Description	A measurement of amyloid beta protein which is composed of peptides 1 to 40 in a biological specimen. (CDISC) - Standard units: pg/mL
rdfs:conceptCode	# 45768723
Classes	
CSF Measures	

Fig. 3.10.: Node on the ontology to define a standard concept. The rdfs:label identifies the node in the ontology and, for this example, the desired input values from the cohort raw data are positive numbers. The standard code is represented in the rdfs:conceptCode, which belongs to the SNOMED vocabulary with the identifier 45768723 [124].

3.6.2 Cohort harmonisation

The harmonisation workflow was validated in an initial stage using two synthetic datasets that were generated from real data. These cohorts had small numbers of

patients and a reduced number of concepts. However, we were able to test and validate the efficiency of the automatised components. This initial validation was required to ensure that the system was developed with quality and that the outputs produced were as expected. This validation was made manually with the collaboration of the elements from Alzheimer Centre Amsterdam, in the Netherlands. They received a small sample of the database generated with the methodology and identified possible structural errors, namely in the mapping of the concepts in the OMOP CDM schema.

With the full pipeline consolidated, we used two heterogeneous cohorts from the EMIF-AD project. Those cohorts are the Berlin Memory Clinic (BMC) cohort related to the Charité University Hospital in Berlin containing 6583 individuals, and a small set of 86 patients from the BioBank Alzheimer Center Limburg (BBACL) cohort in close collaboration with the Maastricht University Medical Centre. Both cohorts were mapped following the pipeline described in section 3.3. All the attributes were analysed, but we only mapped the variables of interest for Alzheimer’s disease studies.

The BMC cohort provided 85 attributes, of which 59 were mapped and 26 discarded. The BBACL cohort contains 313 variables, but only 113 were included in the minimal clinical dataset. From the mapped variables, we further generated new attributes based on the ontology rules: 8 from the Berlin dataset and 20 from the Maastricht cohort. A summary of these variables is presented in Table 3.2.

Tab. 3.2.: Summary of attributes in both cohorts. The first column contains the sum of all variables in the raw data. The columns discarded and mapped are the number of variables used from the original data, and the composed column represents the number of attributes generated from the ontology rules. The final column contains the number of attributes forming the migrated cohorts.

	Variables	Discarded	Mapped	Composed	Final
Berlin BASE-II	85	26	59	8	67
Maastricht Study	313	200	113	20	133

The variables’ selection was based according to the information considered of interest for future studies. The variables discarded represents noise in the data, which would be difficult for the analysis of the cohort if these were migrated. In the case of Maastricht, a considerable number of variables related to blood analysis were available, but they were not considered relevant by researchers. The composed variables are

the new information that was indirectly present in the cohort, but it was identified and stored in a searchable format.

Similarly to the harmonisation procedure and knowledge representation, the migration pipeline also detected incorrect values in the collected data. This analysis allowed the datasets to be cleaned, providing at the end more accurate information.

3.6.3 Blcenter-AD applied to Alzheimer's diseases datasets

In Blcenter-AD, using the PDI steps and the Usagi component, we were able to implement the methodology described in Section 3.3. In these examples, we had the cohort raw data stored in CSV files, which did not require connecting to any database. However, these datasets had a heterogeneous format. To simplify the ETL flows and reuse parts of the transformation stage, we first reorganised the cohort raw data into a similar format. This operation has divided the ETL into two tasks: one to transform the data to a pre-harmonised format; and a second one, that takes the first one output and proceed with the data

The pre-harmonised format stores the data in a key-value structure, in which both key and values are tuples. Therefore, the fields composing the key are: i) the patient identifier; ii) the visit date; and iii) the exam or cohort attribute. The value would be the entry for that attribute, and in the last stage, the mappings for the cohort attribute and its value. This format, and how the data is reorganised in it, is illustrated in Figure 3.11. The first table represents an example of cohort raw data. The second table is in the pre-harmonised format, and the coloured boxes represent the reorganisation of the columns and fields in the new format. This first transformation is the most complex and ad hoc in the pipelines since it requires the use of several PDI steps to generate this transposition. The third table addresses the mappings of the concepts. Each row in these tables represents a medical attribute collected in a follow-up visit for a single patient. This structure simplifies the harmonisation stage since the medical attributes and their values are clearly identified in the structure.

To demonstrate the ETL flow, and using the BBACL as an example to demonstrate the ETL flow, the first step of this methodology was to identify which columns would constitute the key of the pre-harmonised structure. These fields should be capable of representing the patient in a follow-up visit and correlated to the medical attributes. The next step was the reorganisation of the raw data into the pre-harmonised struc-

Patient ID	Visit date	...	MMSE Total Score	Clock Drawing Test	...
10424	15-01-2013	...	16	1	...
10424	24-02-2013	...	20	3	...
...

Patient ID	Visit date	Original exam	Value	Harmomised Exam	Harmonised Value
10424	15-01-2013	MMSE Total Score	16		
10424	24-02-2013	MMSE Total Score	20		
...		

Patient ID	Visit date	Original exam	Value	Harmomised Exam	Harmonised Value
10424	15-01-2013	MMSE Total Score	16	2000000166	16
10424	24-02-2013	MMSE Total Score	20	2000000166	20
...

Fig. 3.11.: Illustration of cohort raw data (first table) and its representation during transformation stage. The blue box represents the concepts that identify the patient's visit. The green box represents the new position of the cohort's exams. Both of these fields represent the key of the keyvalue structure, for the value of the exam (orange box). The yellow box represents the fields that would receive the harmonised concept codes.

ture, as explained before. Once the cohort is in this structure, the transformation and loading stages are similar for all cohorts. The Usagi component loads the mappings and applies the transformation for all the entries. The last part of the ETL task gathers the structure and reorganizes the data in order to fit into the data schema of the OMOP CDM database.

3.7 Discussion

This work shows the benefits in adopting the proposed methodology to migrate tabular medical data into OMOP CDM databases. Applying a graphical ETL tool to design the cohort migration pipelines also provides some advantages. Although some parts of the tasks presented would be simpler using a programming language, as shown using the CMToolkit, this may not be the best option when non-technical people need to understand what is happening with the data. In this section, we discuss the advantages of having the cohort data in this format focusing on data quality and analysis, the strategy adopted to have interoperability between the resulting databases, how data privacy is ensured and the impact of the collaborative features.

3.7.1 Data quality and analysis

One advantage of using this workflow is data quality. At the end of the ETL procedure, the system was able to provide a migration report that includes statistical information about the data migrated, including inconsistencies in the source data. This information was helpful for the cohort owners so they could rectify these issues, since the values were collected manually during the patient's follow-up visits.

Besides this data quality control and the adoption of a common model, the methodology facilitates data sharing in multiple cohort studies. The research question can be defined in one dataset, where the cohort details are specified, and the resulting query can be shared and executed in the remaining cohorts to assess whether the medical findings are replicable in different populations. This query can be manually defined in the database through SQL language, or by using ATLAS.

ATLAS is an open-source web platform that allows to conduct scientific analyses on OMOP CDM databases. It can also be considered a web user-friendly query builder for these databases. For instance, consider the following scenario: a researcher wants to study a patient dataset based on several medications and exams, patients' personal information (such as age and gender) and correlating these conditions in a temporal window of events. If the data are stored in institutional systems, the support of IT teams may be necessary to query the databases, which is time-consuming and not always feasible. Both strategies are currently used in several institutions, but there is a considerable delay associated with data collection. Additionally, neither approach allows data interoperability, which is the main requirement of our methodology. Using ATLAS, the researcher can easily define the cohort entry events, inclusion and exclusion criteria, the concepts studied and other conditions.

3.7.2 Datasets interoperability

The advantages of the proposed methodology are not limited to the simplification of the data analysis. This also allows the use of the same analytical tools in distinct cohorts. The OHDSI community includes specialised tools to show statistical information about the dataset graphically using the ACHILLES¹⁴ tool, which is an R package that performs broad database characterisation. Therefore, ATLAS and ACHILLES provide a web environment with analytic features to work with migrated

¹⁴<https://github.com/OHDSI/Achilles>

datasets individually, but by being in a homogeneous data schema, these analyses are easily replicated. Additionally, there are other tools, namely the EHDEN Network Dashboards¹⁵, which are focused on comparing OMOP CDM databases, and these features can also be incorporated to compare different cohort datasets in order to understand which are feasible as part of a multi-cohort study.

One of the key points of harmonising cohort data is the use of a common data schema as the output of this procedure. By using the OMOP CDM schema, we were able to apply a well-established data schema that is currently used to store EHR information in an interoperable format for observational studies. Alzheimer’s disease cohorts can be mapped to this structure without any adaptations in the original data schema. This ensures that the resulting databases are compliant with OHDSI principles, and cohort owners can use the OHDSI analytical tools to interact with the data.

The interoperability lies in the use of the original OMOP CDM data schema. This schema is fully detailed at the OHDSI book [28]. However, in this case, we are only required to populate three tables, namely the “Person”, “Observation” and “Observation Period”. The “Person” table can store the patient’s personal information, *i.e.* gender, date of birth, race and ethnicity. However, we did not require all of these fields in the Alzheimer’s disease cohorts. The “Observation” table maintains all the measures made during the study, which we defined previously as exams. Each entry in this table contains: i) a numerical entry for patient identification, generated during the ETL procedure and only used in this database; ii) the standard code for the observation concept, *i.e.* specific exam conducted during the patient’s visit; iii) the standard code for the observation type concept, which characterises the measure/exam done on the patient when it can be represented using a standard code; and iv) the date and value of the observation. This value can be characterised by its type, *i.e.* it can be numeric, text or a code. The “Observation Period” table contains the time interval each patient was under observation, starting from the date of the first entry in the cohort and ending with the date of the last follow-up visit.

3.7.3 Data privacy

Cohorts’ data schema is typically distinct and the integration of multiple cohorts is always an ad-hoc procedure that typically needs to be repeated for each new study. With the proposed methodology, *i.e.*, data harmonisation into a standard schema,

¹⁵<https://github.com/EHDEN/NetworkDashboards>

we can avoid this problem and speed up research. At the same time, since the data transformation is performed locally by each data team, we ensure the privacy of combined data. Therefore, our methodology can overcome some existent barriers in medical research regarding ethical, legal and social issues. The ethical and legal aspects related to patients' data privacy and the second use of this information are settled because the OMOP CDM format is compliant with General Data Protection Regulation (GDPR) guidelines. The social issue that might arise from the fact that researchers do not want to share data is also addressed because we consider a scenario in which the data do not need to be shared at all.

The level of anonymity using OMOP CDM is dependent on the organization's privacy policies. The OMOP CDM can store patients' information without exposing sensitive data. In the case of sensitive attributes that would affect this directly, these were discarded during the migration. This was a manual procedure, in which the cohort owners identified the patients' attributes that did not contribute to studying the disease, but could identify the patient. The idea of this operation was to hide these attributes and aggregate the necessary fields in generic groups of data. For instance, we used patients' age and discarded their date of birth, which did not affect the data value.

The resulting databases from the work proposed in this chapter contain harmonised patient information in a standard format. Although the data was pseudo-anonymised, the institutions kept the data isolated and inaccessible. However, the people interested in querying the databases can define their study request, send it to the cohort owners and wait for the results. The cohort owners can execute the SQL against the database and analyse whether they can reveal the results. Currently, this methodology for performing distributed studies is used by the OHDSI community at the EHR database level.

3.7.4 Multi-institutional environments

The use of BICenter leveraged this methodology to new possibilities, leading to a collaborative and multi-institutional environment. BICenter was initially developed to have different roles assigned to different institutions. This strategy allows the use of a single installation to define the migration pipelines of all cohorts with the possibility of splitting users by institutions or cohorts. Therefore, the existing RBAC mechanisms maintain sets of permissions to access the different features of the application. For

instance, it allows specific users to visualise the results of each transformation, or write them in the target database.

The mechanisms to access and manage the ETL tasks and institutions can be characterized by four distinct types of users: data analyst, task manager, resource manager and administrator. The data analyst is the most limited role in the system. Users with this role can inspect task execution history, namely the aggregations of resulting data, execution logs and performance metrics. These users cannot execute the ETL pipelines. Therefore, the medical teams that only contribute to the ETL validation have this role. The task manager is the entity capable of building and executing ETL tasks within a specific institution. Some elements of medical teams have this role when they collaborate more actively with the technical teams during the ETL implementations. The resource manager is the entity responsible for managing the private data sources and execution servers at a deeper level than the task manager. Finally, the administrator is responsible for moderating the platform.

The collaborative environment is centralized in the ETL Task Editor. This workspace allows the definition of the ETL pipelines. Therefore, users with permission to edit an ETL task can work collaboratively in the same workspace. Although BICenter does not create real-time working sessions, the system provides a user-friendly environment where multiple users can work collaboratively.

3.7.5 Impact of web ETL tools

Many software applications aim to conduct ETL workflows. Although some of these are currently being used in complex scenarios, they lack flexibility in collaborative environments. BICenter aims to fill some of these gaps by providing similar features as well-established ETL tools, *e.g.*, Kettle. The collaborative environment was a core requirement to develop this application, allowing the same workflow to be defined and shared within a team as well as allowing the pipelines to be executed in multiple institutions.

The harmonisation of health data into a common data schema is one scenario that can benefit from this approach. Currently, some initiatives aim to reuse health data to perform analysis of the data in clinical studies [73]. Some of these initiatives include the creation of a network of databases from several institutions [65]. This requires a common data schema and the source data needs to be processed in ETL workflows.

However, the data owners do not know the target data schema or how to map their database into this schema. On the other hand, the teams specialised in the target data schema usually do not know the source data or how to map the medical concepts into their standard definition. The current solutions for this problem are not flexible and do not provide a collaborative environment. The collaborative environment provided by BICenter can solve this teamwork problem in this and similar scenarios.

BICenter was developed not to only replace tools currently used for ETL workflows, but to extend the diversity of solutions using a web-based approach. To remain compatible with existing tools, it takes advantage of Kettle features to manage the ETL processes locally, instead of building a new core. Despite having a collaborative environment, BICenter also allows the execution and definition of ETL workflows remotely, without accessing the source and target servers. This strategy simplifies the management of ETL workflows and the strategy used in the implementation provides a layer of security to access private servers. These features simplify the tasks of technical teams responsible for handling data.

3.8 Final considerations

Multicentre studies empower clinical research by extending the research to different populations with similar characteristics. In the study of rare conditions or diseases with a low number of subjects to be studied, the reduced number of participants is normally the most significant drawback to attaining a solid investigation and a higher impact of results. However, using similar cohorts from distinct and independent studies has the potential to increase the research value and validate the findings.

To simplify this research scenario, in the first instance, we developed a migration pipeline that relies on a standard data schema (the OMOP CDM), on normalized vocabularies (Unified Medical Language System), and on open-source analytic tools (the OHDSI ecosystem). The result of this work helps foster collaboration between different clinical institutions studying the same disease, respecting patients' data privacy. Additionally, this pipeline simplifies data filtering and sharing, necessary to answer specific research questions without making a new clinical trial.

These efforts contributed to the possibility of conducting a multi-centre cohort study due to simplification in harmonising cohorts' raw data into a common data model. However, as we experienced, such ETL procedures require collaboration between a

technical team and cohort owners, who are usually people with a medical background. The development of these procedures requires the above-mentioned collaboration during the design, implementation and validation of the ETL, due to the data scope.

BIcenter is a web collaborative ETL tool capable of reproducing the components of PDI using a responsive HTML interface. This tool provides a workspace where both teams can work and understand what is happening with the data. The goal is to have a platform to set the ETL pipelines without using programming languages, which are not understood by the medical peers involved in the process. This simplifies some phases of the pipelines, reducing time, and ensuring a deeper validation of what is happening with the data during each stage.

In the latter stage, we created the BIcenter-AD, which uses the same core as BIcenter. However, this version is more focused on the Alzheimer's disease problem, having extra components capable of addressing the migration pipeline proposed and validated in this work.

From unstructured text to ontology-based registers

The content of the clinical notes that have been continuously collected along with patients' health history has the potential to provide relevant information about treatments and diseases and to increase the value of structured data available in EHR databases. These databases are currently being used in clinical studies which lead to important findings in medical and biomedical sciences. However, the information present in clinical notes is not being used in those studies, since the computational analysis of this unstructured data is much more complex in comparison to structured data. In this chapter, we present strategies for solving some of the existent gaps in ETL procedures regarding the harmonisation of clinical concepts extracted from clinical notes into a relational database.

Medicine has long enjoyed the benefits of technological developments and so has the quality of life of the population in general. Healthcare improvements were accompanied by the creation of new tools and data sources, which brought new knowledge and capabilities to physicians and impacted aspects such as disease prevention, diagnosis, treatment and patient follow-up [125]. These new resources brought the possibility of improving areas such as health research studies, which are composed of many time-consuming and expensive stages (*e.g.* identifying and recruiting subjects that consent to the study, and monitoring them over long periods), by lowering their cost and time through the exploration of already existing data, such as data obtained from previous studies or data stored in health-related registry systems [126].

However, challenges also arose with the need to cope with the resulting scale and diversity of medical data. EHR systems were created to provide an electronic infrastructure capable of storing administrative and medical data from diverse modalities, centralising data at the patient level [127] and providing a longitudinal view of the patient medical history. The resulting healthcare databases can be explored in health research studies to help increase the quality of the research, especially when combining data from several databases [128].

Apart from structured information (*e.g.* form fields), data in EHR systems can also be stored in unstructured form. Unstructured text is frequently used to document the patient's medical status and progress through time using a flexible format. Owing to this fact, free text makes up a significant part of the data stored in EHR systems, especially in chronic diseases in which clinical notes outweigh structured data [40], and can contain unique information that is not detectable in other data sources [129]. While some data can be structured using standards vocabularies such as SNOMED CT [78], RxNorm [130], or Drugbank [131], mapping these concepts is a task that can be complex and time-consuming. Moreover, the nature of the free text makes difficult the development of automatic information-retrieving solutions for clinical text.

Nonetheless, since clinical text poses great interest, some approaches have been developed for extracting relevant information. Even though this process consists in having clinical experts manually review clinical notes, a process which cannot scale and keep up with the growing rate of generation of medical data [40]. Much research has been made during the past years in fields such as clinical NLP to create solutions capable of annotating and extracting relevant concepts in clinical notes [41, 132]. Since multicentre medical studies usually do not explore the large amounts of data stored in clinical notes, there exists an opportunity to leverage those documents to complement structured data with additional information.

In the previous chapter, we proposed different strategies to migrate heterogeneous data into a common data schema. Following this research direction, we identified some gaps in these ETL procedures regarding non-structured medical information.

4.1 Contribution

This chapter explores methodologies to support the ETL procedures presented previously, which focused on reusing the information on patient medication present in the clinical notes. Summarily, our main contributions in this domain are the proposal of:

- A methodology to extract relevant concepts from clinical notes to enrich structured OMOP CDM databases, enabling the use of SQL queries or query builders for analysing the clinical text information. This can help medical researchers in the definition of patient cohorts sharing similar characteristics.

The implementation of this methodology is available at <https://github.com/bioinformatics-ua/DrAC>;

- A system that combines text mining with language detection techniques, aiming to optimise ETL migration pipelines using non-English concepts. This system was designed to be integrated into already existing migration workflows, without the need of adapting them. This multi-language system was integrated into the CMToolkit available at <https://bioinformatics-ua.github.io/CMToolkit/>;
- A methodology to unify and extract family's health history information from clinical notes using rule-based techniques in NLP. This methodology raised new strategies to automatically annotate large amounts of EHR, facilitating the detection of comorbidities within family relations. The implementation of this methodology is available at <https://github.com/bioinformatics-ua/PatientFM>.

This chapter is mainly based on the following publications:

- **João Rafael Almeida**, João Figueira Silva, Sérgio Matos and José Luís Oliveira, *A two-stage workflow to extract and harmonize drug mentions from clinical notes into observational databases*, Journal of Biomedical Informatics, 2021, DOI: 10.1016/j.jbi.2021.103849;
- João Figueira Silva, **João Rafael Almeida** and Sérgio Matos, *Extraction of Family History Information From Clinical Notes: Deep Learning and Heuristics Approach*, JMIR Medical Informatics, 2021, DOI: 10.2196/22898;
- **João Rafael Almeida** and Sérgio Matos, *Rule-based extraction of family history information from clinical notes*, in proceedings of the 35th Annual ACM Symposium on Applied Computing, 2020, DOI: 10.1145/3341105.3374000;
- **João Rafael Almeida** and José Luís Oliveira, *Multi-language Concept Normalisation of Clinical Cohorts*, in proceedings of the IEEE 33rd International Symposium on Computer-Based Medical Systems, 2020, DOI: 10.1109/CBMS49503.2020.00056.

4.2 Background

NLP algorithms are constantly evolving and they are applied to new problems in diverse sciences, and since several solutions may solve the described issue, we divided the background into three parts: i) methods for extracting drug mentions from clinical notes; ii) techniques for cross-language identification; and iii) strategies to identify patients' relatives and their health conditions.

4.2.1 Retrieving patient information

Clinical notes are an important resource as they let physicians document patient status with descriptions throughout time, which enables the monitoring of the patient trajectory. These notes can contain relevant information such as family history, prescribed medication and medication intake, diagnosis, and followed procedures. While this wealth of knowledge stored in clinical free-text remains underexplored, developments in NLP can help leverage this source of data by effectively extracting and structuring relevant information contained in clinical narratives [40].

Information extraction can typically be divided into two components. The first one is Named Entity Recognition (NER) and consists of the detection of entities of interest in the text. In the clinical text, these entities can involve mentions from family history, prescribed medication, disorders, laboratory measurements, and others. The second component is Named Entity Normalisation (NEN) and aims to further structure the extracted text by normalising entities according to coding standards [133]. When dealing with clinical text, this process can leverage existing medical terminologies such as RxNorm or Drugbank.

Similarly to other NLP problems, medication extraction from clinical narratives can currently follow two main paths: heuristics-based solutions and machine/deep learning-based solutions. However, deep learning-based solutions still struggle when annotating certain information such as duration, adverse drug events and reasons, similar to what human annotators experience [134, 135]. Since our objective was to extract patient information from complete clinical notes, we opted for customisable frameworks capable of generalising and extracting medical concepts from diversified clinical notes.

MedXN [136], MedExtractR [137] and cTAKES [138] are examples of open-source Unstructured Information Management Architecture (UIMA)-based solutions for information extraction, with cTAKES being a modular and extensible framework and MedXN being a solution specifically designed for medication extraction, whereas MedExtractR is an R programming language package that follows a similar approach but sacrifices some generalising capability by narrowing down the scope of drugs to search for [137]. Another flexible and modular framework for text processing and annotation is Neji [139]. This open-source system provides annotation services that can be easily configured and extended with new dictionaries and machine learning models, having already been used in previous work to extract family history information from clinical narratives [16]. Resources from Neji and MedXN were used in this work to extract drug-related information, as described in more detail in Section 4.3.1.

4.2.2 Cross-language matching

The problem of multilanguage has been studied over the last years mainly due to the amount of non-English information spread over the internet. Indexing information following the semantic web principles allows cross-language search and domain language identification. Trojahn *et al.* [140] analysed the state-of-the-art in cross-language ontology matching and described different methodologies using semantic web. These authors concluded that there is no perfect solution to solve multilingual and cross-lingual matching problems.

Bella *et al.* [141] proposed a solution based on semantic matching where labels are parsed by multilingual natural language processing. This solution works using the background knowledge of the domain languages relying on offline multilingual NLP and lexical-semantic resources. The solution may be able to solve our problem, however, we do not need to use such sophisticated techniques, at least at an early stage in which the data source dimension does not justify.

The analysis of free text and concept detection in clinical texts can follow different approaches. Typically those problems were solved using ad-hoc solutions or using general information extraction frameworks [142], which are complex to integrate and customised for specific scenarios. Some authors explored the concept search based on similarity or exact matching in order to map clinical terms in the standard definition. Nunes *et al.* [143] developed a system capable of recognising and annotating more

than 1.2 million concepts extracted from more than 1.6 million external references in 30 online resources. This system provides an external service which simplifies its integration, but it was designed to work with English concepts.

Silva *et al.* [144] used supervised and knowledge-based disambiguation methods to identify the correct meaning of biomedical terms. The authors used MEDLINE abstracts to train word embedding models, and the UMLS Metathesaurus to calculate Concept Unique Identifier (CUI) embedding vectors from UMLS textual definitions. We believe in the potential of this approach for a more comprehensive scenario using English terms, but in the proposed context, where non-English words need to be considered, it is necessary to train new models for each language.

4.2.3 Patient Relatives Extraction Approaches

The patient's family history retrieval from clinical notes can be divided into two strategies: i) the identification of the patient's relatives; and ii) the diseases of each relative. Concerning this, we split our study of the current methodologies into these two paths.

The patients' relatives' extraction could be addressed as a task to identify specific words in the clinical notes. However, this was not straightforward because of two main issues: i) the text can have information about the relatives of the patient's partner; and ii) the relation of the family member could not be directly expressed. For instance, there are clinical notes where the patient is a baby born, and the first person in the clinical notes are the parents. For these cases, all the detected members need to be considered following the translation of the expressed relation. Also, there are situations where the relationship is quite complex to understand, because there are so many kinship degrees, that the computational system eventually loses the context.

The use of rule-based models is mainly the preferred architecture to solve this type of issue. A good set of rules, in theory, will have a good concept coverage, producing excellent results. Goryachevet *al.* [145] proposed a rule-based algorithm and evaluated it using 1 000 sentences. The good results showed the success that this kind of architecture could produce, although the validation dataset is small.

Friedlinet *al.* [146] also adopted a rule-based model for extracting and coding clinical data from free-text reports. Their system identifies the family history section if present and then processes the identification of disease mentions. However, their approach was concentrated on specific diseases.

Billet *al.* [147] already explored these problems with different data sets, considering the patient's relatives and their diseases. They used the UIMA-based approaches in NLP. Our methodology follows a similar philosophy, but we decide to invest in rule-based models.

4.3 Extract and harmonize drug mentions

In this section, we propose a two-stage workflow for incorporating some of this non-structured information into the harmonised databases (denominated as DrAC). The first stage of the workflow extracts prescriptions present in patients' clinical notes, while the second stage harmonises the extracted information into their standard definition and stores the resulting information in a common database schema, namely the OMOP CDM.

4.3.1 Clinical Notes Analysis

The first part of the pipeline is responsible for the extraction of relevant medical information from free text in clinical notes, and for the storage of extracted data in a matrix structure to be used in the second part of the pipeline. Figure 4.1 illustrates an overview of this process. Here, a system reader initially receives clinical notes as input, reads their content and stores it according to a fixed structure. This reader is implemented using the factory programming pattern, thus a new dataset reader needs to be implemented whenever a new clinical note dataset is to be used. After reading the clinical notes, a Neji web service is used to annotate medication entities in each note, and the resulting annotations are stored and post-processed. Finally, the extracted information is stored in a matrix to be used in Section 4.3.2.

Although we used Neji to annotate the clinical notes, other tools/frameworks can be integrated into this pipeline or used to replace Neji. Furthermore, this pipeline is not dependent on a specific technology. The main condition is that the output provided at the end of this first part of the pipeline should match the expected input

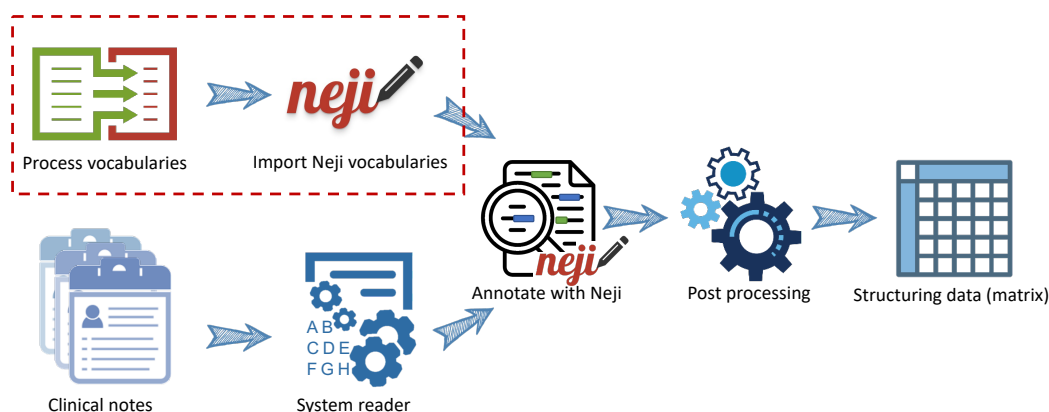


Fig. 4.1.: Overview of the extraction of information from clinical text into the matrix format. The red box represents an initial setup phase where the vocabularies are processed and imported in the Neji annotator.

format of the harmonization component. To further facilitate the use of different strategies in the annotation component, the post-processing module incorporates a programmatic parser that can be easily extended to reformat the annotation output into the expected format.

Annotating Clinical Notes

The red dashed box presented in Figure 4.1 concerns the setting up of the annotation mechanisms. Our goal was to create a pipeline capable of generalising and working with any type of clinical note, hence we needed a flexible framework for text processing and annotation that could be easily configured with new resources, such as dictionaries and machine learning models. Neji [139] fulfilled the above-mentioned requirements and provided an annotation viewer along with easy access to its annotation mechanisms through web services. Furthermore, Neji can be installed locally and used through the command line interface or a web service, so that users can use it to annotate sensitive data without depending on external services. This was a key aspect of its integration in the pipeline, considering the privacy concerns associated with the manipulation of sensitive patient information.

To set up Neji as a medication annotator, we first extracted three drug-related medical terminologies from the UMLS Metathesaurus [148]: RxNorm, DrugBank and Alcohol and Other Drug Thesaurus (AOD). However, these terminologies cover many semantic types and groups, thus to narrow down the scope of the dictionaries we filtered them keeping only entries from the “Chemicals & Drugs” semantic group. The resulting dictionaries were imported to Neji, and a Neji annotation service was configured for the extraction of drug mentions in clinical text. After passing all clinical notes

through the system reader, the Neji web service was used to annotate medication entities in each note and the resulting annotations were stored.

Post-processing disambiguation

To filter annotations and perform a further search for additional drug-related information, namely drug strength, dosage and administration route, a post-processing module was developed. This module explores specific vocabularies and integrates resources from Athena and MedXN [136], namely vocabularies and regular expressions.

The post-processing module begins by checking for ambiguous annotations. Since Neji was supplied with three different drug-related dictionaries, which may have concept overlap, it is possible to have ambiguous situations where Neji creates multiple annotations for a mention. As an example, in the sentence “the patient took aspirin 600mg orally”, Neji can annotate “aspirin” with a DrugBank code and “aspirin 600mg” with a RxNorm code. When there exist multiple annotations associated with a mention, the post-processing module gives higher priority to RxNorm annotations as they are more complex and specific, enabling the distinction of mentions that have strength information incorporated.

Since there may exist irrelevant entries in the dictionaries, which results in the annotation of many false positives, disambiguated annotations are subjected to an additional filtering process where possible false positives are removed by checking each annotation against a false positive vocabulary. This vocabulary was manually compiled and integrates part of the MedXN [136] vocabulary along with a list of common medical abbreviations used in clinical text.

Retrieving additional information

Afterwards, considering the sentence where Neji detected an entity, the post-processing module uses a vocabulary of possible administration routes to search for the route used to administer the drug within the sentence. This vocabulary was compiled from three main sources: the MedXN vocabulary, a manual list of common route abbreviations and their expansion, and finally a list of SNOMED routes retrieved from Athena, which was obtained by searching codes with type “Routes”. Route annotation is of utmost importance since the drug administration route is a mandatory field in the Drug Exposure table from OMOP CDM. Therefore, when the post-processing

module cannot detect a route for a drug annotation, this field is annotated with “N/A”.

The final post-processing step is responsible for extracting strength and dosage information. To extract drug strength, the system first checks if the annotated drug mention contains strength information, and if so it directly extracts the strength component, whereas if not the system uses an adjusted version of a MedXN regular expression to try to identify drug strength in the full sentence. Finally, the sentence is processed with a list of regular expressions to detect the presence of dosage information.

Storing extracted information

Once the information extraction process is completed, all extracted information is stored in a matrix structured by patient and drug, where each cell holds information on a drug mentioned (strength, dosage and route). The reason for storing extracted data in this particular format lies in the fact that the resulting structure is similar to that already used in cohort studies, greatly simplifying the process of migrating it into an OMOP CDM database, as described in the next section.

Practical example

Figure 4.2 presents the annotation and conversion into a structured matrix of an example clinical note. The clinical note is firstly annotated using Neji, as demonstrated in the second element of the image extracted from the Neji interface. Then the post-processing stage searches for additional drug-related information in the clinical note, such as dosage, strength and route. The resulting annotations are finally cleaned and restructured in a matrix format, as shown in the bottom element of Figure 4.2.

4.3.2 Data Harmonisation

Concept extraction from the text is only the first part of this process. Since the goal is to reuse the extracted information, the second part of the pipeline is responsible for gathering the extracted information from the matrix and storing it into the OMOP CDM data schema. Despite having the data represented in the previously defined structured format, this information still needs to be harmonised and cleaned, which is one of the main tasks of this second component in the proposed methodology.

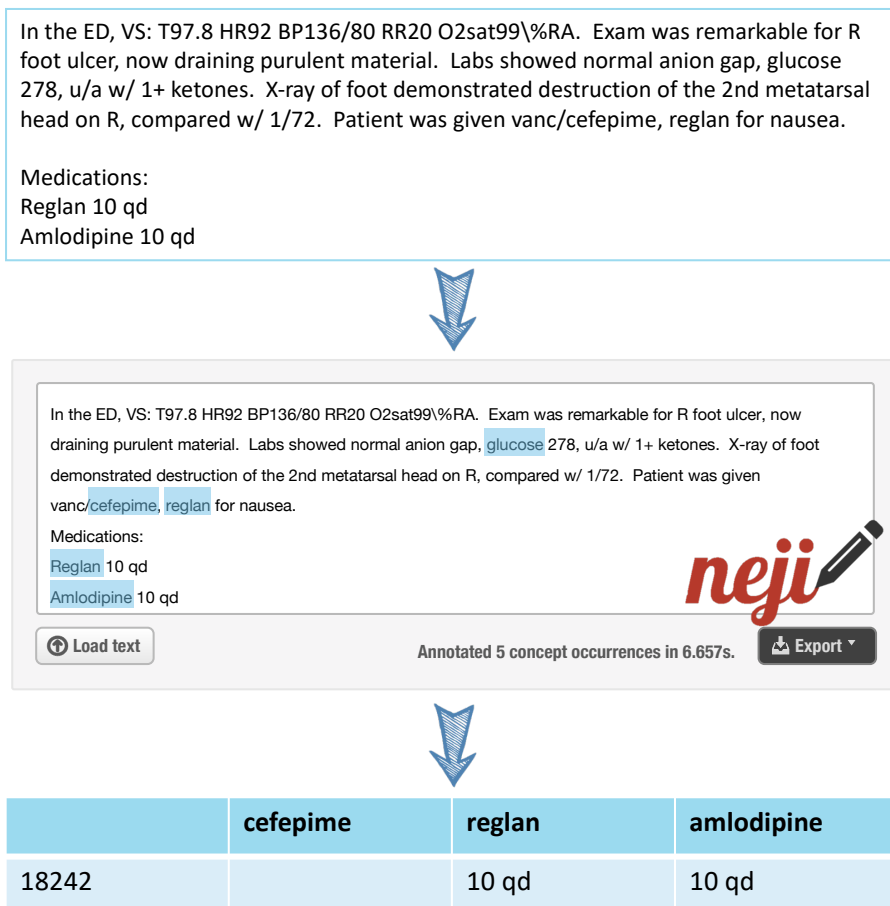


Fig. 4.2.: Example of a clinical note processed with the first part of the presented pipeline. The clinical note is firstly annotated with Neji; for illustrative purposes, the detected drug mentions are shown highlighted in Neji's graphical interface. The post-processing module searches the clinical note further for additional drug related information such as dosage, strength and route. Finally, all annotations are restructured into a matrix to be forwarded to the second part of the pipeline.

Migration Workflow

The component in the proposed methodology responsible for migrating the data to a relational database followed similar principles as represented in Figure 3.2 (Chapter 3). We improved this pipeline to a more generic solution and for this scenario, we used different output tables. Therefore, as presented in Figure 4.3, this second part is divided into two stages: i) vocabulary loading; and ii) raw data harmonisation.

Vocabulary loading stage

The vocabulary loading stage requires an initial manual procedure, where the user needs to download from the Athena platform the desired vocabularies to use in the methodology. These vocabularies are used in the OHDSI Databases Network in order

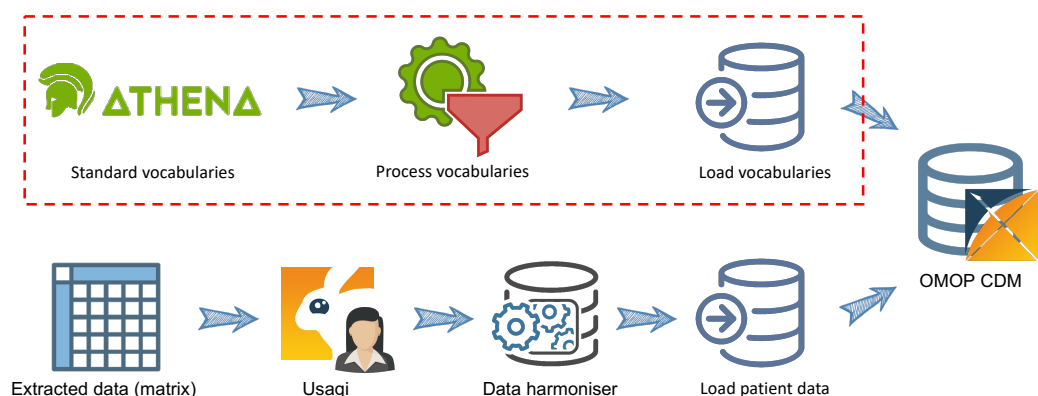


Fig. 4.3.: Overview of the data harmonisation pipeline used to read the extracted data matrix, process and harmonise its concepts into a relational database using the OMOP CDM data schema. The red box represents the vocabulary loading process that can be executed in the setup stage.

to allow federated and distributed queries over multiple databases from different countries. In the case of building a database only for clinical notes, which is not very common, this stage will load those vocabularies automatically. The vocabularies used in this methodology were RxNorm, which provides normalised names for clinical drugs, and SNOMED, which contains medical terms particularly useful for the standard definition of routes among others.

Vocabularies were also useful in the second stage to feed the Usagi tool, which maps the concepts in raw data to their standard definition. However, this second stage is complex because the information needs to be harmonised on different levels: i) the standard definition for drugs; ii) the standard definition for routes; and iii) the correct field in the data schema. This last harmonisation level is attained automatically by the system, based on the structure of the matrix resulting from the clinical notes analysis component. The remaining harmonisation levels are achieved using Usagi, as this tool already provides suggestions for each concept based on the textual similarity with standard concepts.

Concept mapping stage

The mapping stage is the most time-consuming part of this pipeline. However, it ensures that the mapped concepts are validated, while also discarding wrongly annotated concepts. Despite requiring the health professional to validate each mapping individually, empirical experience shows that Usagi's suggestions are correct for a large portion of the cases with those cases requiring very little time to validate. The tool also provides search mechanisms to simplify the correction of the remaining mappings, in order to accelerate the process.

The proposed methodology was built to be integrated into the OHDSI ETL procedures. In these procedures, concepts existing in the database are mapped to their standard definition using Usagi. The bottleneck in those procedures is the mapping stage due to a large number of concepts. However, our proposal is focused on medications extracted from clinical notes, thus it is possible to reuse some of the mappings made in a previous EHR migration (in case such migration was performed), which reduces the time required in our approach. Another aspect concerning concept mapping is that terms are aggregated, *i.e.*, even though a term can occur multiple times in the whole dataset, it is only mapped once in Usagi.

Upon completing the mapping stage, the system receives the Usagi output and creates the mappings. While this is the only file being currently used as input, in more complex scenarios an ontology containing more information about the concepts could also be used. An example of such could be the use of range-of-values in order to automatically validate drug strength for each entry of a specific drug. Despite not being explored in our use case, the proposed system was designed taking into account this possibility.

Another requirement for this part of the methodology is the need for some of the patient personal information, *i.e.*, birth date, ethnicity, race, location, provider and death date (in case of dead patients). This information is already part of the EHR structured data and can be collected together with the clinical notes processing.

Data Schema

The output of this pipeline is a relational database that adopts the OMOP CDM standards. Therefore, our methodology relies on the set of OMOP CDM tables presented in Figure 4.4, which are the following:

- Person: contains the patients' personal information (*i.e.* birthday, race, gender and ethnicity).
- Drug Exposure: captures the records related to the utilisation of a drug by the patient.
- Visit Occurrence: contains the interval times of a Person that received medical services. In our scenario, we may not be able to define the time span due to the end date of those visits.

- Note: stores the clinical note in the database. It keeps some information that characterises the note, and in one field it captures the unstructured information recorded by the provider about a patient in free text format.
- Note NLP: encodes all output from NLP processes on clinical notes. Each row represents an extracted term.

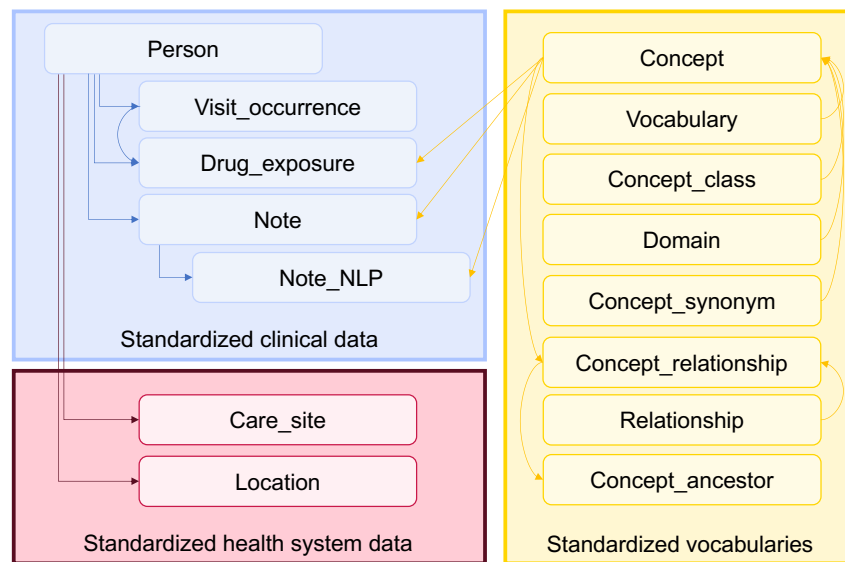


Fig. 4.4.: Tables used from OMOP CDM schema in the proposed methodology. The complete data schema was presented on Section 2.1.2.

Regarding the Drug Exposure table, it is important to highlight some of its characteristics as these may impact the text extraction procedure. The Drug Exposure table stores patient information associated with written prescriptions, orders, and pharmacy dispensing, among other situations concerning patient-drug relations. Its structure contains several mandatory fields such as “drug_exposure_id”, “person_id”, “drug_concept_id”, “drug_exposure_start_datetime”, “drug_type_concept_id”, “route_concept_id” and “drug_source_concept_id”. Some of these fields contain references to the standard vocabularies, which helps keep the record harmonised. Moreover, the table contains additional fields that can be used to characterise drug utilisation, yet these are not mandatory.

In addition to this information, the OMOP CDM schema also has a set of tables named “Standardised Vocabularies” which are designed to store the standard vocabularies as well as additional information, such as hierarchical concept relations for example. Additionally, OHDSI provides the Athena¹ web platform, which contains the most

¹<https://athena.ohdsi.org/>

common vocabularies available in the OMOP CDM schema and facilitates selecting the desired vocabularies to be used in migrated databases.

4.4 Multi-language Concept Normalisation

The idea of performing multicentre studies is focused on exploring multiple databases to answer new research questions using more substantial clinical data. However, this is only possible if the databases are interoperable, as we described in the previous chapter. One of the problems of this procedure is the effort necessary to map the original concepts into their standard definitions. While several automatic mapping solutions can help in this task, their complexity increases when dealing with multi-language databases, leading to a significant manual effort in translating and mapping. In this section, we propose a strategy that combines text mining with language detection techniques, aiming to optimise these migration pipelines. This system was designed to be integrated into already existing migration workflows, as proposed before.

4.4.1 Supportive open-source tools

Our proposal uses two open-source tools in order to: i) provide a user interface to validate the mappings; and ii) supply a web collaborative platform to manage the ontologies used in our proposal. We used the Usagi² as an interface for validating the mappings. It provides suggestive mappings based on word similarity through a simple but intuitive interface, in which the non-technical teams can validate the mappings. Usagi's suggestion only compares the concept with the standard vocabulary, leading to many wrong suggestions that end up being modified manually. Nevertheless, the user interface is intuitive and reusable for the proposed approach and is currently used in several migration workflows, including in the methodologies proposed in the previous chapter.

In this proposal, the users need to manage the ontologies containing the medical concepts over time to set up new languages in the system and detect whenever the vocabulary dictionary is extended. For the ontology, it was used the WebProtégé³, a web platform designed to simplify the development of collaborative ontologies [117].

²<https://github.com/OHDSI/usagi>

³<https://webprotege.stanford.edu/>

The use of a web platform is helpful because there is no sensitive data present in ontologies and the building process is done in cooperation with the institutions involved in the cohort data collection and the technical teams.

4.4.2 Multi-language mapper

In the harmonisation workflows, there is one step that requires human interaction, mainly by medical teams familiarised with the original database. These teams have the mission of helping during the concept mapping stage and validating the data at the end of the migration. However, the problem is the time spent in the mapping stage. The tools designed to support this stage are limited because the usual behaviour is based on a search-by-word similarity. When the database is in English, these tools only provide valid suggestions when the original terms match or are much similar to, their standard definitions. Otherwise, this procedure needs to be fully manual. In addition, all these semi-automatic approaches will fail in databases from non-English institutions due to the lack of dictionaries in the source domain language.

A possible solution could be translating the source data, however, we are dealing with very sensitive data, which sometimes invalidates the use of external resources for data translation. Therefore, the current solution is to manually map the concepts without benefiting from the possible optimisation offered by using concept recognition techniques. The proposed solution combines two types of resources in order to reduce the time spent in the mapping stage: i) multi-language detection approaches adapted for medical scenarios; and ii) text mining techniques to identify the standard concepts for each term.

System Description

The proposed system identifies terms in the dataset and tries to relate them to their standard definition, independently of the source language. However, each dataset in its original form has a different structure and the initial information about each is different. There are databases with one language or multiple languages, *i.e.*, the collected data may have been recorded in different languages. For this reason, several ontologies can be created in WebProtégé to deal with these different situations. Then, when the system runs, they are supplied as input to the system, along with the dataset raw data.

The Vocabulary Ontology contains the concepts extracted from the standard vocabularies related to the dataset scope. Those concepts are then grouped in classes, leading to a more organised and reduced vocabulary which simplifies the mapping task. This yields two main benefits. Firstly, it provides an elegant structure to manage the rules to apply over the concepts in the migration process, as well as it decreases the number of concepts in the dictionary that need to be translated. In addition, the concepts on the ontology are complemented with extra information that characterises them, for instance, the concept type, range-of-values, the translation for the desired languages, and in some cases a brief description defining the concept. Figure 4.5 shows an example of how a concept was defined in this ontology.

Annotations

















 rdfs:label	►  CDR Sum of Boxes
 Range_or_Possible_Values	►  0-18
 rdfs:Description	►  The CDR-SOB score is obtained by summing each of the 6 CDR domain box scores.
 rdfs:conceptCode	►  # 273367002
 rdfs:en	►  Clinical Dementia Rating sum of boxes
 rdfs:pt	►  Classificação de demência clínica soma das caixas
 rdfs:es	►  Clasificación de la demencia clínica suma de cajas
 rdfs:nl	►  Clinical Dementia Rating som van dozen

Fig. 4.5.: Node on the ontology to define a standard concept. In this example, there are available translations for Portuguese, Spanish and Dutch cohorts. The rdfs:label identifies the node in the ontology and for this example the range of values is 0-18. The standard code is represented in the rdfs:conceptCode, which belongs to the SNOMED vocabulary with the identifier 4164818 [124].

To complement the vocabulary ontology, we created another one focused on language domain detection. This ontology contains a set of words that occurs frequently in the dataset. For instance, words like “yes/no” or “male/female” are included in this ontology, as well as their translation for the available languages in the concept ontology. This second ontology simplifies the language identification because of the dataset composition, *i.e.*, in some situations the language was detected based on acronyms or abbreviations commonly used in the database’s country.

Operationalisation

The system operates in three different modes (Figure 4.6): i) with the dataset language defined, which skips the language detection stage; ii) having multiple languages defined, requiring the identification of the concepts only considering those

languages; and iii) without any language defined, where the system tries to identify which languages are present in the dataset. The third execution mode was designed for cases where data owners have doubts about which languages are present in the dataset.

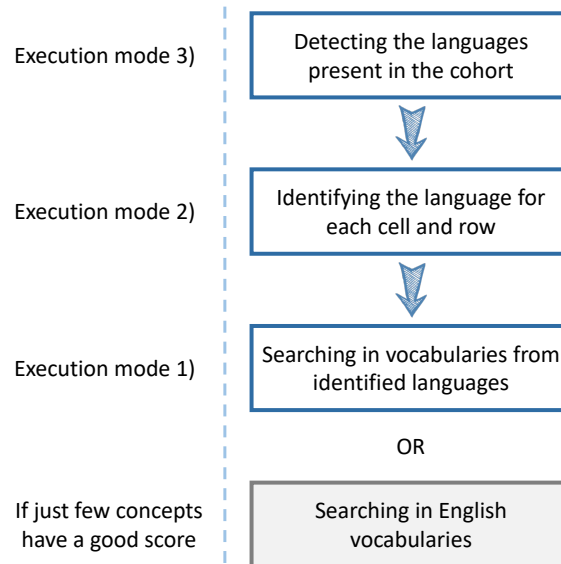


Fig. 4.6.: The system operation modes and the relations between them. The system behaves differently depending on the information available about the cohort at the beginning. In the case of no good score and the English vocabularies were not used, the system tries a final attempt using those vocabularies.

Whenever language information is not available, the dataset data is loaded and all the non-numeric values are analysed. During this analysis, there are free-text entries also identified in the dataset structure, which may require extra processing during the mapping phase. This detection is based on the number of words present in the answers for that column. Then, the system tries to infer each language by looking at the bag of words present in the language detection ontology and classifying each cell with the language classifier result. This operation creates a matrix with none, one, or several languages. The cells without any language defined are processed in all languages. However, it is assumed English by default if any language was detected.

After knowing the domain language, the system suggests the mappings by searching for the concepts in the vocabulary ontology based on their similarity. We used the Levenshtein distance to calculate the similarity between the concepts and the nodes in the vocabulary ontology, considering only the ones greater than 70 %. The concepts without mapping are defined as empty. Then, the system exports the mappings respecting the same structure used in the Usagi tool, which will then be used by the medical team to validate the mappings through the graphical interface. In addition,

one can export the validated mappings to the formats used in the harmonisation workflows, making our proposal interoperable in those scenarios.

4.5 Extraction of family history information

Despite the efforts to structure all the patient’s clinical data, clinical reports and notes containing essential information about the family’s health history, which may be highly relevant for diagnosis and prognosis. In this section, we propose two methodologies to unify this knowledge and extract family history information from clinical notes using rule-based techniques in NLP. With these methods, we intend to collect the family members informations mentioned in the text as well as associations with diseases and living status. The implementation of these methods resulted in a tool denominated PatientFM.

The family history extraction system was originally developed under the scope of the 2019 national NLP clinical challenges (n2c2)/open health NLP track on family history extraction, which had the objective of extracting family history information from EHR clinical notes [149]. This task was divided into two sub-tasks. The first sub-task aims at the identification of entities, *i.e.*, the family members mentioned in the text, and observations in the family history. The second sub-task focuses on the extraction of relations between family members, observations and their living status.

4.5.1 Dependency parsing rules

For this first approach, we pre-processed the documents with Stanford CoreNLP [150] using the dependency parsing and co-reference resolution steps. Figure 4.7 illustrates the result of applying these annotators to an example text fragment.

For the family member identification subtask, we compiled a lexicon including all considered family members and also other members such as a partner, nephew, great grandparents and half-siblings. Although these family members were not considered in the final evaluation, they were included to avoid erroneous associations with the family members considered and were filtered out when creating the final annotations. After annotating the explicit mentions, we used the co-reference graph to add the corresponding annotations to pronouns. For example, in the sentence “Her paternal

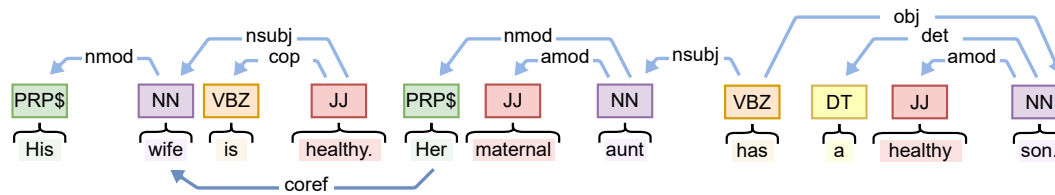


Fig. 4.7.: Illustrative example of dependency parsing and coreference resolution from Stanford CoreNLP. amod: adjectival modifier; cop: copula; coref: coreference; det: determiner; DT: determiner; JJ: adjective; nmod: nominal modifier; NN: noun; nsubj: nominal subject, obj: object; PRP\$: possessive pronoun; VBZ: verb third person singular present.

uncle has two healthy daughters”, the pronoun “Her” would receive the same family member annotation as the referred mention. Additionally, a set of rules was applied to map mentions to the corresponding family link. For example, in the example sentence above, the mention “children” would receive the annotation “Cousin”. Also, the text pattern “paternal” in the sentence is used to assign the family side to the annotation “Uncle”, which is then carried to the annotation “Cousin”.

Disease mentions were identified using Neji annotation server [139] with a disease dictionary compiled from the UMLS Metathesaurus. To improve precision, a blacklist was created by annotating the corpus used in the SemEval task on Analysis of Clinical Text [151] and identifying false positives.

For the family history subtask, we followed the shortest path in the dependency graph to associate disease mentions with family members. The same approach was used to determine living status, using a small lexicon extracted from the training data (e.g. “doing well”, “passed away”). Also, each living status mention was assigned a score (0, 2 or 4) following the task guidelines and by examining similar annotations in the training data. This approach was used in two official submissions for the task, with the second one including disease annotations from a dictionary compiled from mentions found in the training data.

4.5.2 Phrase characteristics extraction

The second approach involved the creation of rules for family member recognition and dictionaries for observation extraction and processed both subtasks as an end-to-end system outputting the required submission files for both subtasks. The engine processed each sentence in a document sequentially, aiming to link sentences when one of the system processing flows did not detect family members in a sentence.

Therefore, using this approach, we created a system that tried to answer the following three questions:

1. Who is the subject of the sentence?
2. Which observations are in the sentence?
3. Is the subject alive?

Although answering these questions does not entirely solve the proposed problem, the process of finding answers for them simplifies the procedure of establishing relations between annotated concepts.

Relatives' detection

The first step of this flow splits the document into sentences and removes a considerable set of pre-identified words. The chosen words are the most common English verbs and conjugations, several adjectives, and names. This procedure preserved relevant words and reduced the distance between these words allowing the correct identification of family members and their respective family sides.

After cleaning the sentences, the system applied exact matching rules to identify subjects in the most trivial cases. When no subject was detected, more complex rules were applied. In this case, rules have different properties, namely, collections of words that should exist before and/or after the detected subject; and if the identified subject is relevant or not for this scope. These properties generated a set of very precise rules, that when applied, increased the potential of the system for the challenge specifications at the cost of reducing its reuse in other scenarios.

When none of the previous rules was able to identify a possible family member, the system executed another component that tries to correlate the current sentence with the previous one. In case of being the first sentence in the document and no subject is identified, the system is configured by default to consider the patient as the subject in the sentence. The final component, which is always executed, tries to relate the subjects identified in the sentence to the patient or the patient's partners. If the sentence was associated with the partner, the extracted family member is discarded.

Extracting observations

The process of extracting observations is simpler than family member detection. However, this process followed similar principles and used the initial preprocessing stage for cleaning undesired words. For the n2c2 challenge, we created a vocabulary based on the observations annotated in the training set and used it in the test set. Simultaneously, the system applied rules to map the detected observation to the identified subject in the sentence. When it was not possible to establish the relationship between the observation and the detected family members, the observation was kept to be used in the first subtask of the challenge.

Identifying living status

Living status identification was performed using only two sets of rules. One set for targeting deceased subjects and the other targeting healthy and alive subjects. We did not try to identify cases where subjects were alive but not healthy because based on statistical analysis, mentions for this group of entries represented only 12.2 % (46/376) of the living status entries in the gold standard of the training set.

4.5.3 Rule-based engine

The rule-based engine pipeline processes documents individually and sentence by sentence following a sequential flow. In this pipeline, the detected words have different levels of importance. For instance, terms like partner and patient coexisting in the same sentences are weighted differently. These weights were considered by the complementary rules during subject identification in a sentence. Disambiguation was performed using a set of verbs and specific words in situations where it was not clear whether the sentence was related to the patient, the patient's relatives, the patient's partner, or the partner's relatives. Figure 4.8 shows an excerpt of a clinical note that illustrates clearly how the system processes original sentences and what is the result of this processing.

The rule-based engine provided good results in the annotation of the family members of the patient. However, the methodology used to extract observations was not the best, regardless of possible improvements to produce more accurate results. Therefore, in a second version, we rebuilt the component responsible for extracting the family members. Following the initial principles we removed specific sets of rules that were generated from the training set of the challenge, reducing possible overfitting. The system pipeline and how components are interconnected is presented in Figure 4.9.

Family history information **was obtained** from the **patient** and her **partner** this morning. Details from the family histories are on file in the Department of Medical Genetics. Pertinent information is as follows: **Ms. Benjamin** has one **sister**, age 32, and two **brothers**, ages 34 and 17, who are all reportedly healthy. One of her **brothers** **has** a **son** diagnosed with **Dubowitz syndrome**. Her **parents**, ages 53 and 50, are **alive** and well. ...

Fig. 4.8.: In the text, there are highlighted the words which the system considers relevant to make a decision. In purple, it is represented the auxiliary words which will help to understand who will be the subject of the sentence. In yellow is highlighted the patient, which indicate that the words defined as relatives in the sentence are related to the patient. If there is some information related to the patient's partner, this is highlighted in blue and indicates that there is a probability of the relative in the sentence be associated with the partner. In green is represented the possible family members and in red the diseases. Finally, in grey is highlighted keywords to indicate if the member is alive or dead.

This flow starts by trying to identify if the subject in the sentence is the patient. If the patient is not identified, the previously described complex rules are executed. The third component performs exact matching over a clean sentence for trivial annotations, and the output of these components is filtered to disambiguate relations between family members and to remove any relations that should be discarded.

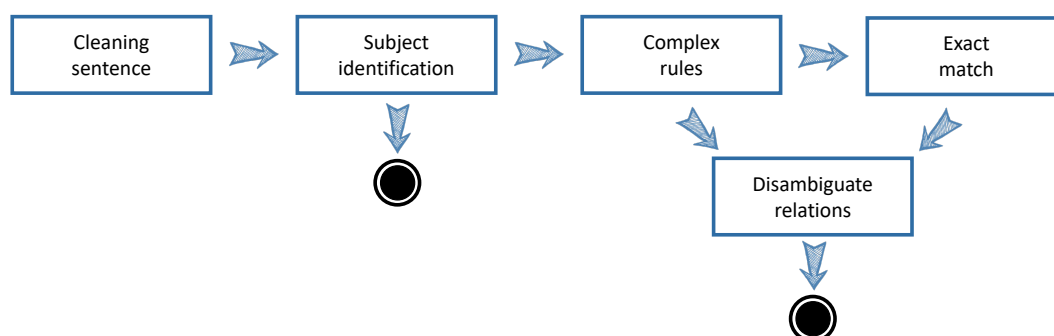


Fig. 4.9.: Overview of the processing workflow responsible for family members detection.

In the complex rules component, rules follow a six-part structure where it was defined keywords that triggers the rule, for instance, father or grandparent. In these rules, it is also defined that a list of terms needs to appear before or after each keyword. This structure also contains a flag that indicates whether the annotated relative must be considered or discarded and the terminology for the detected relative. Regarding the disambiguation component, the system contains a set of rules composed of four elements. These rules have two relatives and a mapping to the real relation of this subject to the patient. Besides the rule-based system contains a more extensive list of rules that were used for the processes of partial and exact match search.

4.6 Results

The methodologies proposed in the previous sections were validated with distinct datasets. Therefore, in this section, we present the results of each methodology, individually.

4.6.1 Drug mentions extraction and harmonization

DrAC was validated on a medication extraction use case using two public datasets from previous text-mining research challenges. This system was not implemented focusing on a particular dataset, *i.e.*, the methodology was tested on these two datasets without any prior training on them.

Use case overview

The present work focused on extracting information regarding medication from clinical narratives. This is a area of great interest since several international organisations have promoted research challenges. For instance, the 2009 i2b2 medication extraction challenge had the objective of extracting medications, dosages, modes of administration, frequency of administration and reason for administration [152], while the n2c2 2018 track 2 on Adverse Drug Events (ADE) and medication extraction in EHR systems, also added to that information the relations between drugs and ADE [135].

To validate our proposal, we used the datasets provided in these two challenges. The objective of this work was not to develop a top-performing NLP approach for information extraction, but to abstract and have a generalisable annotating system for extracting information from clinical notes, which enabled the validation of the pipeline as a whole. Therefore, we used the full datasets to validate the system.

The 2009 i2b2 dataset contains 1 249 discharge summaries from which only 252 have gold standard annotations. Even though this dataset has 9 003 drug annotations each with additional information (*e.g.* dosage, route), the challenge enabled the annotation of additional information with “N/A” whenever that information was not present in the text. Table 4.1 provides statistics on the number of annotated entities in this dataset.

Tab. 4.1.: Dataset statistics for the 2009 i2b2 medication extraction challenge full dataset, which is provided with the train and test partitions combined.

Concept	Valid annotations
Drug	9 003
Route	3 406
Dosage	4 482

Tab. 4.2.: Dataset statistics detailing the number of annotated concepts and relations in the 2018 n2c2 ADE and medication extraction challenge dataset.

	Concept			Relations to Drug		
	Training	Test	Total	Training	Test	Total
Drug	16 225	10 575	26 800			
Strength	6 691	4 230	10 921	6 702	4 244	10 946
Route	5 476	3 513	8 989	5 538	3 546	9 084
Dosage	4 221	2 681	6 902	4 225	2 695	6 920

The 2018 N2C2 dataset contains 505 discharge summaries from the Medical Information Mart for Intensive Care-III (MIMIC-III) database, annotated regarding entities and relations, and was originally split into train and test partitions containing 303 and 202 annotated documents, respectively. Even though the dataset contains annotations for a wider variety of drug-related information, in this work we only focused on drugs, dosage, strength and route, and on the relations between drugs and the remaining entity types. Table 4.2 provides statistics on the number of annotated concepts and relations present in the dataset.

Tab. 4.3.: Evaluation results from the medication extraction component applied in the validation datasets. PP: Post Processed.

Source	2018 n2c2			2009 i2b2		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Neji	0.426	0.797	0.555	0.544	0.776	0.640
PP	0.802	0.705	0.751	0.667	0.556	0.607

Medication extraction

The medication annotation component was designed to extract as many drug entries as possible to populate the OMOP CDM data schema. Since the route field in the Drug Exposure table is mandatory, a “N/A” route is attributed whenever it is not possible to detect the route used to administer a drug. While the 2009 dataset provided

the possibility of annotating various entities with “N/A”, the 2018 dataset did not. Therefore, to perform a consistent validation throughout both datasets, we decided to remove all “N/A” annotated entities during the validation process.

Moreover, since this component was developed considering the extraction of data from different datasets, it was necessary to develop a single evaluator for assessing system performance across various datasets. The resulting evaluator assesses extraction performance based solely on extracted drug and route mentions, as these two fields are mandatory for the OMOP CDM data schema (dosage and strength are informative yet not mandatory). Therefore, its analysis only considers true positives for those drugs which have an associated route annotated in the gold standard, which considerably reduces the number of true positives. For instance, despite the existence of 26 800 annotated drugs in the 2018 dataset, there are only 8 989 annotated routes and 9 084 annotated route-drug relations, which translates to a final number of valid drug annotations close to a third of the total annotated drugs. Similarly, the 2009 dataset contains 9003 drug annotations from which only 3 406 drug annotations possess valid route annotations.

The results obtained from the evaluation of the extraction component in all validation datasets are presented in Table 4.3. It is possible to observe in all datasets with Neji annotations have a higher recall, while post-processed annotations suffer a decrease in recall but an increase in precision. This behaviour is expected since Neji is used to detect drug mentions indiscriminately (high recall), whereas the post-processing module is responsible for connecting drugs with their respective routes and filtering out drugs without mention of the administration route.

Furthermore, Table 4.3 shows that the system obtained a higher F1-score in the 2018 dataset than in the 2009 dataset. However, a manual revision of some gold standard annotations of the 2009 dataset revealed some annotation inconsistencies, which can impact the performance of this component. This could be partly explained by the fact that gold standard annotations were not created by a group of medical experts but instead by challenge participants after the challenge terminated.

The obtained results represent the baseline of our annotator. These results can be improved, by training models for each data set, which was not our main goal in this work. These results were already quite promising and showed consistency since in both datasets the F1-score was very similar (a difference of 0.03 points was observed).

Importantly, these data sets are very different, thus allowing us to verify that the proposed annotator is flexible enough to serve as the baseline for this methodology.

Migrated data

The second part of the methodology was validated differently. In this case, there is no available gold standard to calculate the usual performance metrics. However, the idea of harmonising patient data from raw data stored in the matrix format into OMOP CDM databases has already been explored in other scenarios. In the OHDSI approach, the validation of this migration process is usually performed through manual searches in the resulting database.

The cohort harmonization pipeline used in the EMIF-AD project had the goal of converting the data into a common schema which was not compliant with the OMOP CDM. The ETL core of this methodology is similar to the ETL methodology proposed in Section 3.3. Overall, both pipelines share identical processes at a high-level context, *i.e.*, the use of a similar structure for the data source (matrix) and the manual validation by health professionals.

Following the proposed pipeline, the information extracted from the datasets was used in the second part of the system to validate this methodology. Table 4.4 presents some metrics regarding the data harmonisation component. When using Usagi, we defined filters to ensure that only the Drug and Route domains from the RxNorm and SNOMED CT vocabularies were used. This procedure is important as it disables similarity searches with other concepts which could have a high level of textual similarity but are not related to the medication domain.

Tab. 4.4.: Results of the mapped concepts in the second part of the methodology, including the database entries in the Drug Exposure table and the predicted amount of entries that were not mapped.

	2018 n2c2	2009 i2b2
Unique concepts	961	1049
Mapped (score equals to 1.0)	470 (48.9 %)	448 (42.7 %)
Mapped (score less than 1)	87 (9.1 %)	94 (9.0 %)
Mapped (manually)	221 (23 %)	215 (20.5 %)
Not mapped	183 (19 %)	292 (27.8 %)
Database entries	5316	8998
Discarded entries	1246	3855

The mappings were divided into categories since some required more effort to validate than others. These categories are: i) mappings with a score of 1, where the health professional only needs to confirm the automatic mapping; ii) direct mappings that had a similarity score lower than 1; and iii) the concepts that required a manual search in the tool's dictionaries. With the help of two experts, we were able to obtain the mapping values presented in Table 4.4. As observed in this table, the health professionals were not able to provide a mapping for around 20 % of the concepts. Since both professionals were not familiar with the datasets used in this work, in case of ambiguous mappings or uncertainty about the concept meaning, they decided to not perform the mapping, hence resulting in a subset of unmapped entries.

4.6.2 Multi-language cohort harmonisation

The multi-language concept normaliser was applied to harmonise two distinct cohorts (focused on studying Alzheimer's disease), which were already presented in Section 3.6.

Use case overview

One was the Berlin BASE-II [153] containing 6 583 individuals. The cohort structure is composed of 85 attributes, from which 59 were mapped and 26 were discarded. In this cohort, the data was collected and maintained in English. The second cohort is a smaller dataset from the Maastricht Study [154], containing 86 individuals. It consists of an extensive study focused on type 2 diabetes, and even though its main target is distinct, it contains valuable information regarding Alzheimer's disease. The Maastricht Study was collected in Dutch and contains 313 variables, of which 113 are of interest for this scope. During the harmonisation procedure, the mappings were done manually by the medical experts and a summary of the cohorts' attributes is presented in Table 3.2.

At the end of this process, our gold standard is constituted by the two sets of mapped attributes, which were manually validated and identified as relevant for the cohort scope. Discarding attributes is usual during the harmonisation procedures because not all the variables are considered feasible across studies. For any new cohort, if the system provides an empty mapping, it can represent one of three things: i) the ontology needs to be extended to have this new concept; ii) the concept does not fit in the cohort scope; or iii) the original concept needs to be verified manually because

it does not fit in any standard definition available in the vocabulary ontology. In both cases, the user is notified and this behaviour is expected.

Mapped concepts

Table 4.5 shows the results of our mapping methodology in both cohorts, where the precision is much higher than the recall. These results were influenced mainly by two reasons: i) the abbreviation to represent medical concepts in the cohort; and ii) the existence of similar concepts in the same node, which originated in a classification failure. For instance, the concept “Amyloid Beta 1-42 Abn” has a Levenshtein Distance of 4 when calculated against the concept “Amyloid Beta 1-42” and a distance of 5 for the concept “Amyloid Beta 1-42 Abnormal”. In this case, the right mapping has a bigger distance, because both nodes have similar references since they are both under the same class. However, the system failed in this case due to the use of abbreviations in the concept name. Cases like this occur in almost 5-10 % of the concepts header in the cohort. This can be easily fixed using the Usagi interface during the validation.

Tab. 4.5.: Results for both cohorts using the proposed system. The scores were calculated considering the variables of interest in both cohorts

	Precision	Recall	F1-Score
Berlin BASE-II	0.895	0.378	0.531
The Maastricht Study	0.809	0.371	0.508

The F1-Score of 0.531 and 0.508 for the Berlin BASE-II and The Maastricht Study cohorts, respectively, show that this methodology can provide for the non-English cohorts similar results as the ones in the English cohorts. In the methodology, we decided to prioritise precision which influenced the recall negatively. However, we were able to automatically provide suggestions with precision higher than 0.80 %, which optimises the harmonisation procedures considerably.

4.6.3 Patient family extraction

The methodology for extracting family history information was developed and validated using the corpus provided during the n2c2/OHNLP track on Family History Extraction [149].

Dataset overview

This corpus was composed of 216 clinical notes with 1 250 family members annotated and 1 836 observations (Table 4.6). In the dataset, specifically in the training set, there was a clinical note that seems to reference two different patients, *i.e.*, looks like there were two different clinical notes, wrongly merged into one. Additionally, the family members and their observations were not annotated line by line, but instead, the annotation was made by document, which difficult the training stage of the proposed methodologies.

Regarding the dataset content, the family members considered in the tasks were the parents, siblings, grandparents, uncles and aunts, the remaining relatives should be discarded for the challenge. Some of the clinical notes also referenced information about the patients' partners and their relatives, which were not relevant to the result. Consequently, those entities have also been discarded.

Tab. 4.6.: Detailed dataset statistics of n2c2/OHNL track on Family History Extraction.

	Training	Test	Total
Number of clinical notes	99	117	216
Number of annotated family members	667	583	1 250
Number of annotated observations	930	906	1 836

Subjects and observations extraction

Table 4.7 shows the results for the first subtask, obtained using the test set. The dependency parsing rules method had lower precision compared with the phrase characteristics extraction method because this second approach was less flexible in the classification of family members. The rules were restricted and selected only the persons with a high rate of confidence. A similar methodology was applied to disease extraction. This reduced the recall in the second approach because of the adversities in disease detection, *i.e.*, the enormous datasets of diseases existent were not well prepared in this approach.

When analysing the results obtained, we noticed several false positives in the third-degree relatives and unisex nouns. Therefore, we decide to measure the impact of the algorithms if the family side was ignored. The F-score for the family members increased considerably, from 0.7903 to 0.8195 and 0.8246 to 0.8614 for the dependency parsing rules and phrase characteristics extraction approaches, respectively. However, the impact of this change in the overall results was only approximately 0.02 in the F-score.

Tab. 4.7.: Results of both approaches for the subtask 1 of the n2c2 challenge. A1: Approach using dependency parsing rules. A2: Approach using phrase characteristics extraction.

		Precision	Recall	F-Score
	Overall	0.6501	0.8892	0.7510
A1	Family Members	0.7095	0.8918	0.7903
	Observations	0.6162	0.8874	0.7273
	Overall	0.8507	0.6211	0.7180
A2	Family Members	0.8514	0.7994	0.8246
	Observations	0.8500	0.5046	0.6333

Relationship between subjects and observations

The results obtained in the second sub-task are expressed in Table 4.8. The accuracy of these results was influenced by the efficiency of the methods in the previous sub-task, mainly because the data set used was the same. However, we consider the methodologies applied to assign the diseases to the patient's relatives helpful, as well as the discovery of the relatives living status. Overall, the proposed methodologies lose their efficiency because of a lack of success in the detection of the relatives. This observation was not possible to recognise only by analysing these scores, but with a filtered analysis, we could identify some details worthy of improvements.

Tab. 4.8.: Results of both approaches for the subtask 2 of the n2c2 challenge.

	Precision	Recall	F-Score
Dependency parsing rules	0.5406	0.5005	0.5198
Phrase characteristics extraction	0.6468	0.5992	0.6221

Additionally, we decided to analyse the obtained results more deeply in order to understand what we can improve to achieve exceptional results. In the training stage, the annotation was made to the clinical note, instead of the sentence. This lack of information leads to wrong rules originated automatically by analysing the dataset. If the annotation were made to the sentence, the training stage was more precise, which could create fewer rules but more accurate, increasing the precision.

4.7 Discussion

The proposed methodologies create new opportunities for reusing the information present in clinical notes, namely to complement research studies. Currently, during

the migration of EHR databases into OMOP CDM databases, the clinical notes are migrated but the information stored in them is rarely used. Although this schema has two tables for NLP extracted concepts, researchers usually do not consider this information during the study design. One of the reasons for this is the fact that the information stored in these tables is not focused on a specific domain, since these tables can store any kind of medical data that is extracted from clinical notes. This property makes it difficult to replicate studies in distributed databases, which is one of the core OHDSI fundamentals.

DrAC was designed focused on drug extraction and this system is capable of harmonising extracted information following the OHDSI principles. With this information mapped to its standard definition following the validation practices using Usagi, there is no reason not to use this data in research studies. If this system is applied to support the migration from an already existing table, some of the Usagi mappings can be reused from the EHR migration, since the source data is the same. While reusing previous EHR mappings does not guarantee that all concepts are directly mapped, mainly because of abbreviations and free-text annotations present in clinical notes, this process can considerably reduce the number of concepts in the mapping stage.

The system applied to multilanguage datasets can play an important role at this stage since its procedures output is compliant with the Usagi structure. Originally, this component was created to integrate the ETL pipeline proposed in Section 3.3. However, since DrAC shares the same principles, this component is also compliant with this tool.

Besides this novel concept regarding the use of clinical notes to increase the content of OMOP CDM databases, it is possible, with the text information stored in an interoperable data schema and mapped to their standard definitions, to simply analyse the dataset using SQL queries or BI tools. Aiming to increase the content of these databases, in a different branch, we invested some efforts trying to retrieve information from patients' family members. Although we had success and created a tool capable of accomplishing this task, the extracted information did not fit into the OMOP CDM scope. The literature shows that almost all medical studies were focused on patient data, which led us to not invest more efforts in trying to integrate the information about the patient's family members into the structured format.

4.7.1 Systems synopsis

The system was carefully designed to divide its responsibilities into two components, clinical notes annotation and data harmonisation. The main reason for this strong separation of responsibilities was to provide the possibility of performing future improvements in each component individually whilst maintaining a fully functional pipeline. Since NLP techniques are progressing at a rapid pace, this flexibility enables the information extraction component to be constantly updated, thus helping maintain the complete system up to date. For instance, this enables the future integration of deep learning-based approaches, which have already been shown to be successful in clinical text extraction tasks.

In this proposal, we used an English clinical text annotator that was not trained on any dataset since our goal was not to develop a state-of-the-art annotating system. Instead, our objective was to have a generic annotator capable of producing satisfactory and consistent results in various datasets. This way, we could solve a current problem and create new opportunities in the exploration of information available in clinical notes. However, to be useful in many realistic scenarios, it is necessary to be able to switch the English text annotator to another designed for a different language. The OHDSI community is currently spread over the world with many OMOP CDM databases existing in several non-English speaking countries.

Based on the error analysis performed in Section 4.7.2, we were able to identify a few points in which the system can be optimised. Although we used Neji for the information extraction task and this system does not currently support deep learning models, our methodology was designed to incorporate multiple annotators as well.

The decision to use a matrix for storing extracted information was made based on previous experience. In the past, we needed to migrate patient clinical data collected in medical studies that were stored in spreadsheets. After studying different alternatives for performing a clean and solid data harmonisation, the principles that we applied in the second component of our methodology were the most aligned with ETL procedures.

One key aspect of this methodology is the manual validation using a graphical interface, which cleans wrongly annotated concepts and facilitates the correction of concepts that were incorrectly mapped to other standard definitions. As previously mentioned, this is the current procedure used when migrating EHR databases into

the OMOP CDM schema. We tried to optimise this procedure as much as possible in our methodology since it requires manual interaction with the system. The idea was to incorporate the possibility of loading other mappings in the system, such as previous mappings made in the institution during an EHR migration.

An additional possibility would be to develop a pre-mapping in the annotator component, which would then be loaded in Usagi. With this approach, Usagi's suggestions would be skipped and only the annotation features would be used. However, this tool has already been validated in several migration procedures and its operation considers mappings based on a hierarchy defined in the standard vocabularies, an aspect that is not so deeply explored in text annotators including Neji. An illustrative example of this was the existence of the term “marijuana” in the 2018 n2c2 dataset. Usagi was able to map with a suggestive mapping score of 0.815 to “Cannabis sativa seed oil” because the vocabulary contains synonyms for this standard concept that is more similar to the mention than the proposed mapping. This type of feature could be developed in the Neji annotator, however, this would result in losing future community contributions in the Usagi tool.

The proposed evolution of the Usagi becomes essential when dealing with multi-language data sources. Although this component was developed using NLP techniques, it is a great asset for the ETL methodologies proposed in Chapter 3. The main motivation for this component was the effort necessary to map the original concepts into their standard definitions. While several automatic mapping solutions can help in this task, their complexity increases when dealing with multi-language cohorts, leading to a significant manual effort in translating and mapping.

One of the features of EHR systems is to store the patient clinical data. Some of this information refers to the family's health history and may be highly relevant for diagnosis and prognosis. PatientFM was developed to unify this knowledge and extract family history information from clinical notes using rule-based techniques in NLP. With these methods, we intended to collect the family members mentioned in the text as well as associations with diseases and living status. As result, we were able to properly filter and store the extracted data that was migrated to the relational databases.

Overall, the proposed systems enhance the information present in observational databases that use the OMOP CDM data schema. The work of Liu *et al.* [155] is very useful to retrieve clinical notes from the repository based on conditions defined

in a cohort. Park *et al.* [156] used the OMOP CDM database to extract the notes from a standard schema in free-text to be then annotated. Although both works were focused on using NLP to leverage the information of OMOP CDM databases, neither of these integrated the resulting data with the data already existing and extracted from the relational model of the EHR system.

4.7.2 Main limitations

Despite the efforts in developing the most accurate systems, we recognise that some components, still have some limitations, and the system is not free of errors. Therefore, we did a deep analysis of the errors that occurred during the extraction and migration process, as well as, in the extraction of the patients' relatives' health status.

Error analysis of DrAC

DrAC was built to be specialised in extracting medication information from clinical notes, without being designed for a specific dataset. Despite this goal, analysing the results of the validation datasets allows the identification of possible limitations in the annotation component. Table 4.9 presents some examples of the most frequent errors.

Tab. 4.9.: Analysis of some of the false positives and false negatives annotated by the proposed system. Mentions annotated by the system are highlighted in bold.

Missing information	Sentence from the clinical note
aspirin / by mouth	The patient is taking aspirin and enalapril by mouth .
ins (Insulin) / p.o (by mouth) / 3 cap (capsules)	5. ins: 3 Cap p.o
10 milligrams	3. lasix : 10 miligrams po
iv	The patient took an iv dosage after the breakfast containing aspart insulin .
PO	2. omeprazole 20 mg Capsule, Delayed Release (E.C.) Sig: Two (2) Capsule, Delayed Release (E.C.) PO DAILY (Daily).

The first example is due to the presence of coordination, which is not currently being processed. In some situations, the text contains more than one medication and only references the way that these are administered at the end because all mentioned drugs share the same administration route. In this example, the system only detects the route for the last drug, whereas the initial drug (aspirin) is annotated without any route, which leads to false positives and false negatives in the evaluation metrics.

The problem represented in the second example is related to unknown abbreviations, which are employed by institutional staff and most commonly present in enumerated points. Since Neji performs exact matching, these are only extracted by the annotation component if they are present in the vocabularies used.

Another issue concerning abbreviations or misspelt words is during the extraction of the drug strength or dosage. In this situation, we followed two different approaches. The first involved mention of disambiguation prioritising RxNorm annotations, when that integrated detailed information such as drug strength. Since this technique only works in specific situations, the other approach consisted in using a conditional regular expression. However, the vocabulary used in this regular expression does not consider spelling errors, such as the one presented in the third example of Table 4.9. This example shows the impact of the missing “l” that affects the detection of the lasix’s strength.

The last two examples in Table 4.9 are related to the window size used before and after concepts annotated by Neji. These words are used to identify more information about the medication (such as the route, dosage, and strength, among others). The first example shows the occurrence of “iv” outside the word window considered before a drug annotation. Similarly, the second example illustrates missed information (“PO” occurring outside the word window considered after a drug annotation).

False positives at PatientFM

In Table 4.10, we present several sentences extracted from the clinical notes that are good representative examples of how our approaches gave false positives. In the first example, it detected the word “children” related with the pronoun “he”, which provided a high probability of the sentence discussing the patient’s children. However, the pronoun “he” references to the patient’s father and in reality, these children are the patient’s half-siblings, a family member not considered by us. The second example is similar to the previous one. The problem is the main subject of the sentence. In this case, the daughter is not the patient’s daughter but instead the

great-aunt's daughter. Also, the words “maternal/paternal” before the great-aunt in the text are a good example of how the system can lose context.

Tab. 4.10.: Analyses of the most common false positives.

Relative		Sentences from the clinical note
Child	NA	Mr. Parsons' father, age 59, suffers from diabetes and has an elevated cholesterol level. He has several children through several other women...
Daughter	NA	The maternal/paternal great-aunt who was affected with ovarian cancer had three children. One of these individuals had a cancer of an unknown type and is deceased. The second daughter is the individual with ovarian cancer who was BRCA tested...
Cousin	NA	While living in Alabama, they lived with extended family, including Gabriel's grandparents, two aunts, and one cousin . Mom reports that she and Gabriel have always been open about their relationship with the children and show affection appropriately in front of the children...
Parent	NA	William's parents are both reportedly healthy at age 63, but they have not seen a physician in approximately 30 years. William's mother had one second trimester miscarriage...
Sibling	NA	Lucas's father is a 38-year-old man who is a college graduate and who has a total of 12 siblings ...

The third example shows the occurrence of a cousin in the text without detailing the family side. However, in this case, the information in the clinical note only indicates that the patient lived with the cousin. There is no clinical history about him in the text, so this member should be considered. In the fourth example, a new problem is presented. In this case, the clinical note is referring to the patient's parents. However, in that sentence, there is no clinical information about them, and in the next sentence, there is some information about the health of the patient's mother. Thus, in these scenarios, the family member to consider is the mother. The fifth example is similar to the first, the “siblings” in the text are the patient's uncles or aunts, but without mention of the right gender.

4.8 Final considerations

Clinical text enriches and expands physicians' knowledge about their patients. During patient admission, important information is recorded in the clinical notes which are not currently exploited in medical studies. Although several initiatives to improve information extraction in clinical text exist, this data is not commonly used to reach new health findings in observational studies.

This chapter proposed a methodology that was implemented in the Python programming language aiming at: i) the extraction of medication information in clinical notes; and ii) the migration of extracted information into a relational standard data model. Complementary to this, two other systems were proposed, to normalise multi-language concepts, and to extract patients' relatives' medical information. The former aimed to optimise the harmonisation process by providing more accurate mapping suggestions that can reduce the manual mapping and validation phase performed by the researchers. The latter had the objective of increasing the clinical decision support system to provide more precise outcomes, as well as to predict the probability of the patient suffering from hereditary diseases.

These tools promote new strategies to automatically annotate large amounts of EHR data. We also created new opportunities mainly related to EHR exploration by fostering the discovery of new relationships and pathways between diseases and parental phenotypes.

Scalable database profiling for multicentre studies

Database profiling allows extracting relevant characteristics from databases without revealing their contents. The collected metadata can then be stored and searched in specific data catalogues. However, when dealing with health databases, keeping these catalogues updated, while exposing enough information without raising privacy issues is a challenging task. In this chapter, we propose a strategy to help data owners publish characteristics about their databases. A centralised platform has been proposed to simplify the discovery of these data sources. We also proposed complementary strategies to enhance this platform by providing tools to orchestrate multicentre studies, as well as, to select the most suitable databases for the study.

The secondary use of health data is a research strategy applied in some observational studies. This has the potential to expand the knowledge about medical procedures and the efficiency of treatments for specific diseases, which may lead to more personalized healthcare [73, 4]. One of the challenges when reusing health databases for research is the correct selection of the data sources. This is a complex problem since it requires strategies to characterize data sources without revealing their content, and platforms for disseminating the databases' characteristics [157, 158]. For the database characterization issue, there are already some guidelines when dealing with this type of data. Depending on the project or institution's policies, the data owners can share aggregated information about their data. This can provide a summarization of the patients in the databases. Other characteristics can also be provided, namely data governance policies and contact details. These summarization guidelines are not standard and may differ depending on the context. For instance, a community focused on studying Alzheimer's Disease would have datasets with different characteristics compared with a more generic domain [70]. Figure 5.1 represents the main idea of the concept of this summarization, which is defined as fingerprinting.

Profiling databases (or fingerprinting) is the action of representing a database using a set of characteristics that combined can create a singular conception of the database. Defining these characteristics raises some issues that vary depending on the project scope. While these issues have complex solutions, we propose a different strategy

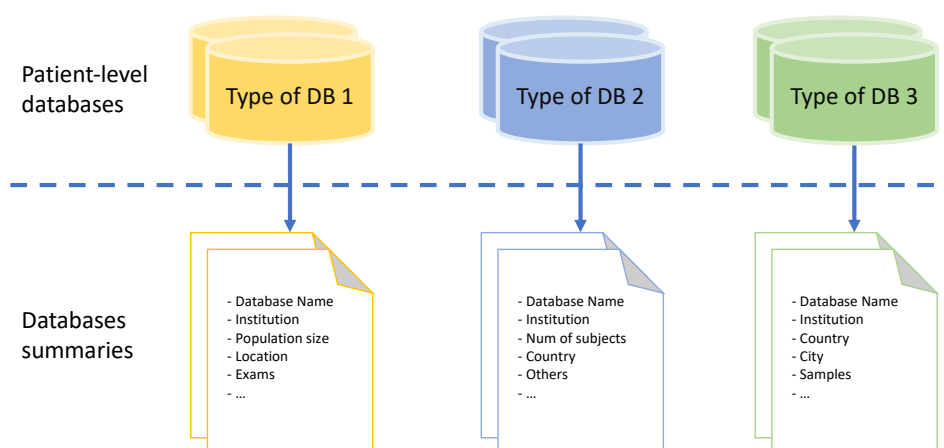


Fig. 5.1.: The concept of fingerprinting databases focuses on extracting characterises from databases of the same type.

to help the discovery of medical databases. It aims to provide enough information about the databases, that can characterise them at a deeper level, without sharing sensitive information.

5.1 Contribution

This final chapter continues the work described in the previous chapters. Although the solutions proposed in this chapter can be generalised, it was assumed the adoption of OMOP CDM as the standard data schema. In this chapter, it is proposed a solution for streamlining multicentre studies, by profiling, publishing and sharing metadata regarding OMOP CDM. It is also proposed a solution for supporting the database selection and for coordinating multicentre between all the entities involved. Summarily, our main contributions in this domain are the proposal of:

- A platform for cataloguing metadata extracted from biomedical databases, focusing on data sharing (denominated as MONTRA 2). This solution was originally created for EMIF project, and it was extended in this work to handle new challenges raised within EHDEN project, namely to also include features to support study management.
- A visual platform to help researchers obtain deeper information about each data source and about the data network. This tool aims to complement the information available on MONTRA 2. It can also support the selection of databases for specific research questions. The tool was developed within EHDEN

project, in close collaboration with all project stakeholders, including OHD-SI. The source code is currently available at <https://github.com/EHDEN/NetworkDashboards>.

- A command-line solution to extract relevant information from ACHILLES output, to populate the EHDEN Network Dashboards. This is a python-based tool that is available at <https://github.com/bioinformatics-ua/AchillesLite>. This tool was later replaced by CatalogueExport¹, which was developed by EHDEN partners.

This chapter is mainly based on the following publications:

- **João Rafael Almeida**, Eriksson Monteiro, Luís Bastião Silva, Alejandro Pazos Sierra and José Luís Oliveira, *A Recommender System to Help Discovering Cohorts in Rare Diseases*, in proceedings of the IEEE 33rd International Symposium on Computer-Based Medical Systems, 2020, DOI: 10.1109/CBMS49503.2020.00012;
- **João Rafael Almeida**, João Paulo Barraca and José Luís Oliveira, *A secure architecture for exploring patient-level databases from distributed institutions*, in proceedings of the IEEE 35th International Symposium on Computer-Based Medical Systems, 2022, DOI: 10.1109/CBMS55023.2022.00086;
- **João Rafael Almeida** and José Luís Oliveira, *MONTRA 2: A flexible framework for profiling health databases*, Submitted.

5.2 Background

Multicentre studies are usually formed by several steps, as described in Section 2.1.3. Although this pipeline is divided into seven steps, we recognised that it can be technically supported by dividing the problem into three types of applications: i) tools for database profiling; ii) web platforms for publishing databases' metadata; and iii) mechanisms to streamline studies between all the entities involved. Therefore, in this section, we present the current strategies used for profiling databases, existing platforms to enable the discovery of these databases, and the state-of-the-art tools to orchestrate tasks and workflows when conducting a multicentre study.

¹<https://github.com/EHDEN/CatalogueExport>

5.2.1 Database profiling

Characterising databases is a process that, if done manually, is time-consuming and may not produce the best results, *i.e.* the characterisation may not contain useful data to help researchers selecting data sources. Therefore, to help exposing information about data sources, without revealing sensitive information, we analysed which tools are currently available for database profiling.

DataMed is a tool composed of two major components: i) data ingestion and indexing pipelines; and ii) searching engine component [159]. This tool aims to build a data discovery indexing system, to support users when searching for existent datasets spread across repositories. The first component of this tool consists of a metadata ingestion pipeline, with extracting, mapping and indexing features. It adopts a unified data model, designated as Data Tag Suite (DATS). This model was developed based on the community inputs, and the analysis of the existing metadata from the most common data repositories. It is used to describe the metadata of the data sources, including its structure [160]. Considering that different data sources may have distinct data schemas, this tool was developed to extract this information following abstract retrieval modes and data formats. The implementation is based on different ingestors created to extract specific data formats, that are then combined. Each ingestor transforms the original data, through a ETL procedure to the DATS model. Although this system presents great flexibility, we have defined that the common data model used to support the data sources in this scope is the OMOP CDM. Gonzalez-Beltran *et al.* [161] have already done some work mapping OMOP CDM datasets into DATS.

Xtract is a serverless middleware developed to extract metadata from distributed files, enabling the centralisation of the indexed information [162]. It aims to create a scalable and decentralized metadata extraction system. This solution enables the automatization of processes to create searchable data hubs from disorganised repositories. Xtract uses built-in extractors that can identify the values on tabular files, *i.e.*, determining if there are null values in these tabular structures, or extracting keywords from unstructured files. The metadata is extracted using a crawler to fetch the files' properties in a repository. Therefore, this tool can dynamically profile files, when these are added to the repository. To simplify the integration with other tools, Xtract has some endpoints to execute the functions of extraction from the registered repositories, which can be deployed across heterogeneous data sources.

Skluma is another metadata extractor, developed to handle disorganised data [163]. It can gather metadata information from scientific sources automatically, supporting different systems and repositories. Skluma is mainly composed of three components: i) a crawler for processing data collections; ii) extractors to obtain the files' metadata; and iii) an orchestrator to launch the crawler, manage the extractors and expose an Application Programming Interfaces (API). This API enables the integration of this tool in third-party solutions, namely to request metadata extractions. The output of this extraction pipeline is a JSON document, that includes all the extracted metadata. This information can be then used for cataloguing the repositories, enabling some search features.

In the OHDSI ecosystem, there are some data analytical tools to support research studies. The OHDSI Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems (ACHILLES) was created to characterise OMOP CDM databases, producing summaries about the database content. The information present in the outputs of this tool is very valuable to the OHDSI research community, since this tool focus on extracting characteristics to support the study's feasibility in prior stages. Besides this characterisation, this tool also enables a quality assessment of the data present in the database. It is implemented using R programming to execute a set of SQL queries, defined within this community. Since we focused this work on harmonising the health information to the OMOP CDM, this tool is more valuable for our work compared with the previously described.

One of the main concerns regarding the ACHILLES outcomes is the data sensibility that can be included in the output. Due to this lack of confidence from the data owners, this tool is used and the results are kept private. One of the goals of this characterisation is to generate a reliable profile, in order to support medical researchers in selecting the databases of interest. Therefore, the CatalogueExport² was proposed. This package limits the information extracted from the databases to the minimal necessary that enables their characterisation to understand their feasibility for a new study. Therefore, in this work, we contributed to the development of this package in order to incorporate it into the methodologies that we propose in this chapter.

²<https://github.com/EHDEN/CatalogueExport>

5.2.2 Discovery of medical databases

Database profiling is the first step in assisting with the discovery of medical databases. There are already some solutions to simplify this discovery by publishing the databases' metadata. Trifan *et al.* [164] conducted a study to identify possible solutions for this problem in the biomedical field. These authors identified 20 unique publications focused on data discovery solutions, which are mainly focused on exposing different levels of aggregated information to web platforms. From the identified platforms in this work, we have selected those that are open-source and designed for more general-purpose data sharing, namely Cafe Variome, FAIRSharing and EMIF Catalogue.

Cafe Variome is a data discovery platform designed for general purposes. This platform is prepared to be adopted by any data owner, enabling the sensitive content discovery [165]. The platform can be customised for different domains, due to its flexibility. It has RBAC and different levels of data access, namely: i) open access, in which the researcher is allowed to see the data; ii) linked access, which is a scenario where the researcher can only access the data through an external data source link; and iii) restricted access, which is reserved to users with permissions. From a technical point of view, this platform was developed following the design principles, which can be enhanced with new features easier. However, it does not provide any SDK to simplify the integration of features for managing network studies.

FAIRsharing is an enhanced version of the BioSharing [166] platform. In its current version, it is a web informative and educational resource to describe and interlink community-driven standards, databases, repositories and data policies [167]. This platform presents these types of data, detailing the relations between them. The records on this platform are manually curated, *i.e.*, the data owner needs to characterize their databases and publish the metadata manually. The source code for this platform and adjacent tools is available on GitHub. Therefore, this tool is a potential candidate to serve as the base for metadata publishing and database selection in the study workflow.

Dataverse is one of the tools that was not included in the systematic review, but that is an open-source web platform to store, share, explore and analyze research data [168]. This platform is not close to the medical domain, as it is more of a generic platform for integrating meta-information about heterogeneous data sources. The purpose of this application is to provide a ready-to-use system, that can be deployed

on the institution's infrastructures for publishing information about the datasets. The application is another candidate to support the catalogue features of our proposal.

EMIF Catalogue is an online platform developed in the context of the EMIF project, designed as a marketplace of biomedical databases. In this platform, the databases are the main entity to be characterized and presented as possible data sources for conducting studies. This platform supports the concept of community, which is used to support distinct projects. The community concept allows data owners to characterize their databases in more refined domains. This platform follows the Findability, Accessibility, Interoperability, and Reusability (FAIR) principles in order to take advantage of data and enforce data reuse and interoperability [158]. The EMIF Catalogue core is based on MONTRA Framework, which is a web system developed adopting a plugin-based architecture to allow dynamic composition of services over-represented datasets [169]. This framework is another potential candidate to serve as the base of our work due to its flexibility and adoption by the community.

5.2.3 Streamlining multicentre studies

The last component of this procedure aims to streamline studies between all the entities involved. To increase the study quality, when working on each of the phases described in Section 2.1.3, these must be carefully planned. This usually involves a multi-disciplinary team of statisticians, clinical researchers and laboratory scientists, among others [170].

To gain access to clinical digital data, researchers have to deal with complex processes that include study submission, governance approval, data harmonisation, data extraction and many other tasks [171, 172, 173]. This process can be simplified by using task and workflow management systems. Furthermore, they can also be used to streamline all the processes associated with a health research study.

Scientific workflow systems allow the composition and execution of a set of computational processes, in cascade, and over a distributed environment. Some of these systems may be used to simplify research studies [174, 175, 18]. Taverna³ is a scientific workflow management system, available as a suite of open-source tools, which is used to facilitate computer simulation of repeatable scientific experiments. It can be executed in a self-hosted server or as a desktop client. The system follows a

³<https://taverna.apache.org>

Service-Oriented Architecture (SOA) approach, which makes the various web interfaces available for external software integration. It is a highly specialised and widely adopted platform, but is less suited to the diverse set of steps in a typical health research study [176].

Galaxy⁴ is another popular scientific workflow management system. This cloud-based platform is oriented to facilitate the execution of computational processes over biomedical datasets. The main purpose of the system is to be easy to use by people without technological knowledge, to allow reproducibility of experiments and to facilitate sharing of results. Galaxy integrates external tools in an user-friendly web interface, allowing the linear cascading of processes and providing, at the same time, access to several bioinformatics datasets. It allows collaborative discussion of results and studies' replication, but the system architecture is mainly oriented to computational process pipelines [177].

Besides these two scientific-oriented applications, there are several workflow management systems with a broader scope. However, most of them are commercial and do not allow integration with other external systems. Wrike⁵, for instance, is a collaborative platform, where users can assign tasks and track deadlines and schedules. It follows the workflow model and allows integration with document management solutions. Asana⁶ is another cloud-based solution, targeted at project and task management, which can be helpful for teams that handle multiple projects at the same time.

Whenever integration within another system is the main requirement [178], a workflow engine may be a good solution. This kind of engine does not offer a ready-to-use solution, but only the base blocks to build the final system. Although this brings the obvious disadvantage of having to develop the end-user application, it also brings several advantages, mainly due to the flexibility to integrate other software modules.

FireWorks⁷ is another open-source project for management and execution of scientific workflows [179]. It provides integration with other task queuing platforms, but is focused mostly on parallel work execution and job scripting and processing.

⁴<https://galaxyproject.org>

⁵<https://www.wrike.com>

⁶<https://www.asana.com>

⁷<https://github.com/materialsproject/fireworks>

jBPM⁸ is an open-source business process management suite, which runs as a Java EE application to execute repeatable workflows [180]. The system supports multi-user collaboration, using groups of users, but its configuration is rather complex for users without technical skills. The Activiti BPMN platform⁹ is a lightweight engine focused on open source Business Process Management (BPM), targeted at the needs of business professionals, developers and system administrators. This platform allows complex repeatable workflows with different kinds of tasks, but with only one assignee at a time, even though it enables reassignments in the middle of a process.

These task and workflow-oriented systems have distinct features and goals, and there is a need to combine some key aspects of both systems, namely asynchronous manual/automatic tasks and the integration with external tools. Furthermore, existing workflow engines do not support multi-user features such as users' collaboration over the same workflow, discussion of results and workflow sharing between different users.

5.3 Framework for profiling databases

The first version of the MONTRA Framework was created in the context of the EMIF project aiming to support biomedical data sharing. In its first version, MONTRA had the potential to simplify the creation of web-based catalogues, independently of the data scenario. However, the EHDEN project had other requirements, demanding a refactoring of the system core to incorporate the needs of project stakeholders.

5.3.1 Functional requirements

Based on the needs of the EHDEN partners, we defined a set of functional requirements that need to be addressed to fill the existent needs of this project. These requirements are the following:

- Data discovery: The system needs to facilitate data discovery, providing a set of features that support users in identifying data sources aligned with their researcher interests. From a technical perspective, these features include exposing metadata in a catalogue, searching mechanisms and metadata comparison.

⁸<http://www.jbpm.org/>

⁹<https://www.activiti.org/>

This is an umbrella requirement that can be split into specific features defined in the development roadmap, that we decided to omit in this document.

- **Dynamic skeleton:** Although we are interested in exposing OMOP CDM databases due to the research application, the system should have mechanisms to store metadata following a dynamic structure. In other words, the catalogue skeleton should be defined and updated without recoding the system. This flexibility is essential to keep the system compliant with different types of information, that can vary over time. For instance, new updates in the OMOP CDM schema, or in case of adding non-interoperable data sources (that are not completely migrated to OMOP CDM but can bring value to the network). Therefore, the system architecture should enable the definition of catalogue skeletons by users, with the support of daily tools, for instance, a spreadsheet.
- **Data aggregation:** Since the number of characteristics to define a data source can grow, the system should label or group concepts that belong to the same topic. This aggregation should support associations to Resource Description Framework (RDF) ontologies, if necessary.
- **Communities:** A different level of data aggregation is through communities, *i.e.* the existence of disease-specific communities in the system. These should be maintained by entities responsible for moderating their communities. This concept should segregate data and users within a specific scope.
- **Data visualisation:** The system should integrate graphical features to characterise the databases visually, *e.g.* through a web dashboard.
- **Privacy and data security:** The information extracted and exposed from the databases should not violate subjects' privacy. Since the use cases for this system are focused on clinical data, ensuring data privacy and security is essential.
- **Access control policies:** Although the system should not expose sensitive information, it may contain different levels of information that can be exposed to different groups of users. Therefore, the system should incorporate access control policies, that are controlled by an entity in the system.
- **Third-party integration:** In clinical research, different tools are requested depending on the studies' scope. To simplify the aggregation of all the required

tools by medical researchers, the system should incorporate a feature to easily integrate third-party applications.

- **Single sign-on integration:** Some of the tools may have built-in authentication services, while others may support Single Sign-On (SSO) integration. One of the system requirements is to be compliant with the most modern SSO authentication protocols used in web applications.
- **Metrics:** The system utilisation should produce and store a history of actions that can be used as statistical information about the system usage. For instance, for enhancing the system usability, namely by incorporating recommender systems.

5.3.2 System overview

MONTRA 2 aims to be a Rapid Application Development (RAD) [181] system, enabling the fast development of solutions for specific use cases. RAD systems evolve with the project requirements, unlikely the conventional software solutions. This type of software does not rely on rigid specifications, as is the case of critical software. Instead, it is developed to be continuously adjustable to the users' needs, and the development follows the general guidelines for biomedical software development [182] to fit new requirements without rebuilding the system core.

Main components

MONTRA 2 is built in a three-tier software architecture, as illustrated in Figure 5.2. The top layer is responsible for the user interface interactions, including integration with third-party applications. At this level, five components were implemented: i) system management; ii) community management; iii) browse data catalogue; iv) fingerprint templating; and v) API web services.

The system management component aims to define policies to control all the system's features that are associated with database operations. The community management component also contains administrative features, however, these are limited to the community scope. For instance, community visibility, plugins, users, and databases, among other operations within the community.

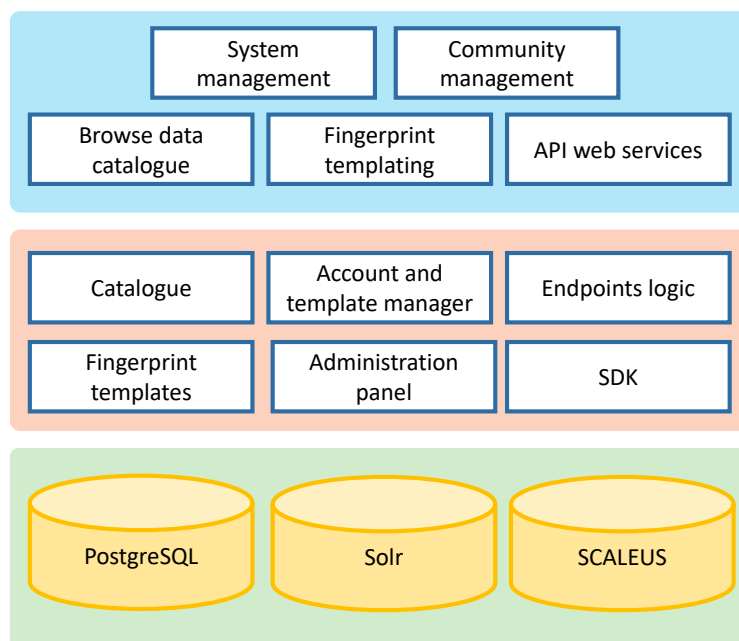


Fig. 5.2.: Three-layer MONTRA 2 architecture that includes: i) presentation tier, which is represent on the top (blue box); ii) logic tier, in the middle (red box); and iii) data tier, in the bottom (green box).

The module for browsing the data catalogue creates a workspace where researchers can navigate the databases' characteristics of a specific community. The features of this module include searching and querying operations over the exposed metadata, as well as, data aggregations and comparison.

The interface for the creation, visualisation and edition of database fingerprints is the responsibility of the fingerprint templating component. Since the fingerprint's structure is defined dynamically, the data is stored in a dynamic schema. This data schema is generated from a skeleton, that is defined by non-technical users, using the spreadsheet format. Therefore, data owners can use daily tools, such as Microsoft Excel, to edit and construct the skeleton template that can be imported into the system to generate the fingerprint structure. All these operations are supported by the fingerprint templating module. The last component, responsible for handling and exposing API services, is better detailed in Section 5.3.5.

The logic layer contains the business logic and the models of the system's entities. This tier is responsible for defining the template schemas, community and fingerprint instances, access control policies and endpoints that can be used by the web API or the SDK. The models defined in this tier communicate directly with the different components used in the data tier. For instance, users' data is stored in the PostgreSQL

database for account management. Fingerprint data are also stored in this database. However, the information stored in the tables associated with fingerprints is also indexed in an Apache Solr instance. SCALEUS play a different role in this architecture, which is better described in Section 5.3.5.

Technologies

MONTRA 2 was developed using Django¹⁰, a python-based web framework, which encourages rapid development and supports clean programmatic design. The user interfaces were developed using front-end technologies, namely supported by Hyper-Text Markup Language (HTML), Cascading Style Sheets (CSS) and JavaScript. The system web interfaces adopted Bootstrap¹¹, which is an open-source and responsive CSS framework.

In the backend, the system incorporated additional components to Django, namely for optimisation. All heavy tasks, that take some time to be completed were added to a queue message system, namely RabbitMQ¹². Celery¹³ was responsible for executing these tasks in the backend.

The deployment of MONTRA 2 is based on containerized technologies, namely Docker¹⁴. To simplify the orchestration of all components, docker-compose specifications were used.

5.3.3 MONTRA Software Development Kit (SDK)

To simplify the integrations of third-party applications, MONTRA 2 includes a SDK. It enables the creation and integration of additional components, without changing the system core. These components, designed as plugins, can be integrated at different levels of the system. There are essentially three types of plugins, namely global, database-related and third-party full-fledged.

Global plugins usually reflect a general view of all the databases for a given user. These plugins are available at the root of the platform. They provide information that takes into consideration all the databases, and aggregate data in a certain way.

¹⁰<https://www.djangoproject.com>

¹¹<https://getbootstrap.com>

¹²<https://www.rabbitmq.com>

¹³<https://docs.celeryq.dev/en/stable/>

¹⁴<https://www.docker.com>

The database-related plugins only reflect views over specific database data. These plugins are available when a user opens a specific database. Finally, the third-party full-fledged applications do not have any real data integration with the platform. They are complete, full-fledged applications that are linked to the system, through the navigation menu. They usually provide completely different functionality, though they may share some features like authentication with the platform. The main goal of these plugins is to integrate their application features into the environment.

The MONTRA SDK offers an environment in which is possible to create new plugins based on these types. The plugin system has a very simple lifecycle, based on a rendering system. After the initial rendering, each time an event occurs, the plugin updates the content. This behaviour is caught through the usual javascript event listeners, as shown in Figure 5.3. Therefore, integrating a new plugin in the system core needs to respect this life cycle, which is very common in several JavaScript Web Frameworks.

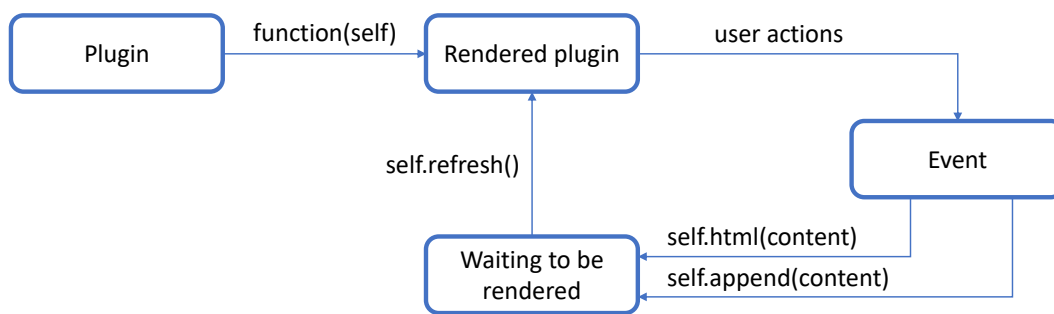


Fig. 5.3.: The life cycle of MONTRA 2 plugins.

5.3.4 Data representation

Data owners need to answer a set of questions about the databases when publishing the database characteristics. This information is then exposed in the data catalogue. Additionally, the data owner can upload a generated file to populate an additional system to show the data in a dashboard format. Both strategies are briefly described in this section.

Database catalogue

The database catalogue can be considered one of the core features of MONTRA 2. In this catalogue, it is represented each database through the concept of fingerprinting, as it was already described. Therefore, the data owners can define the catalogue

structure that better fits their needs in that scope, and the system generates the web catalogue based on that file. Figure 5.4 represents the view of an empty fingerprint to insert the database characteristics.

Fig. 5.4.: View of an empty fingerprint of the database characteristics with several categories of questions available.

The skeleton structure is flexible and contains fields (questions) to be filled by the data owners. Several questions can be aggregated in a “QuestionSet”, creating a hierarchical data representation. Each question can store different types of data, for instance, dates, numbers, strings, multiple-choice values, geographic location, among others.

These fields, which represent the metadata about the health databases in the catalogue, are used for free text search, advance search, dataset comparison, and other features of the catalogue. Figure 5.5 shows one of the many views of the database catalogue showing the list of databases.

Graphical dashboards

An additional component of the database catalogue, is the Network Dashboards. This component was developed outside of the MONTRA architecture, but integrated using the MONTRA SDK. It allow data owners to upload a CSV file with aggregated data that characterise databases in a deeper level than the fingerprint. The file is exported from the CatalogueExport, a tool with similar features as ACHILLES, that allows data partners to control the data they expose.

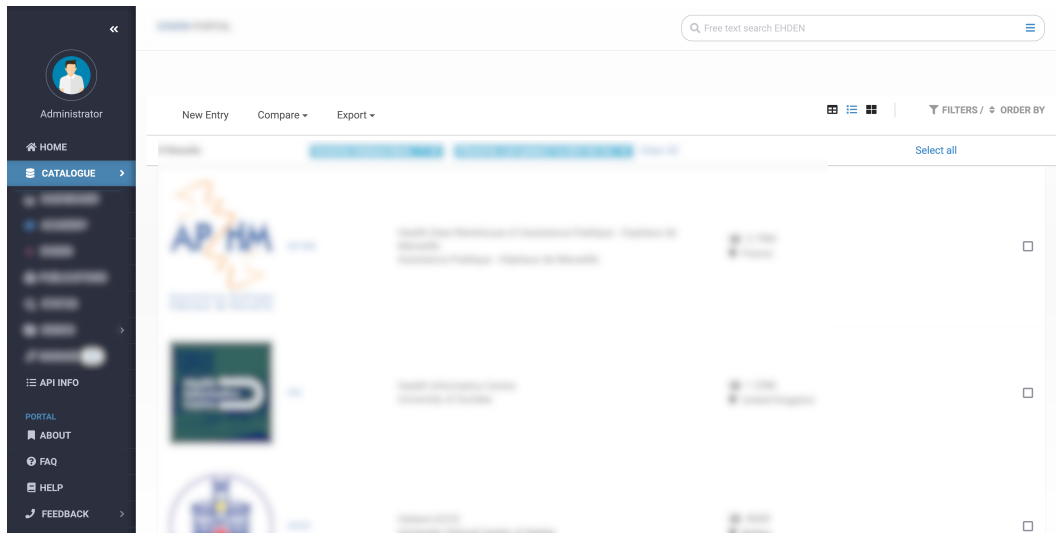


Fig. 5.5.: Database catalogue list view (intentionally blurred due to privacy issues).

This tool was developed aiming to display the OMOP CDM characteristics in a visual form. However, to simplify the uploading procedures, we developed specific a user-friendly interface. This component also validates the format of the data and automatically integrates it into existing data. This also triggers a process for updating the graphic visualisations.

The component responsible for rendering the charts and dashboards was Apache Superset¹⁵, an open-source visualisation platform with a rich set of graphs, filtering and cross-filtering, and easy to customize. Each visualisation in Superset is backed by a SQL query. Since Superset requests the data from the database every time a chart is rendered, the information present in the dashboards is updated by refreshing the database content when a new upload is concluded.

The tool can aggregate data from several OMOP CDM databases, simplifying the comparison process of databases from the OHDSI community through graphical dashboards. It generates two types of dashboards: i) database-level dashboard; and ii) network dashboard. The database-level dashboard (Figure 5.6) contains a set of charts capable of providing a quick overview of the database content, similar to some of the charts used in Achilles Web tool. This dashboard enables the analysis of data from a single database.

¹⁵<https://superset.apache.org>

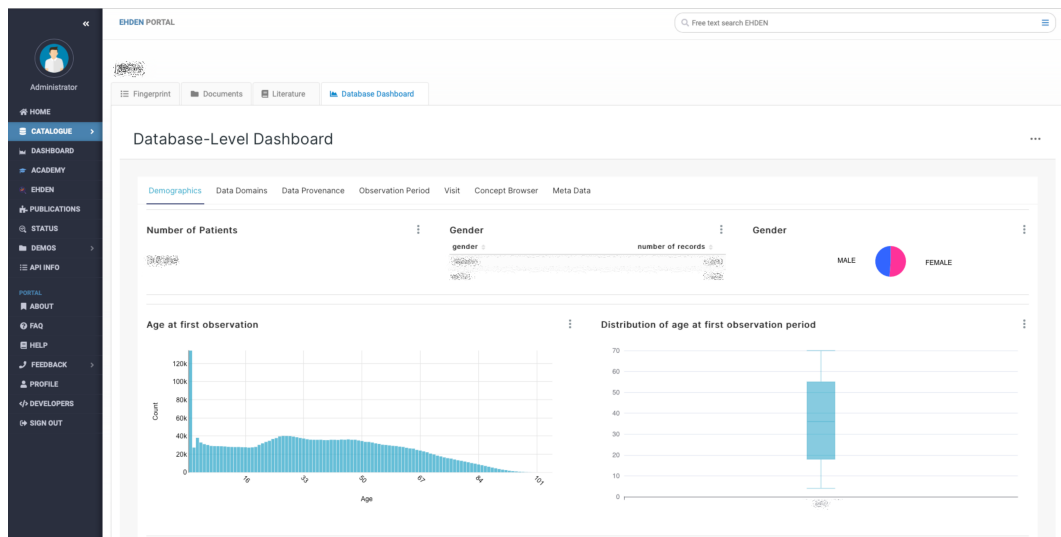


Fig. 5.6.: Overview of the Database-level Dashboard (intentionally blurred due to privacy issues).

The network dashboard (Figure 5.7) displays data from multiple databases, enabling data comparison. The visualisations within the dashboards are divided into the following characteristics groups:

- Demographics: charts that show the distribution of gender and age;
- Data domains: visualisations to analyse the distribution of data domains;
- Data provenance: charts to show where the data is originating from;
- Observation period: visualisations showing the distributions of patients' observations period;
- Visit: visualisations to compare visit occurrence records;
- Concept Browser: charts to analyse concept data.

Both the Database and the Network dashboards follow the same group structure with slight differences. The latter includes two additional groups of charts: one that contains overall metrics of the network and another with some information about the system. The database level dashboard has an additional Metadata group containing extra information about the files uploaded. Furthermore, the network

level dashboard contains several filters, allowing the user to dynamically restrict the data being displayed on the visualisations.

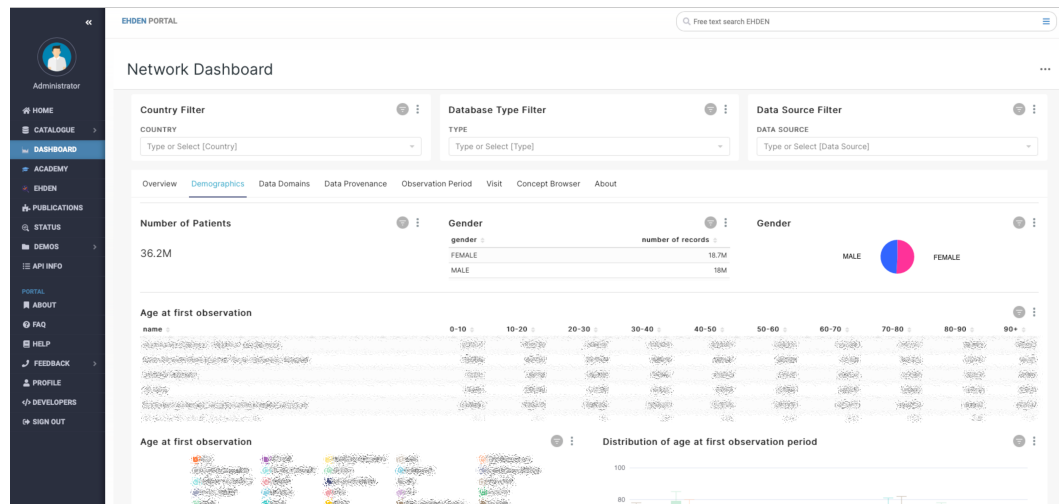


Fig. 5.7.: Overview of the Network Dashboard (intentionally blurred due to privacy issues).

5.3.5 Endpoints for interoperability

The systems' interoperability with other database catalogues is simplified with the creation of endpoints. A RESTful API can simplify communication with other applications, as well as, can be used to provide federated endpoints, if these followed a federated specification.

RESTful API

By providing a RESTful API, MONTRA 2 makes available endpoints to be consumed by plugins or third-party applications. For instance, it can make available metadata about the registered databases in the catalogue, that can be consulted using the defined endpoints.

The API is essential for developing plugins that use parts of the databases' characterisations, *e.g.* to populate specific information in the dashboards previously described, or to support the Study Manager proposed in Section 5.5. More details about the authentication mechanisms to access the API are available in Section 5.3.6.

Federated endpoint support

A catalogue of biomedical datasets, such as those that can be built using MONTRA 2, provides users with a centralized access point to descriptions that help them make

decisions that have a profound impact on their research. Conveniently, these descriptions can be found using suitable user interfaces to facilitate this work. Mapping data in a semantic format using an ontology allows linking and relating the metadata, supporting federation of endpoints.

In the spreadsheet used to define the data schema that generates the database catalogue, data owners can also define the Uniform Resource Identifier (URI) that characterises each entry of the data skeleton, *e.g.* using Data Catalog Vocabulary (DCAT)¹⁶ to annotate essential information about the data sources described on the platform. The name of the database can be mapped to the DCAT property `http://purl.org/dc/terms/title`, and the `http://purl.org/dc/terms/accessRights` term provides access privileges and security status information.

Ontology repository

The management of multiple semantic datasets can be performed using a tool such as SCALEUS-FD, which allows the conversion of tabular data into semantic data. In addition to this primary function, it is a robust solution when used as an ontology repository. Software agents can load and access ontologies in SCALEUS-FD since it also offers a RESTful API to perform these operations [183].

The publication of ontologies must ensure that they can be registered or indexed by search engines. Their findability is crucial for researchers to benefit from their information. In addition, we need to ensure they can be accessed using open communication protocols that allow machine-machine interactions.

The databases' metadata normally includes data use conditions, *i.e.* how the data can be accessed and reused. To create access points to catalogues described by metadata and allow their interoperability, they must follow a standard vocabulary (*e.g.* DCAT).

5.3.6 Access control mechanisms

MONTRA 2 also contains distinct features to define access control policies. Since we aim to implement a Federated Identity Management (FIdM) solution for aggregating multiple stand-alone applications, we evaluated several standards that are currently used for this task, namely Security Assertion Markup Language (SAML) [184],

¹⁶<https://www.w3.org/TR/vocab-dcat-2/>

Open Authentication (OAuth) 2.0 [185] and OpenID Connect (OIDC) [186]. These standards share similar features, using security tokens in their services. The security tokens, also known as Identity Tokens, Authentication Tokens, and Authorisation Tokens, are the key concept in a FIdM implementation because they are responsible for authenticating and authorising users [187]. Therefore, in this work, we integrate the OIDC since it only requires the definition of a new account provider.

OpenID Connect (OIDC) support

The OIDC is an extension of the OAuth 2.0 protocol, more precisely an identity layer on the top of this protocol [188, 186]. This framework contains a group of specifications for transmitting users' identity using RESTful services [186], and facilitates the process of clients confirming the users' identity depending on a chosen OAuth 2.0 Authorization Server.

This protocol involves three parties, namely the Identity Provider (IdP), the Relying Parties (RP) and the users. The IdP manages users' accounts and authenticates them. An authenticated user can request an access token in the IdP in order to use it to log in to the RP [186]. Even though OIDC has more features than OAuth 2.0, some older systems only support the latter in their authentication components. MONTRA 2 includes a module for account management and another for SSO integration. The SSO supports OIDC which enabled the creation of a federated identity over multiple platforms, integrated into a MONTRA 2 instance.

Users' profiles and interactions

The system requires the existence of different profiles since there are specific features for the different roles. The administrator role has permission to manage the platform, including users, database fingerprints, and study flows.

The researcher role is granted to users that need access to the data. With this role, an user can see the database characteristics and create studies, but they cannot accept study requests, since this entity is not considered a data owner.

The data owner can create entries in the database catalogue, can upload dashboards statistics, and overall keeping updated the database characteristics. This entity is the only one capable of answering study requests.

Finally, the “new user” role can register in the system and ask for another role. Until one of the administrators approves this registry, this user does not have permission to access other features.

Role-based access control policies

Associated with the users’ accounts, the platform supports RBAC policies, *i.e.*, different roles can have different privileges. Complementary, the system also supports Access Control List (ACL) for managing access to each plugin. All these control mechanisms were implemented in the system, and only require the right configuration depending on the system objective.

The presented REST API also enforces the use of the access control mechanisms. To consume the available endpoints, the third-party applications need to have two distinct tokens: i) user token and ii) registry key. By combining both keys, access is granted, being possible to use the operations defined for each endpoint.

5.4 Recommending health databases

Researchers need to periodically analyse the updates in the available databases, looking for new datasets of interest. Manual filtering is required because new studies can be conducted following different practices, generating unrelated datasets focusing on the same disease. Aiming to simplify the correct identification of new data sources of interest, we proposed a solution to suggest similar datasets or publications to the users involved in a clinical study, augmenting the information of interest. This solution recommends new data sources based on user profiles, keeping researchers updated about similar studies conducted using data from the platform proposed in Section 5.3.

5.4.1 Feature extraction

Although we focused the work of this chapter in OMOP CDM databases, we recognise that some of the research studies do not follow a standard database schema. These studies are built for specific purposes, with particular inclusion and exclusion criteria, and are normally stored and maintained in ad-hoc solutions. To allow the reproducibility of research questions in different data sources, the proposed strategy uses a common template to characterise variables and values (columns and rows).

This method enables the comparison of multiple data sources using the medical concepts that were mapped into the ontology classes. Since this ontology supports relationships between concepts, this comparison is possible by following a hierarchical tree with root entities that represent core categories that are being followed (*i.e.* patient demographic data, neuropsychiatry, and laboratory results, among others). Moreover, the template is enough flexible to allow combining, extending or creating new variables. The ontology management is maintained by community managers using RDF format [189].

The concepts with top hierarchical position enable the calculation of the similarity as an anatomic method, *e.g.* if the patient demographics field contains two sub-concepts, both can be used to calculate the similarity of the demographics branch. Using as an example two data sources with different concepts mapped under the same branch, these two have some similarities since medical researchers have previously mapped concepts that can be found within the ontology branches. To classify this similarity, we attributed different weights to the levels in the ontology. Therefore, the lower the branch level, the higher the similarity between the concepts. This flexibility avoids structural weight inconsistency, for instance, if there are two variables in the same hierarchical position, that have the same purpose, they should have the same weight, to contribute equitably to the similarity score.

The Apache Solr provides a good foundation for a large-scale search engine and a basis to implement a useful and scalable recommender engine [190]. This framework was used in this context to index all the data source's features combined with the ontology, and the scientific publications related to those features extracted from external sources (PubMed/MEDLINE). With this framework, we can calculate the similarity between the data source using the Minimum Weighted Tree Reconstruction (MWTR) problem. This algorithm consists of discovering the minimum length weighted tree connection for a set of nodes [191]. The nodes in our ontology were the RDF classes, and the leaves were the mapped variables.

5.4.2 Collaborative filtering

Collaborative filtering in recommender systems produces target suggestions to users, based on patterns of usage or ratings. These suggestions are possible to make after collecting the preferences from several users that are considered with similar interests [192]. Therefore, users' rating history is essential to build the user profile. The

profile determines which users are similar within the system, by processing this as the nearest neighbourhood estimation problem [193, 194].

In an unrated system, like our catalogue, we can use the users' clicks to classify their interest in each database. This technique can produce considerable outcomes when there are a great diversity of active users on the platform. However, for users that rarely explore the data sources available on the system and focused their interactions on consulting a reduced number of databases, giving precise recommendations is more challenging.

This method is implemented using matrix factorisation of the matrix representing the pair user and score (*i.e.* user interest in a given database or article). Matrix factorization allows dealing with sparsity and scalability, which are the two biggest challenges in recommenders that use collaborative filtering features. One can define the dimensions of the latent feature space to keep control of the complexity of a model. Singular Value Decomposition (SVD) is one of the most used matrix factorization algorithms when implementing a collaborative filtering recommender system. Using SVD, users and scores are represented by latent feature vectors $q_i, p_u \in \mathbb{R}^k$ of dimensionality k . With this, the inner product of the latent vectors is used to predict the rating user u would give to the item i :

$$\hat{r}_{u_i} = q_i^T p_u \quad (5.1)$$

The use of this approach in the proposed system can be beneficial due to the diversity of users in the catalogue communities (users from several institutions). This diversity allows the definition of patterns from the institutional users. The fact that they were from the same institution increases the similarity of interests in the datasets, or publications.

5.4.3 Content-based retrieval

A content-based recommendation system tries to give a suggestion based on the user's rating and on item's content and their similarity. This is calculated based on the most relevant features [195]. To predict suggestions, the system uses these metrics considering that there is a relation between the items' similarity and the

user's preferences. This can be solved as a classification problem considering the users' likes and dislikes for each item [196].

The items in this proposal are the data sources, and their concepts are the features to compare the similarity. To identify the interest of the users in the data sources, we used a normalised metric based on the clicks to identify how much the user “likes” each data source. This originated a matrix with users and data sources, that we then defined as a binary classification task (with labels $C = \{c^+, c^-\}$) regarding the user preferences. This was solved as a classification problem where the classifier has to consider what data sources the users' likes (c^+) and dislikes (c^-) based on the items features [197].

The use of probabilistic methods to define user profiles is simple but effective. Bayesian classifiers can define a probabilistic model using previous data, which estimates a *posteriori* probability, $P(c|s)$, of data source (or study) s belonging to class c . This is calculated based on: i) the probability of observing an item with the label c , $P(c)$; ii) the probability of item s given class c , $P(s|c)$; and iii) the probability of observing item s , $P(s)$. Therefore, the Bayes theorem can be used to calculate $P(c|s)$ as:

$$P(c|s) = \frac{P(c)P(s|c)}{P(s)} \quad (5.2)$$

The label prediction for a new data source s is defined by the class with the highest probability using the function:

$$c = \underset{c_j}{\operatorname{argmax}} \frac{P(c_j)P(s|c_j)}{P(s)} \quad (5.3)$$

This approach is good to classify the data sources by their similarity and suggest others when there are just a few users on the platform. However, when there is a large range of concepts to compare, the diversity of data sources may interfere with the recommendation due to the lack of similar concepts.

5.4.4 System overview

The proposed recommendation system combines the two techniques presented to fill gaps of each isolate methodology. Collaborative filtering can detect similar user

profiles and provide recommendations when the data sources structure varies significantly. On the other hand, context-based retrieval can provide better suggestions, just relying only on the data sources' similarity. Therefore, we applied metrics to first measure each approach and then combine both.

In Figure 5.8, it is presented an example considering four users and three data sources. In this case, the recommender system needs to predict if user D should receive a suggestion to access the data source Z. Based on the clicks, the algorithm identifies a similarity with user A, concluding that this data source can be interesting for user D.








	 Medical Records X	 Medical Records Y	 Medical Records Z
 A	43	2	24
 B	32	25	4
 C	3	17	0
 D	26	1	?

Fig. 5.8.: The collaborative filtering matrix correlates the number of clicks from user A to user D, with the data sources in the catalogue.

Figure 5.9 aims to represent a matrix that classifies the similarity between three distinct data sources. An empirical similarity metric was defined, for a threshold superior to 0.7, indicating that the data source should be considered a data source of interest. After defining these values, the system calculates the product between the matrix to define the prediction score. If this value is superior to the threshold, the suggestion is presented to the user. For this example, the system should suggest the data source Z to users that usually access data source X.

5.5 Explore distributed patient-level databases

The methodology proposed to streamline the execution of multicentre studies is based on MONTRA 2. To accomplish this, we developed an additional tool that was integrated in MONTRA 2 as a plugin. It aims to simplify the execution of health studies as well as centralise and coordinate the operations between all the entities involved.







	 Medical Records X	 Medical Records Y	 Medical Records Z
 Medical Records X	1	0.189	0.731
 Medical Records Y	0.189	1	0.327
 Medical Records Z	0.731	0.327	1

Fig. 5.9.: The content-base matrix compares the similarity between all the data source. The higher the value, the higher will be the similarity.

5.5.1 Methodology overview

Figure 5.10 presents an overview of the proposed methodology for managing distributed queries. In the first step, the researcher defines the research question in a query builder on top of a synthetic database that shares the same schema of all the databases present in the network. The second and third steps are focused on selecting databases of interest in the catalogue. The fourth step is processed locally in each database and is responsible for the retrieval of the query results. This is a manual step after which the data owner can decide whether or not to share the results. The final step is the response to the query by each database owner, where the researcher can aggregate the results. Since this is just an overview of the proposed methodology for distributing queries, some concepts were omitted. For instance, the query builder can be a specialized tool prepared to work only in a specific domain. Therefore, in that scope, the presented methodology would integrate such tool to simplify the workflow.

5.5.2 Functional requirements

The analysis of the functional requirements was conducted based on the needs to have a study manager on the EHDEN project. These requirements are mainly the following:

- Defining new studies: Researchers should be able to define a new study. This definition requires the selection of the databases of interest, the study goal and the query or package that needs to be executed locally in each database. The

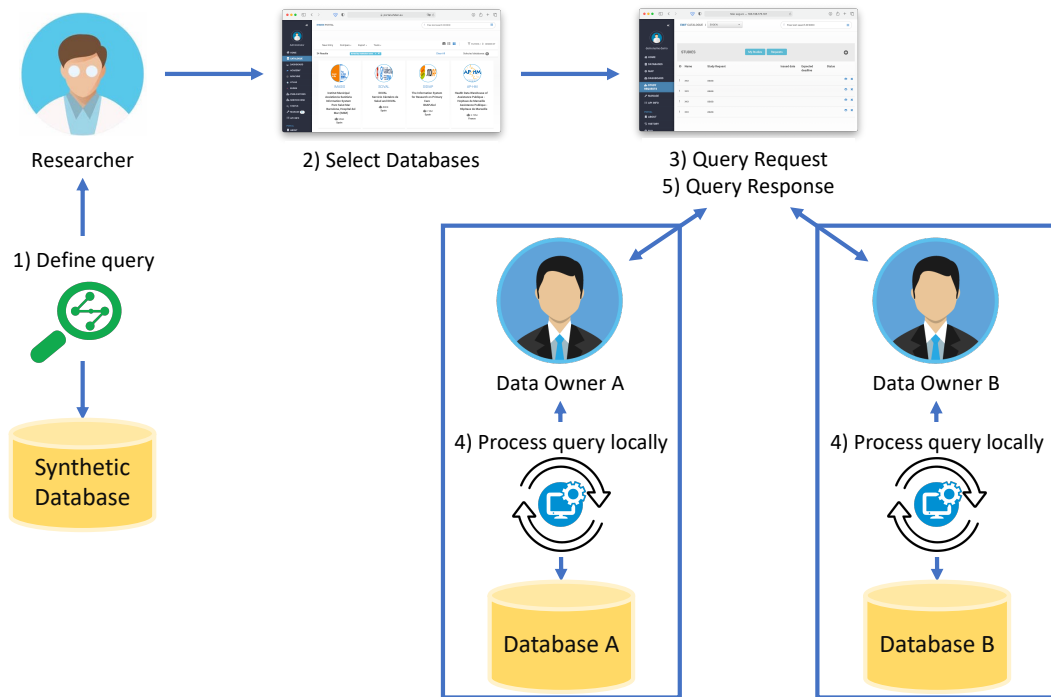


Fig. 5.10.: Methodology overview for managing the distributed queries. 1) the researcher defines the query using a synthetic database with a data schema similar to the ones maintained by the data owners. 2 and 3) are focused on selecting the databases and sharing the query with their owners. 4) is the processing of the query locally in each database. 5) is the response to the query by each database owner.

creation of a new study should notify the data owners, in order to alert them about the need for their support to execute the query in their facilities.

- **Uploading complementary information:** The study definition may need extra information regarding the study. For instance, documents with complementary information about the query, or data governance policies. To meet this requirement, the system should be capable of supporting the upload of documents during the study creation. This feature should support the upload of documents in the following formats: PDF, DOCX, XLSX or ZIP.
- **Keep the study's history:** With the system's evolution and continuous usage, several studies would be created. Therefore, the system should keep the history of studies conducted by each user, and enable the analysis and reuse of past studies and templates.
- **Managing study requests:** Data owners should be able to manage the study requests addressed to them. Although the study can be created, each data

owner should be able to decide about their participation in the study. Therefore, the system should be prepared to enable the management of study requests, namely to answer positively or to reject the participation in the study.

- Uploading data sets: When accepting the participation in the study, the system should provide a strategy to upload the datasets in order to be only accessible to the researcher. This requires the encryption of the dataset using symmetric and asymmetric algorithms.

5.5.3 Study Manager architecture

The proposed system, designated as Study Manager, adopted the same technologies used in MONTRA 2, namely Django in its core. To simplify the integration between systems, this tool was implemented to be compliant with the MONTRA SDK, following a Model-View-Controller (MVC) software pattern. This pattern segregates the application logic into three main elements: i) the model, responsible for handling the data storage; ii) the view, that generates the data representation for the client; and iii) the controller, which contains the business layer. This architecture is illustrated in Figure 5.11.

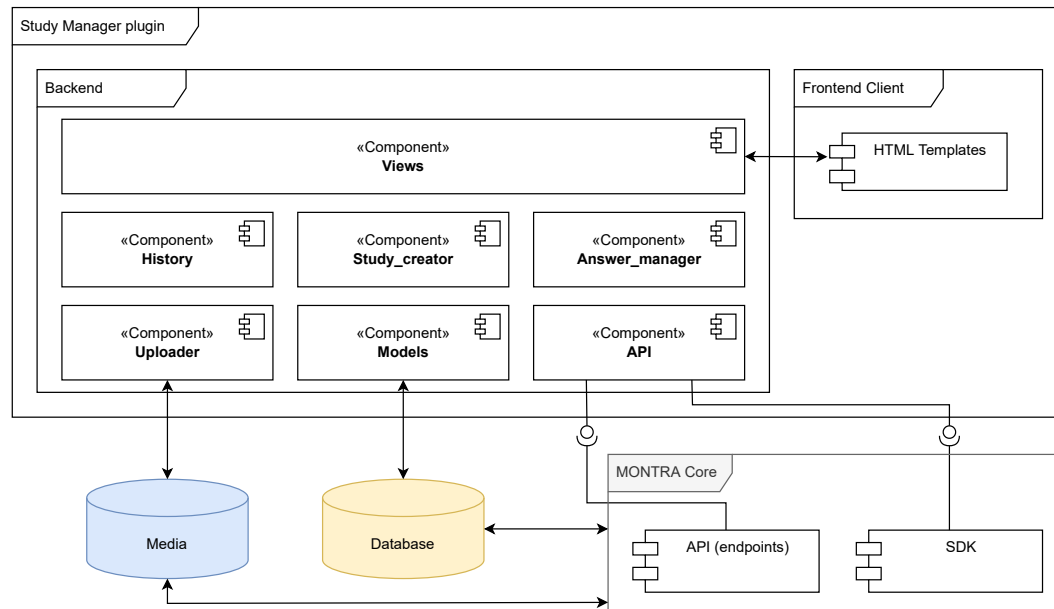


Fig. 5.11.: Architecture of the Study Manager plugin, containing the core components of this system and the integration into MONTRA 2 Core, using its SDK.

This application contains two components for managing the studies, namely the “Study_creator” and “Answer_manager”. To support these, the system has also a

component to keep the history of all actions, e.g., to keep track of the study status alterations. On top of these components, the “Views” component was created. It is responsible for handling the client requests and sending the necessary information to generate the HTML pages. The communication with MONTRA core is made using the endpoints provided in the API, enabling access to the databases’ characteristics available on the catalogue.

The “Uploader” component is responsible for ensuring persistence of the files uploaded during the study creation. The repository for storing these documents is the same as already used by MONTRA Core for other scenarios. Regarding the tabular data persistence, a component responsible for it was also created.

5.5.4 Features and user experience

To control the users access to this feature, a new role was created in the Montra framework - Study Manager. This plugin is only accessible to this group of users. The first view of this system contains a list of all studies created by the user. In this section, we detail the workflow for the main features associated with the core operations when conducting a research study.

Study creation

The first of the seven stages illustrated in Figure 2.4, is based on the definition of a research question, which is not entirely done in our system. However, the database catalogue is used to understand the study’s feasibility. The next stage aims to establish the study design and protocol. In this stage, it is defined the inclusion and exclusion criteria, and described the expected study outcomes. All this information is inserted in the Study Manager when creating a study. Before defining the study, researchers need to find the datasets of interest.

After concluding the definition of the study, the data owners are contacted (stage four). The study creation ends at this stage, but in the following sections, we described the features to handle stages five and six. Stage seven (marked in grey) was not included in our proposal.

Answering study requests

The stage five represented in Figure 2.4 is processed by the data owners that are accepted to be part of the study. The interactions with the system involve receiving,

processing and responding to a research question (supported by a query). These interactions are illustrated in the diagram represented in Figure 5.12. The data owner accesses pending requests in a web platform and processes the query locally against the protected databases. At this stage, this entity can analyse the outputs and decide if the results are compliant with the institution's privacy policies. In case they are compliant, the data owner encrypts and uploads the results in a moderated repository that is accessible by the researcher.

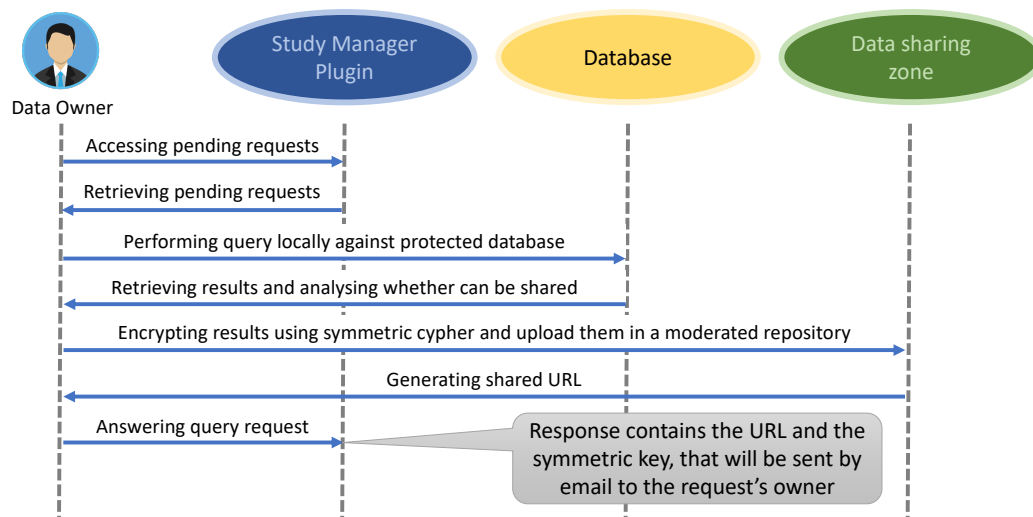


Fig. 5.12.: Interaction diagram where data owners process a query locally without exposing the database. The queries are received in the query manager plugin, processed locally and the results are uploaded in a sharing zone, that is moderated with credentials.

The idea is to use the Study Manager to streamline the study, by centralizing the information into a web platform. However, the patient-level data is not uploaded to this system. Only the study details and necessary information to access the data.

Data sharing

Due to the high number of solid solutions available for data storage and sharing, we decided to adopt one that is compliant with our requirements, and keep this part of the work open for further research. The process of encrypting the data is frequently based on symmetric algorithms, using a randomly generated key for each request. The key is then encrypted using an asymmetric algorithm such as RivestShamirAdleman (RSA). The latter can use the researcher's public key to ensure that only this entity can access the data, even if the cryptogram (the data) is intercepted. The use of symmetric algorithms for ciphering the dataset aims to optimise the encryption process since they are efficient in terms of computational performance. However, asymmetric algorithms are better for key distribution among peers.

Based on these conditions to encrypt and share the data, we rely on a cloud solution for file transfer: Tresorit¹⁷, that ensures that each file stored is encrypted individually, with different keys. The encryption process occurs before the file is uploaded and uses a 256-bit Advanced Encryption Standard (AES) cypher. For sharing, the system creates directories, in which the owners can give access to invited users, and these directories are encrypted using 4096-bit RSA public key algorithms. However, this is a cloud solution, which sometimes is not well accepted by the data owners.

Alternative solutions can be ownCloud¹⁸, or myQNAPcloud¹⁹, or a custom developed service. All can be installed in the data owner's institution with controlled access to the data using credentials or temporary links. In our methodology, we did not focus on the technical strategy for sharing the data, only on having centralized point for exchanging the information, *i.e.*, the access to the data repository. We used myQNAPcloud for testing purposes, however, in one of the research applications of this work, ownCloud was used instead.

Aggregation of query results

The procedure for aggregating the query outputs cannot be addressed by creating a single solution that would fit every scenario. Although this is an important topic for the proposed methodology, this task may require ad-hoc solutions. Depending on the data domain and data results structure, this process can be resumed by inserting records in a single table. In other cases where several tables need to be shared, this may require more complex actions, that need to be addressed using an ETL pipeline. An example of this last case, that is frequent in medical data, is when the databases use distinct standard vocabularies to represent the medical procedures, drugs, among others. This requires a harmonisation procedure when aggregating the query results.

From all the stages depicted in Figure 2.4, this represents part of the sixth task. Since we focused this work on medical data, and we developed this system based on the needs of the projects that motivated this work, we employed the data aggregation methods already used in these projects for this task. These methods are briefly described in the next section when presenting the research applications for this work.

¹⁷<https://tresorit.com>

¹⁸<https://owncloud.com>

¹⁹<https://www.myqnapcloud.com>

5.6 Results

The main result of this chapter is a platform for cataloguing metadata extracted from biomedical databases, facilitating data sharing. This result can be presented in two parts: i) portals for publishing and promoting the discovery of biomedical data; and ii) tools for orchestrating and streamlining multicentre studies.

5.6.1 Portals for biomedical data sharing

The framework created by this work is MONTRA 2, but since it is a software application, it can also be used to support different other projects. Currently, this system has three instances in production, to support different platforms, namely the EHDEN Portal, EMIF Catalogue and MSDA Portal.

EHDEN Portal

The EHDEN Portal²⁰ is a web platform that centralizes the entry points for all available services within the EHDEN project. The portal is currently online and provides an ideal starting point for a project portal to hold additional functionality and tools focussed on the services provided by EHDEN. For instance, information for the Small and Medium-sized Enterprise (SME), and analytical tools can be made available through the portal. It integrates a common security layer on top of the individual components. Access to the information available in this portal is protected using RBAC policies, and it is supported by LifeScience AAI.

This portal currently contains information about 76 databases, which are publicly accessible to all registered users, with 40 more being added to the system. The portal is currently being used by more than 550 registered users from the EHDEN partners, and will soon be openly available to the whole community. The tools integrated into this portal provide an ecosystem capable of supporting the different stages of a medical study.

EMIF Catalogue

One of the outcomes of the EMIF project was the EMIF Catalogue, an online platform designed to expose the characteristics of the databases involved in the project. In this platform, the databases are the main entity to be characterized and presented as

²⁰<https://portal.ehden.eu/>

possible data sources for conducting studies. The platform supports the concept of community, which is used to support distinct projects. When the EMIF project came to an end, so did support for this platform. Since EMIF Catalogue was built on top of MONTRA, we were able to migrate the outdated version to a new EMIF Catalogue²¹, supported by MONTRA 2.

The database catalogue was extended with interoperability measures according to the FAIR Data Principles. New components were developed to enrich the fingerprint template, as well as more extensive access profiles, for end-users. Moreover, this extension in the software platform addressed other aspects such as usability, semantic data annotation, and data retrieval. The management of the data catalogue framework was simplified to facilitate the creation of new communities. Currently, in the EMIF Catalogue, there are 11 active communities, with more than 500 users registered. Each community is focused on a health domain.

MSDA Portal

One of the EMIF Catalogue communities started to have similar needs compared with the EHDEN project, namely the isolation of information in a distinct instance of EMIF Catalogue. Based on this need, this community was extracted from its original host, and deployed in a self-contained portal, similar to the one created for EHDEN.

The MSDA Portal²² is currently available to support a multi-stakeholder collaboration that is working to accelerate research insights for innovative care and treatments for people with multiple sclerosis. The Portal enables the discovery of cohorts and datasets related to this disease.

As an instance of MONTRA 2, it was customised to incorporate the needs of this community, namely by defining the sets of RBAC policies, plugins accessible and the fingerprint template. One of the features that distinguish this portal from the two previously described is the capability of incorporating multiple fingerprint templates, *i.e.* the database catalogue, which incorporates multi-catalogue features. This can be seen as a catalogue of catalogues within the portal.

²¹<https://emif-catalogue.eu>

²²<https://msda.emif-catalogue.eu>

5.6.2 Study Manager, a plugin for study orchestration

The Study Manager system was developed with the aim of improving the process of coordinating a multicentre studies. This tool was integrated into EMIF Catalogue and a beta version of EHDEN Portal.

EMIF Catalogue integration

The EMIF-EHR community intends to explore the abundance of data available in European EHR systems. The community was initially prepared to leverage data on around 40 million of European adults and children by integrating healthcare databases from different countries. In the community, there are characteristics to represent the different types of existing data sources, such as population-based registries, hospital-based databases, cohorts, national registries, among others. Another community in the system is the EMIF-AD community. The community is focused on exposing datasets of patients that suffer from Alzheimer's disease. One of the goals was to set up a large data repository of patient data to allow biomarker discovery studies within the EMIF project.

We used this portal, more precisely these two communities, to integrate the Study Manager developed in the context of this work. We limited the integration of this tool to communities that already deployed features to support study orchestration. More precisely, a plugin similar to the Study Manager to support the methodology proposed by Fajarda *et al.* [86]. However, as we described in Section 2.1.3, this methodology required three entities to conduct a study, including extra manual steps that delayed the study's progress. Besides, its development strategy does not enable easy extensibility of the system as we proposed in this work.

The strategy adopted for data sharing was kept since this strategy was already defined within the consortium during the project. The EMIF-EHR community used a strategy focused on Private Remote Research Environment (PRRE), which is a remote environment, with controlled access that intends to protect the data from being taken out of the environment, while providing analytical tools to work with the data. The data aggregation used was the same as proposed by Fajarda *et al.* [86], which is based on loading the data from all institutions into a database dedicated to the study. Thus, for each study, a new empty database is created. EMIF-AD adopted a different solution, by using ownCloud for data sharing, and a protected instance of tranSMART²³ for

²³<https://i2b2transmart.org>

data analysis. The data was aggregated on tranSMART and subsequently, an approach to harmonise the data was proposed by Almeida *et al.* [11].

EHDEN Portal integration

Since within the EHDEN project it was initially defined the use of the ARACHNE (described in Section 2.1.3) for orchestrating the distributed studies, we were not able to officially use the developed Study Manager tool in the production environment. However, it was integrated as a demonstration tool, with the potential to be included in the project roadmap as an auxiliary tool. For not being an official tool in the project, neither the strategy for sharing the data nor the methodology for aggregating the query results was established.

5.7 Discussion

The proposed systems create new opportunities for discovering and sharing biomedical information. MONTRA 2 represents part of the work described in this chapter, although we complemented this framework with additional features, namely recommending health databases and orchestrating studies.

5.7.1 Evolution from the first version of MONTRA

The first version of the MONTRA Framework was created in the context of the EMIF project aiming to support biomedical data sharing. However, the EHDEN project had other requirements, which were incorporated into the framework proposed in this work. Almost four years of development separate these two versions, in which we evolved this application based on the feedback obtained from the data owners, researchers, and work-package colleagues.

In its first version, MONTRA had the potential to simplify the creation of web-based catalogues, independently of the data scenario. This version included the concept of community, where the databases were segregated based on their community. This framework was the core of the EMIF-Catalogue, previously described as one of the research applications for this new version. MONTRA 2 kept the same principles but expanded to another level, namely by including semantic features in the catalogue definition. This enabled the definition of semantic catalogues, which simplifies the federation between distinct database catalogues.

MONTRA 2 was enhanced to also support the creation of an environment to integrate distinct tools in a centralised platform. The goal of this paradigm was to provide the researchers with a workplace with all required tools to: i) compare and identify the databases of interest for clinical studies; ii) streamline a study over the network; and iii) retrieve the results and aggregate them. All of these tools are protected under a federated SSO mechanism with profile verification. In medical research projects, this level of protection is desired and MONTRA 2 is fully compliant with such mechanisms.

This proposed version was restructured in its core, enabling the possibility of having multiple catalogues by community. An example of a use case of this feature was in the MSDA Catalogue, which required distinct catalogues in the scope of this community. One is for collecting descriptive information and metadata collected in Multiple Sclerosis initiatives. Another for Multiple Sclerosis-specific databases in which COVID-19 data was collected. In addition to these two, others have been planned in this community, and other communities in the EMIF Catalogue are also evolving in this course.

With these new developments, the access control mechanisms were improved aiming to provide RBAC control policies to all users on the platform. However, we also incorporated access control lists to ensure the community manager can define different levels of access to different roles and actions in the system.

5.7.2 Compliance with FAIR principles

The Findability, Accessibility, Interoperability, and Reusability (FAIR) principles are subdivided into 13 items [198]. When MONTRA 2 was created, we evaluated these items and developed the system to be fully compliant with these principles. However, MONTRA 2 instances can be customised by community managers, and some of the principles may depend on specific configurations.

In the EHDEN Portal, the framework is implemented to make the data FAIR at an unprecedented level:

- Findable: Data in the project is findable by publishing meta-data about each data source, which is complemented with profiles generated directly from the OMOP CDM databases.

- **Accessible:** With the Study Manager, data owners have full control of their data, including when sharing study results in the EHDEN network. Researchers can establish their requests, which are then received by the data owners in this web tool. They have full access to the package that follows the request, being capable of reviewing the analytical code. In short, data owners can accept or decline the request, run the study, review the results, and approve data sharing. All these steps can be done in an efficient and transparent workflow. This strategy allows controlled and granular data access.
- **Interoperable:** The OMOP CDM and standard vocabularies adopted by OHDSI enable a high-level of interoperability.
- **Reusable:** The history kept in the system allows the creation of mechanisms to ensure the re-use of all study components, *e.g.* cohort definitions, study specifications, and analytical code, among others.

According to the FAIR principles, in the EHDEN project, it was established and collected a minimal set of metadata describing source data provenance. The metadata includes: i) the primary source of the data; ii) when the data was collected; iii) the purpose of data collection; iv) source coding systems; and v) other directly obtained using the CatalogueExport package. Every data source has assigned a globally unique identifier and the metadata is made machine-readable and openly accessible due to the semantic features incorporated in the catalogue.

5.7.3 Recommender system's impact

Recommender systems have been successfully implemented in several scenarios related to online business, in order to stimulate increased profits [199]. Recently, these systems have been also applied in healthcare services aiming at the optimisation of decision support systems to make recommendations and suggestions for preventive interventions [200]. Although the proposed scenario is in the medical field, with this work we tried to aid clinical researchers in their findings.

The proposed system was integrated into the EMIF Catalogue platform and validated in the EMIF-AD community. The 62 cohorts are available to be studied in the platform composed of more than 141 000 patients. However, not all these cohorts can be used in across-trials studies.

Figure 5.13 shows an architectural overview of how this system was integrated into the EMIF Catalogue platform. The metadata is inserted manually into the catalogue, but the statistical data to perform the calculation is automatically extracted from the recommender engine. There, it applies and combines collaborative filtering with context-based retrieval techniques, replying to the predicted cohort recommendations. Additionally, we also present all the publication related to the recommended cohort that is indexed in the EMIF Catalogue system, as well as in the PubMed repository. Despite the chosen use case, the system was designed to work in other biomedical databases.

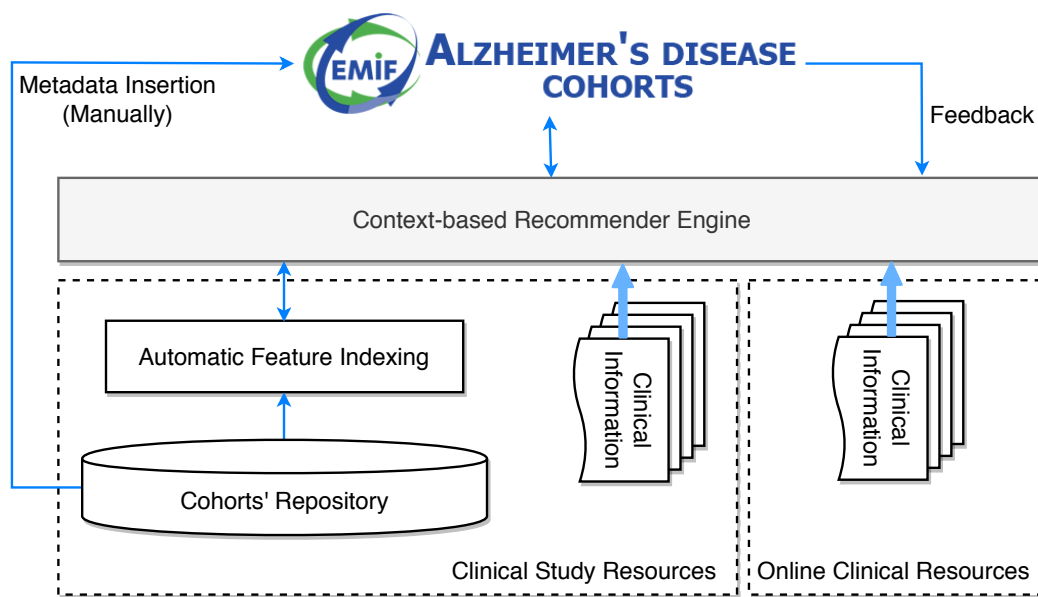


Fig. 5.13.: Architectural overview of the clinical recommender system. The system gets all the information from the EMIF Catalogue and processes it in the recommender engine. Then, it creates predicted cohort recommendations by applying collaborative filtering with context-based retrieval techniques.

5.7.4 Streamlining and orchestrating studies

Conducting multicentre clinical studies typically implies handling several socio-technical issues, from data access policies to its analyses. Coordinating such studies, involves data selection, negotiation, extraction and analyses, which is a complex task. This task cannot be handled using exchanging messages systems, like email, forums or others. Although there are some task-oriented systems for orchestrating processes, we proposed a methodology and a tool that simplifies the process of coordinating a multicentre study. The presentation of this methodology was health-oriented, however, the technical solutions behind this methodology can be applied in to other use cases.

The advantage of this methodology in the health domain is its impact on society. This is a fact since there is a great deal of interest from the medical community to aggregate information from distinct data sources. This data aggregation helps medical researchers to have a bigger dataset to analyse, which can show patterns that are not visible in smaller datasets; or to avoid wrong conclusions from insignificant patterns that are only present in smaller datasets. Therefore, aggregating data is already well-established as a practice, in almost all cases, for providing more accurate insights. By applying this concept in the health domain, we were able to create a solution that can be compared with the current state-of-the-art of methodologies for multicentre data analyses. The solution described in this chapter can optimise the work proposed by Fajarda *et al.* [86], by removing manual steps and extra entities in the pipeline, such as the query manager.

The described tool also solves an intermediate problem when querying distributed and private databases, *i.e.*, the query management is currently handled by the Study Manager. The next step can be focused on ensuring data anonymisation of the retrieved data [19].

5.8 Final considerations

This work was conducted in the context of the EHDEN project to characterise the databases in the project. Although the resulting work also was adopted in other health contexts, we kept the focus on the EHDEN needs. The main goal of the EHDEN consortium was to provide services that enable a federated European data network to perform fast, scalable, and highly reproducible health research while respecting privacy regulations, local data provenance and governance. The proposed portal strongly benefits from the OHDSI tools and principles, being a gateway to showcase the collaboration between the EHDEN consortium and OHDSI.

The main motivation for this work was to facilitate the setup of web data catalogues for distinct applications. MONTRA 2 is based on dynamic skeletons which allow describing any kind of data and is automatically used to create the data stored and to build the web user interface, without requiring coding skills. This framework was used and validated in several applications, such as the EHDEN Portal, EMIF Catalogue and MSDA Portal, to allow the presentation, discovery and sharing of biomedical data sources.

Complementary to the catalogue available in EHDEN Portal, it was created the EHDEN Network Dashboards (that we denominated as Network Dashboards in this document). It was designed to make use of a more restricted version of the standard ACHILLES results files, further encouraging its adoption for all data owners in the OHDSI network. This system can help researchers to perform qualitative data analysis on the available databases and to boost the selection process of more appropriate resources to perform a research study.

In summary, conducting multicentre medical studies is currently a reality. Health and life science researchers have identified several opportunities in sharing data. These opportunities can only be achieved if researchers can share data between them. The strategy proposed in this work empowers them with bigger data sets for each study, which increases the impact of their findings. However, different governmental issues were raised with this idea. Therefore, the proposed strategies aim to facilitate the exploration of patient-level databases, while minimising the risk of violating the patient's privacy.

Conclusions

“Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning.”¹

This chapter summarizes the work presented in this document, providing an overview of what was done during this doctorate. I found this quote capable of describing the feeling after finishing the writing of this document, that we are only at the beginning of a lifetime journey. All the solutions proposed in this document were able to solve a specific problem, but on the other hand, they raised more challenges. Therefore, this final chapter presents a brief analysis of how the initial research questions were answered, as well as, it presents some future work and research directions.

Enriching information extraction pipelines in clinical decision support systems is a research topic that can be addressed from different points of view. In this work, we tried to enrich these pipelines starting by working on the foundations of clinical decision support systems. We recognised that to increase the quality of the treatments, researchers need to study the impact of new drugs, or the efficiency of current treatments. These findings can originate new treatment protocols that can be integrated into the decision-support systems of healthcare institutions. Therefore, in this work, we focused on creating methodologies and tools to help medical researchers conduct more impactful findings, to improve the source of these systems.

We started by specifying the scope of this work, based on the biomedical data formats that we could use. Motivated by EHDEN project, we focused this work on EHR relational data, that we tried to supplement with data extracted from medical narratives. Then, in the later stage, after defining strategies to have an interoperable network of data sources, we proposed solutions to support research using these data sources. In short, we presented several software solutions to integrate biomedical data, and the final product is a platform that facilitates the exploration of this information across databases.

¹Winston Churchill, Lord Mayor’s Luncheon, Mansion House following the victory at El Alamein North Africa, London, 10 November 1942.

6.1 Outcomes overview

In the process of creating methodologies to integrate and share biomedical data sources across Europe, we achieved several results.

The first hypothesis addressed the lack of interoperability between health databases. However, as we found during this work, the problem was not the lack of standard solutions to interconnect these databases. Instead, the problem was the effort required to adopt one of these standards. To answer this problem, we proposed solutions to simplify the migration of EHR data to one of the standard data schemas currently used in medical studies. We validated these solutions using heterogeneous cohorts of patients' data suffering from Alzheimer's disease. The interoperability was ensured by converting data sources to the OMOP CDM schema.

The second hypothesis was about enriching the information stored in the databases, using unstructured data present in clinical narratives. For this, we proposed a solution capable of extracting medical concepts and storing them in an OMOP CDM database. Part of this solution is supported by the work done to answer the first hypothesis. We validated the proposed NLP strategies using scientific challenges, namely organised by n2c2 organisation.

Finally, the third hypothesis was focused on finding the most adequate health databases for specific research studies. To answer this question, we have collaborated during this doctoral program with the EHDEN partners aiming to propose and adjust a solution based on real needs. The result was a flexible framework capable of being extended to support complementary tools. This work was validated in the context of the EHDEN project. Additionally, it also replaced old technologies that have supported the EMIF project in the past. This tool has been validated with thousands of users, with a huge impact on real-life environments.

6.2 Future work and limitations

The methodology proposed in Chapter 3 was developed to generate OMOP CDM databases using cohort raw data. However, changing the output data schema to be completely different from the OMOP CDM may require a restructuring of the loading stage of the proposed solution. Small adjustments in this structure are possible with

minor effects on the developed system. When we developed the workflow, we kept in mind possible adjustments in the OMOP CDM, because OHDSI is an active community that has improved the OMOP CDM aiming to expand to other medical domains.

This methodology was implemented and validated using Alzheimer's disease cohorts. We do not consider this methodology limited to this domain. However, applying this migration workflow using cohorts from other diseases may require adjustments, namely in defining an ontology for this new domain. The methodology is focused on the ETL procedure, which includes dataset harmonisation at different levels, and it adopts well-established tools designed to perform EHR observational studies in cohort datasets. Although these cohorts are disease-specific, the aggregation of results from different institutions has revealed impactful findings [119, 120].

In the work presented in Chapter 4, several important concepts currently being studied in the health informatics field were used. Although we were able to create a methodology respecting the standard principles and using tools already validated in other scenarios, it was possible to identify some future directions and necessities in this subject.

Neji currently supports machine learning modules and vocabularies, but it does not support deep-learning models. This feature could be a very useful asset in specific scenarios, namely if a richly annotated dataset with similar characteristics to the target data was available. In this case, it would be possible to train a model and possibly obtain better results in the information extraction stage. Another interesting feature would be the integration of the post-processing features directly in Neji. This way, the public annotating service would be able to provide the final annotations without the need for posterior processing steps like in our approach. The modular architecture of Neji facilitates such adaptations and extensions.

We also identified the need for a tool that aggregates Neji and Usagi features in a single solution. We were not able to find a tool with such features. However, we believe that merging these features in a unique and collaborative tool could simplify concept extraction and mappings. It could also provide different support to the annotators during the mapping stage of EHR databases migration pipelines. This is mainly because by having these features merged, it would be easier to add customizable vocabularies for specific institutions without affecting the standard vocabularies provided by Athena.

Chapter 5 offers a solution to a problem that is focused on the difficulties in discovering and sharing biomedical data sources. However, some components could be improved in further research. The strategy used for data sharing was briefly analysed and we did not invest too much work in this topic. However, for this end, a tool could be implemented to anonymise the data in the institution using the described algorithms, instead of relying on the encryption algorithms provided by the data-sharing platforms.

Finally, the task that could leverage this work to another level was the automatisation of the query pipeline by removing the interactions of the data owner. This limitation can be technically solved, but with a difficult adoption by the data owners, depending on the data domain. This topic by itself can lead to a research direction focused on data anonymisation.

6.3 Research directions

In the previous section, we presented the limitations of the proposed solutions, as well as, some futures directions. However, the analysis of theses limitations allowed us to identify some challenges and research directions. Herein, we present and discuss some of the possible research lines for future work:

- **Standardising a fingerprinting schema:** A lot of efforts have been conducted to ensure interoperability between data sources, as well as to publish their metadata to facilitate discovery. This resulted in several health database catalogues that cannot communicate and exchange information between them. There are already some initiatives to create federated catalogues in specific domains, however, this is only the beginning. Standard schemas and ontologies to federated this communication is a possible research direction to optimise the creation of health database catalogues.
- **Data exchange:** The strategy adopted for exchanging the released datasets is secure and ensures that the data comes from the correct sources and is addressed to the right data viewer. However, there are more secure strategies that prevent non-repudiation when publishing discoveries using the requested data. With the increase of blockchain technologies and the introduction of Non-Fungible Token (NFT) technologies to ensure data immutability, this work can evolve to the adoption of similar technologies. Adopting NFT technologies

for data sharing would ensure that the data owner shared a dataset that can be referenced in further scientific publications originating from the same data source.

- Automatic definition of ETL workflows: Automatically establishing the mappings between the original data schema to the target is an open research direction that can be applied beyond the health domain. This can be simplified and focused on the medical domain, by using OMOP CDM as the target data schema. In this work, we proposed semi-automatic methodologies, but this proposal can be optimised at different levels.
- Extending OMOP CDM to incorporate other data types: Over the years some initiatives tried to extend the OMOP CDM to incorporate more information. The adoption of these initiatives at a large scale fails due to several issues (ensuring data privacy in complex data formats, breaking the schema interoperability, and raising issues when sharing results, among others). Investing in this direction may leverage medical research to new levels, namely by allowing distributed studies using DICOM images, or genomic data.
- Secure FAIR data: The ultimate goal of FAIR principles is to optimise the reuse of data. The principles emphasise machine-actionability, *i.e.*, the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention [198]. However, we identified a research line in this topic, by combining it with security, *i.e.* applying the FAIR principles following secure guidelines to ensure safe machine-to-machine communication.

Considering the increasing impact of technology in healthcare, along with the rapid developments in this field, we firmly believe in the importance of the presented research topics.

References

- [1] Priya Ranganathan y Rakesh Aggarwal. «Study designs: Part 1–An overview and classification». En: *Perspectives in clinical research* 9.4 (2018), pág. 184. DOI: 10.4103/picr.PICR_124_18 (vid. págs. 1, 171, 185).
- [2] Jae W Song y Kevin C Chung. «Observational studies: cohort and case-control studies». En: *Plastic and reconstructive surgery* 126.6 (2010), pág. 2234. DOI: 10.1097/PRS.0b013e3181f44abc (vid. págs. 1, 171, 185).
- [3] Melissa DA Carlson y R Sean Morrison. «Study design, precision, and validity in observational studies». En: *Journal of palliative medicine* 12.1 (2009), págs. 77-82. DOI: 10.1089/jpm.2008.9690 (vid. págs. 1, 171, 185).
- [4] George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek y col. «Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers». En: *Studies in health technology and informatics* 216 (2015), pág. 574. DOI: 10.3233/978-1-61499-564-7-574 (vid. págs. 1-3, 11, 15, 17, 19, 27, 107, 171, 172, 185, 186).
- [5] Paul A Harris, Robert Taylor, Robert Thielke, Jonathon Payne, Nathaniel Gonzalez y Jose G Conde. «Research electronic data capture (REDCap) a metadata-driven methodology and workflow process for providing translational research informatics support». En: *Journal of biomedical informatics* 42.2 (2009), págs. 377-381. DOI: 10.1016/j.jbi.2008.08.010 (vid. págs. 1, 171, 185).
- [6] C Hendricks Brown, Zili Sloboda, Fabrizio Faggiano, Brent Teasdale, Ferdinand Keller, Gregor Burkhart, Federica Vigna-Taglianti, George Howe, Katherine Masyn, Wei Wang y col. «Methods for synthesizing findings on moderation effects across multiple randomized trials». En: *Prevention science* 14.2 (2013), págs. 144-156. DOI: 10.1007/s11121-011-0207-8 (vid. págs. 2, 15, 172, 186).
- [7] João Rafael Almeida, Luís Bastão Silva, Isabelle Bos, Pieter Jelle Visser y José Luís Oliveira. «A methodology for cohort harmonisation in multicentre clinical research». En: *Informatics in Medicine Unlocked* Volume 27 (2021), pág. 100760. DOI: 10.1016/j.imu.2021.100760 (vid. págs. 2-4, 6, 17, 27, 172, 186).
- [8] Reid Cushman, A Michael Froomkin, Anita Cava, Patricia Abril y Kenneth W Goodman. «Ethical, legal and social issues for personal health records and applications». En: *Journal of biomedical informatics* 43.5 (2010), S51-S55. DOI: 10.1016/j.jbi.2010.05.003 (vid. págs. 2, 172, 186).

- [9] Grace Fox. «"To protect my health or to protect my health privacy?": A mixed-methods investigation of the privacy paradox». En: *Journal of the Association for Information Science and Technology* 71.9 (2020), págs. 1015-1029. DOI: 10.1002/asi.24369 (vid. págs. 2, 172, 186).
- [10] Stephane M Meystre, Christian Lovis, Thomas Bürkle, Gabriella Tognola, Andrius Budrionis y Christoph U Lehmann. «Clinical data reuse or secondary use: current status and potential future progress». En: *Yearbook of medical informatics* 26.01 (2017), págs. 38-52. DOI: 10.15265/IY-2017-007 (vid. págs. 2, 172, 186).
- [11] João Rafael Almeida, Luís Bastião Silva, Alejandro Pazos y José Luís Oliveira. «Combining heterogeneous patient-level data into transSMART to support multicentre studies». En: *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*. 2022, págs. 62-65. DOI: 10.1109/CBMS55023.2022.00018 (vid. págs. 4, 6, 141).
- [12] João Rafael Almeida, Leonardo Coelho y José L. Oliveira. «BIcenter: A collaborative Web ETL solution based on a reflective software approach». En: *SoftwareX* 16 (2021), pág. 100892. ISSN: 2352-7110. DOI: 10.1016/j.softx.2021.100892 (vid. págs. 4, 56).
- [13] João Rafael Almeida, Alejandro Pazos y José Luís Oliveira. «BIcenter-AD: Harmonising Alzheimer's Disease Cohorts using a Common ETL Tool». En: *Informatics in Medicine Unlocked* 35 (2022), pág. 101133. ISSN: 2352-9148. DOI: 10.1016/j.imu.2022.101133 (vid. pág. 4).
- [14] João Rafael Almeida, João Figueira Silva, Sérgio Matos y José Luís Oliveira. «A two-stage workflow to extract and harmonize drug mentions from clinical notes into observational databases». En: *Journal of Biomedical Informatics* 120 (2021), pág. 103849. DOI: 10.1016/j.jbi.2021.103849 (vid. pág. 4).
- [15] João Rafael Almeida y José Luís Oliveira. «Multi-language Concept Normalisation of Clinical Cohorts». En: *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE. 2020, págs. 261-264. DOI: 10.1109/CBMS49503.2020.00056 (vid. págs. 4, 57).
- [16] João Rafael Almeida y Sérgio Matos. «Rule-based extraction of family history information from clinical notes». En: *Proceedings of the 35th Annual ACM Symposium on Applied Computing*. 2020, págs. 670-675. DOI: 10.1145/3341105.3374000 (vid. págs. 4, 73).
- [17] João Figueira Silva, João Rafael Almeida y Sérgio Matos. «Extraction of family history information from clinical notes: deep learning and heuristics approach». En: *JMIR medical informatics* 8.12 (2020), e22898. DOI: 10.2196/22898 (vid. pág. 4).
- [18] João Rafael Almeida, Rosa Gini, Giuseppe Roberto, Peter Rijnbeek y José Luís Oliveira. «TASKA: a modular task management system to support health research studies». En: *BMC medical informatics and decision making* 19.1 (2019), págs. 1-9. DOI: 10.1186/s12911-019-0844-6 (vid. págs. 4, 25, 113).
- [19] João Rafael Almeida, Joao Paulo Barraca y José Luís Oliveira. «A secure architecture for exploring patient-level databases from distributed institutions». En: *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE. 2022, págs. 447-452. DOI: 10.1109/CBMS55023.2022.00086 (vid. págs. 4, 145).

- [20] Umit Topaloglu y Matvey B Topaloglu. «Using a federated network of real-world data to optimize clinical trials operations». En: *JCO clinical cancer informatics* 2 (2018), págs. 1-10. DOI: 10.1200/CCI.17.00067 (vid. págs. 10, 26).
- [21] David C Kaelber, Ashish K Jha, Douglas Johnston, Blackford Middleton y David W Bates. «A research agenda for Personal Health Records (PHRs)». En: *Journal of the American Medical Informatics Association* 15.6 (2008), págs. 729-736. DOI: 10.1197/jamia.M2547 (vid. págs. 10).
- [22] Michael G Kahn, Tiffany J Callahan, Juliana Barnard, Alan E Bauck, Jeff Brown, Bruce N Davidson, Hossein Estiri, Carsten Goerg, Erin Holve, Steven G Johnson y col. «A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data». En: *Egms* 4.1 (2016). DOI: 10.13063/2327-9214.1244 (vid. págs. 10).
- [23] Nicole G Weiskopf, George Hripcsak, Sushmita Swaminathan y Chunhua Weng. «Defining and measuring completeness of electronic health records for secondary use». En: *Journal of biomedical informatics* 46.5 (2013), págs. 830-836. DOI: 10.1016/j.jbi.2013.06.010 (vid. págs. 10).
- [24] MK Ross, Wei Wei y L Ohno-Machado. «"Big data" and the electronic health record». En: *Yearbook of medical informatics* 23.01 (2014), págs. 97-104. DOI: 10.15265/IY-2014-0003 (vid. págs. 10, 11).
- [25] Aya Gamal, Sherif Barakat y Amira Rezk. «Standardized electronic health record data modeling and persistence: A comparative review». En: *Journal of biomedical informatics* 114 (2021), págs. 103670. DOI: 10.1016/j.jbi.2020.103670 (vid. págs. 11).
- [26] Pilar Muñoz, Jesús D Trigo, Ignacio Martínez, Adolfo Muñoz, Javier Escayola y José García. «The ISO/EN 13606 standard for the interoperable exchange of electronic health records». En: *Journal of Healthcare Engineering* 2.1 (2011), págs. 1-24. DOI: 10.1260/2040-2295.2.1.1 (vid. págs. 11).
- [27] Gro-Hilde Ulriksen, Rune Pedersen y Gunnar Ellingsen. «Infrastructuring in health-care through the openEHR architecture». En: *Computer Supported Cooperative Work (CSCW)* 26.1 (2017), págs. 33-69. DOI: 10.1007/s10606-017-9269-x (vid. págs. 11).
- [28] Hripcsak G, Ryan P, Madigan D, Kostka K, Schuemie M, DeFalco F, et al. *The Book of OHDSI: Observational Health Data Sciences and Informatics*. OHDSI, 2019 (vid. págs. 11, 21-23, 40, 42, 64).
- [29] Joel JPC Rodrigues. *Health information systems: concepts, methodologies, tools, and applications: concepts, methodologies, tools, and applications*. Vol. 1. Igi Global, 2009 (vid. págs. 11).
- [30] Lorraine M Fernandes, Michele O'Connor y Victoria Weaver. «Big data, bigger outcomes». En: *Journal of AHIMA* 83.10 (2012), págs. 38-43 (vid. págs. 11).
- [31] Arshia Rehman, Saeeda Naz e Imran Razzak. «Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities». En: *Multimedia Systems* (2021), págs. 1-33. DOI: 10.1007/s00530-020-00736-8 (vid. págs. 11).
- [32] Travis B Murdoch y Allan S Detsky. «The inevitable application of big data to health care». En: *Jama* 309.13 (2013), págs. 1351-1352. DOI: 10.1001/jama.2013.393 (vid. págs. 11).

- [33] Liya Abraham, George C Vilanilam y col. «Big data in clinical sciences-value, impact, and fallacies». En: *Archives of Medicine and Health Sciences* 10.1 (2022), pág. 112. DOI: 10.4103/amhs.amhs_296_21 (vid. pág. 11).
- [34] Peter B Jensen, Lars J Jensen y Søren Brunak. «Mining electronic health records: towards better research applications and clinical care». En: *Nature Reviews Genetics* 13.6 (2012), págs. 395-405. DOI: 10.1038/nrg3208 (vid. pág. 11).
- [35] Jie Xu, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian y Fei Wang. «Federated learning for healthcare informatics». En: *Journal of Healthcare Informatics Research* 5.1 (2021), págs. 1-19. DOI: 10.1007/s41666-020-00082-4 (vid. pág. 11).
- [36] Benjamin CM Fung, Ke Wang, Rui Chen y Philip S Yu. «Privacy-preserving data publishing: A survey of recent developments». En: *ACM Computing Surveys (Csur)* 42.4 (2010), págs. 1-53. DOI: 10.1145/1749603.1749605 (vid. pág. 11).
- [37] Stéphane M Meystre, Guergana K Savova, Karin C Kipper-Schuler y John F Hurdle. «Extracting information from textual documents in the electronic health record: a review of recent research». En: *Yearbook of medical informatics* 17.01 (2008), págs. 128-144. DOI: 10.1055/s-0038-1638592 (vid. pág. 12).
- [38] Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn y col. «Clinical information extraction applications: a literature review». En: *Journal of biomedical informatics* 77 (2018), págs. 34-49. DOI: 10.1016/j.jbi.2017.11.011 (vid. pág. 12).
- [39] Elizabeth Ford, John A Carroll, Helen E Smith, Donia Scott y Jackie A Cassell. «Extracting information from the text of electronic medical records to improve case detection: a systematic review». En: *Journal of the American Medical Informatics Association* 23.5 (2016), págs. 1007-1015. DOI: 10.1093/jamia/ocv180 (vid. págs. 12, 21).
- [40] Seyedmostafa Sheikhalishahi, Riccardo Miotto, Joel T Dudley, Alberto Lavelli, Fabio Rinaldi, Venet Osmani y col. «Natural language processing of clinical notes on chronic diseases: systematic review». En: *JMIR medical informatics* 7.2 (2019), e12239. DOI: 10.2196/12239 (vid. págs. 12, 30, 70, 72).
- [41] Rimma Pivovarov y Noémie Elhadad. «Automated methods for the summarization of electronic health records». En: *Journal of the American Medical Informatics Association* 22.5 (2015), págs. 938-947. DOI: 10.1093/jamia/ocv032 (vid. págs. 12, 70).
- [42] Amy Neustein, S Sagar Imambi, Mário Rodrigues, António Teixeira y Liliana Ferreira. «Application of text mining to biomedical knowledge extraction: analyzing clinical narratives and medical literature». En: *Text Mining of Web-based Medical Content* (2014), págs. 3-32. DOI: 10.1515/9781614513902 (vid. pág. 12).
- [43] Tiago Marques Godinho, Rui Lebre, João Rafael Almeida y Carlos Costa. «ETL framework for real-time business intelligence over medical imaging repositories». En: *Journal of digital imaging* 32.5 (2019), págs. 870-879. DOI: 10.1007/s10278-019-00184-5 (vid. pág. 13).
- [44] Hai K Huang. *PACS and imaging informatics: basic principles and applications*. John Wiley & Sons, 2011. DOI: 10.2345/i0899-8205-40-2-125.1 (vid. pág. 13).

- [45] João Rafael Almeida, Eriksson Monteiro y José Luís Oliveira. «An architecture to define cohorts over medical imaging datasets». En: *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE. 2021, págs. 545-549. DOI: 10.1109/CBMS52027.2021.00088 (vid. pág. 14).
- [46] Elaine R Mardis. «DNA sequencing technologies: 2006–2016». En: *Nature Protocols* 12.2 (2017), pág. 213. DOI: 10.1038/nprot.2016.182 (vid. pág. 14).
- [47] Josephine Johnston, John D Lantos, Aaron Goldenberg, Flavia Chen, Erik Parens, Barbara A Koenig, NSIGHT Ethics y Policy Advisory Board. «Sequencing newborns: a call for nuanced use of genomic technologies». En: *Hastings Center Report* 48 (2018), S2-S6. DOI: 10.1002/hast.874 (vid. pág. 14).
- [48] Jane Kaye. «The tension between data sharing and the protection of privacy in genomics research». En: *Annual review of genomics and human genetics* 13 (2012), págs. 415-431. DOI: 10.1146/annurev-genom-082410-101454 (vid. pág. 14).
- [49] Jan-Eric Litton. «Launch of an infrastructure for health research: BBMRI-ERIC». En: *Biopreservation and biobanking* 16.3 (2018), págs. 233-241. DOI: 10.1089/bio.2018.0027 (vid. pág. 14).
- [50] Angen Liu y Kai Pollard. «Biobanking for personalized medicine». En: *Biobanking in the 21st Century*. Springer, 2015, págs. 55-68. DOI: 10.1007/978-3-319-20579-3_5 (vid. pág. 14).
- [51] Antonio Amorim, Filipe Pereira, Cíntia Alves y Oscar García. «Species assignment in forensics and the challenge of hybrids». En: *Forensic Science International: Genetics* 48 (2020), pág. 102333. DOI: 10.1016/j.fsigen.2020.102333 (vid. pág. 14).
- [52] Holger Langhof, Hannes Kahrass, Sören Sievers y Daniel Strech. «Access policies in biobank research: what criteria do they include and how publicly available are they? A cross-sectional study». En: *European Journal of Human Genetics* 25.3 (2017), págs. 293-300. DOI: 10.1038/ejhg.2016.172 (vid. pág. 14).
- [53] Jennifer Kulynych y Henry T Greely. «Clinical genomics, big data, and electronic medical records: reconciling patient rights with research when privacy and science collide». En: *Journal of Law and the Biosciences* 4.1 (2017), págs. 94-132. DOI: 10.1093/jlb/lsw061 (vid. pág. 14).
- [54] Paul J McLaren, Jean Louis Raisaro, Manel Aouri, Margalida Rotger, Erman Ayday, István Bartha, Maria B Delgado, Yannick Vallet, Huldrych F Günthard, Matthias Cavassini y col. «Privacy-preserving genomic testing in the clinic: a model using HIV treatment». En: *Genetics in medicine* 18.8 (2016), págs. 814-822. DOI: 10.1038/gim.2015.167 (vid. pág. 14).
- [55] Dennis Grishin, Kamal Obbad y George M Church. «Data privacy in the age of personal genomics». En: *Nature biotechnology* 37.10 (2019), págs. 1115-1117. DOI: 10.1038/s41587-019-0271-3 (vid. pág. 14).
- [56] Marco Masseroli, Abdulrahman Kaitoua, Pietro Pinoli y Stefano Ceri. «Modeling and interoperability of heterogeneous genomic big data for integrative processing and querying». En: *Methods* 111 (2016), págs. 3-11. DOI: 10.1016/j.ymeth.2016.09.002 (vid. pág. 14).

- [57] João Rafael Almeida, Diogo Pratas y José Luís Oliveira. «A semi-automatic methodology for analysing distributed and private biobanks». En: *Computers in Biology and Medicine* 130 (2021), pág. 104180. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2020.104180 (vid. págs. 14, 25).
- [58] Sowndarya Palanisamy y P SuvithaVani. «A survey on RDBMS and NoSQL Databases MySQL vs MongoDB». En: *2020 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE. 2020, págs. 1-7. DOI: 10.1109/ICCCI48352.2020.9104047 (vid. pág. 15).
- [59] Bogdan George Tudorica y Cristian Bucur. «A comparison between several NoSQL databases with comments and notes». En: *2011 RoEduNet international conference 10th edition: Networking in education and research*. IEEE. 2011, págs. 1-5. DOI: 10.1109/RoEduNet.2011.5993686 (vid. pág. 15).
- [60] Boyu Hou, Kai Qian, Lei Li, Yong Shi, Lixin Tao y Jigang Liu. «MongoDB NoSQL injection analysis and detection». En: *2016 IEEE 3rd International Conference on Cyber Security and Cloud Computing (CSCloud)*. IEEE. 2016, págs. 75-78. DOI: 10.1109/CSCloud.2016.57 (vid. pág. 15).
- [61] Kanika Goel y Arthur HM Ter Hofstede. «Privacy-Breaching Patterns in NoSQL Databases». En: *IEEE Access* 9 (2021), págs. 35229-35239. DOI: 10.1109/ACCESS.2021.3062034 (vid. pág. 15).
- [62] Tiago Marques Godinho, Rui Lebre, Luís Bastião Silva y Carlos Costa. «An efficient architecture to support digital pathology in standard medical imaging repositories». En: *Journal of biomedical informatics* 71 (2017), págs. 190-197. DOI: 10.1016/j.jbi.2017.06.009 (vid. pág. 16).
- [63] S Trent Rosenbloom, William W Stead, Joshua C Denny, Dario Giuse, Nancy M Lorenzi, Steven H Brown y Kevin B Johnson. «Generating clinical notes for electronic health record systems». En: *Applied clinical informatics* 1.03 (2010), págs. 232-243. DOI: 10.4338/ACI-2010-03-RA-0019 (vid. pág. 16).
- [64] Eriksson Monteiro, Carlos Costa, José Luís Oliveira, David Campos y Luís Bastião Silva. «Caching and prefetching images in a web-based DICOM viewer». En: *2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE. 2016, págs. 241-246. DOI: 10.1109/CBMS.2016.68 (vid. pág. 16).
- [65] João Rafael Almeida, Olga Fajarda, Arnaldo Pereira y José Luís Oliveira. «Strategies to Access Patient Clinical Data from Distributed Databases». En: *HEALTHINF*. SciTePress, 2019, págs. 466-473. DOI: 10.5220/0007576104660473 (vid. págs. 16, 17, 24, 27, 66).
- [66] Shawn N Murphy, Griffin Weber, Michael Mendis, Vivian Gainer, Henry C Chueh, Susanne Churchill e Isaac Kohane. «Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2)». En: *Journal of the American Medical Informatics Association* 17.2 (2010), págs. 124-130. DOI: 10.1136/jamia.2009.000893 (vid. pág. 16).
- [67] Andrew J McMurphy, Shawn N Murphy, Douglas MacFadden, Griffin Weber, William W Simons, John Orechia, Jonathan Bickel, Nich Wattanasin, Clint Gilbert, Philip Trevvett y col. «SHRINE: enabling nationally scalable multi-site disease studies». En: *PloS one* 8.3 (2013), e55811. DOI: 10.1371/journal.pone.0055811 (vid. pág. 16).

- [68] Christel Daniel, David Ouagne, Eric Sadou, Kerstin Forsberg, Mark Mc Gilchrist, Eric Zapletal, Nicolas Paris, Sajjad Hussain, Marie-Christine Jaulent y Dipka Kalra. «Cross border semantic interoperability for clinical research: the EHR4CR semantic resources and services». En: *AMIA Summits on Translational Science Proceedings 2016* (2016), pág. 51. DOI: 10.1002/lrh2.10014 (vid. pág. 16).
- [69] John Weeks y Roy Pardee. «Learning to share health care data: a brief timeline of influential common data models and distributed health data networks in US health care research». En: *eGEMs 7.1* (2019). DOI: 10.5334/egems.279 (vid. pág. 17).
- [70] Isabelle Bos, Stephanie Vos, Rik Vandenbergh, Philip Scheltens, Sebastiaan Engelborghs, Giovanni Frisoni, José Luis Molinuevo, Anders Wallin, Alberto Lleó, Julius Popp y col. «The EMIF-AD Multimodal Biomarker Discovery study: design, methods and cohort characteristics». En: *Alzheimer's research & therapy* 10.1 (2018), pág. 64. DOI: 10.1186/s13195-018-0396-5 (vid. págs. 17, 107, 179, 193).
- [71] Aris M. Ouksel y Amit Sheth. «Semantic interoperability in global information systems». En: *ACM Sigmod Record* 28.1 (1999), págs. 5-12. DOI: 10.1145/309844.309849 (vid. pág. 17).
- [72] Sebastian Garde, Petra Knaup, Evelyn JS Hovenga y Sam Heard. «Towards semantic interoperability for electronic health records». En: *Methods of information in medicine* 46.03 (2007), págs. 332-343. DOI: 10.1160/ME5001 (vid. pág. 17).
- [73] George Hripcsak, Patrick B. Ryan, Jon D. Duke, Nigam H. Shah, Rae Woong Park, Vojtech Huser, Marc A. Suchard, Martijn J. Schuemie, Frank J. DeFalco, Adler Perotte, Juan M. Banda, Christian G. Reich, Lisa M. Schilling, Michael E. Matheny, Daniella Meeker, Nicole Pratt y David Madigan. «Characterizing treatment pathways at scale using the OHDSI network». En: *Proceedings of the National Academy of Sciences* 113.27 (2016), págs. 7329-7336. DOI: 10.1073/pnas.1510502113 (vid. págs. 18, 27, 34, 66, 107, 174, 188).
- [74] Abdul Quamar, Jannik Straube y Yuanyuan Tian. «Enabling Rich Queries Over Heterogeneous Data From Diverse Sources In HealthCare». En: *CIDR*. 2020 (vid. pág. 18).
- [75] Behzad Golshan, Alon Halevy, George Mihaila y Wang-Chiew Tan. «Data integration: After the teenage years». En: *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI symposium on principles of database systems*. 2017, págs. 101-106. DOI: 10.1145/3034786.3056124 (vid. pág. 18).
- [76] Xiaolan Wang, Laura Haas y Alexandra Meliou. «Explaining data integration». En: *Data Engineering Bulletin* 41.2 (2018) (vid. pág. 18).
- [77] Laura M Haas, Eileen Tien Lin y Mary A Roth. «Data integration through database federation». En: *IBM Systems Journal* 41.4 (2002), págs. 578-596. DOI: 10.1147/sj.414.0578 (vid. pág. 18).
- [78] Michael Q Stearns, Colin Price, Kent A Spackman y Amy Y Wang. «SNOMED clinical terms: overview of the development process and project status.» En: *Proceedings of the AMIA Symposium*. American Medical Informatics Association. 2001, pág. 662 (vid. págs. 19, 70).
- [79] WHO. *World Health Organization: International classification of diseases, 11th Revision (ICD-11)*. 2018 (vid. pág. 19).

- [80] Carolyn E Lipscomb. «Medical subject headings (MeSH)». En: *Bulletin of the Medical Library Association* 88.3 (2000), pág. 265 (vid. pág. 19).
- [81] Olivier Bodenreider. «The unified medical language system (UMLS): integrating bio-medical terminology». En: *Nucleic acids research* 32.suppl_1 (2004), págs. D267-D270. DOI: 10.1093/nar/gkh061 (vid. pág. 19).
- [82] J Marc Overhage, Patrick B Ryan, Christian G Reich, Abraham G Hartzema y Paul E Stang. «Validation of a common data model for active safety surveillance research». En: *Journal of the American Medical Informatics Association* 19.1 (2011), págs. 54-60. DOI: 10.1136/amiajnl-2011-000376 (vid. pág. 21).
- [83] Rupa Makadia y Patrick B Ryan. «Transforming the Premier Perspective Hospital Database into the Observational Medical Outcomes Partnership (OMOP) Common Data Model». En: *Egems* 2.1 (2014). DOI: 0.13063/2327-9214.1110 (vid. pág. 21).
- [84] OHDSI. *The Book of OHDSI: Observational Health Data Sciences and Informatics*. OHDSI, 2019. ISBN: 9781088855195 (vid. págs. 22, 26).
- [85] Along Lin y R Brown. «The application of security policy to role-based access control and the common data security architecture». En: *Computer Communications* 23.17 (2000), págs. 1584-1593. DOI: 10.1016/S0140-3664(00)00244-9 (vid. pág. 24).
- [86] Olga Fajarda, Luís A Bastião Silva, Peter R Rijnbeek, Michel Van Speybroeck y José Luís Oliveira. «A Methodology to Perform Semi-automatic Distributed EHR Database Queries». En: *HEALTHINF*. SciTePress, 2018, págs. 127-134 (vid. págs. 26, 140, 145).
- [87] Abdul Majeed y Sungchang Lee. «Anonymization techniques for privacy preserving data publishing: A comprehensive survey». En: *IEEE Access* (2020). DOI: 10.1109/ACCESS.2020.3045700 (vid. pág. 26).
- [88] Romana Talat, Mohammad S Obaidat, Muhammad Muzammal, Ali Hassan Sodhro, Zongwei Luo y Sandeep Pirbhulal. «A decentralised approach to privacy preserving trajectory mining». En: *Future generation computer systems* 102 (2020), págs. 382-392. DOI: 10.1016/j.future.2019.07.068 (vid. pág. 27).
- [89] Chen Fang, Yuanbo Guo, Na Wang y Ankang Ju. «Highly efficient federated learning with strong privacy preservation in cloud computing». En: *Computers & Security* 96 (2020), pág. 101889. DOI: 10.1016/j.cose.2020.101889 (vid. pág. 27).
- [90] Jing Li, Xiaohui Kuang, Shujie Lin, Xu Ma y Yi Tang. «Privacy preservation for machine learning training and classification based on homomorphic encryption schemes». En: *Information Sciences* 526 (2020), págs. 166-179. DOI: 10.1016/j.ins.2020.03.041 (vid. pág. 27).
- [91] Suman Madan y Puneet Goswami. «A privacy preservation model for big data in map-reduced framework based on k-anonymisation and swarm-based algorithms». En: *International Journal of Intelligent Engineering Informatics* 8.1 (2020), págs. 38-53. DOI: 10.1016/j.ins.2020.03.041 (vid. pág. 27).
- [92] Chen-Yi Lin. «Suppression techniques for privacy-preserving trajectory data publishing». En: *Knowledge-Based Systems* 206 (2020), pág. 106354. DOI: 10.1016/j.knsys.2020.106354 (vid. pág. 27).

- [93] Ayong Ye, Qiang Zhang, Yiqing Diao, Jiaomei Zhang, Huina Deng y Baorong Cheng. «A semantic-based approach for privacy-preserving in trajectory publishing». En: *IEEE Access* 8 (2020), págs. 184965-184975. DOI: 10.1109/ACCESS.2020.3030038 (vid. pág. 27).
- [94] Nesrine Kaaniche, Maryline Laurent y Sana Belguith. «Privacy enhancing technologies for solving the privacy-personalization paradox: Taxonomy and survey». En: *Journal of Network and Computer Applications* 171 (2020), pág. 102807. DOI: 10.1016/j.jnca.2020.102807 (vid. pág. 27).
- [95] Maqbool Khan, Xiaotong Wu, Xiaolong Xu y Wanchun Dou. «Big data challenges and opportunities in the hype of Industry 4.0». En: *2017 IEEE International Conference on Communications (ICC)*. IEEE. 2017, págs. 1-6. DOI: 10.1109/ICC.2017.7996801 (vid. pág. 33).
- [96] Yue Zhuang, Fei Wu, Chun Chen y Yun Pan. «Challenges and opportunities: from big data to knowledge in AI 2.0». En: *Frontiers of Information Technology & Electronic Engineering* 18.1 (2017), págs. 3-14. DOI: 10.1631/FITEE.1601883 (vid. pág. 33).
- [97] Konstantinos Vassakis, Emmanuel Petrakis y Ioannis Kopanakis. «Big Data Analytics: Applications, Prospects and Challenges». En: *Mobile big data*. Springer, 2018, págs. 3-20. DOI: 10.1007/978-3-319-67925-9_1 (vid. pág. 33).
- [98] Roberto V Zicari. «Big Data: Challenges and Opportunities». En: *Big data computing* 564 (2014), pág. 103 (vid. pág. 33).
- [99] Abhay Kumar Bhadani y Dhanya Jothimani. «Big Data: Challenges, Opportunities, and Realities». En: *Effective big data management and opportunities for implementation*. IGI Global, 2016, págs. 1-24. DOI: 10.4018/978-1-5225-0182-4.ch001 (vid. pág. 33).
- [100] Pall Rikhardsson y Ogan Yigitbasioglu. «Business intelligence & analytics in management accounting research: Status and future focus». En: *International Journal of Accounting Information Systems* 29 (2018), págs. 37-58. DOI: 10.1016/j.accinf.2018.03.001 (vid. pág. 34).
- [101] Jack G Zheng. «Data visualization for business intelligence». En: *Global Business Intelligence* (2017), págs. 67-82. DOI: 10.4324/9781315471136-6 (vid. pág. 34).
- [102] Marcello Mariani, Rodolfo Baggio, Matthias Fuchs y Wolfram Höepken. «Business intelligence and big data in hospitality and tourism: a systematic literature review». En: *International Journal of Contemporary Hospitality Management* (2018). DOI: 10.1108/IJCHM-07-2017-0461 (vid. págs. 34, 35).
- [103] Pravin Chandra y Manoj K Gupta. «Comprehensive survey on data warehousing research». En: *International Journal of Information Technology* 10.2 (2018), págs. 217-224. DOI: 10.1007/s41870-017-0067-y (vid. págs. 35, 36).
- [104] Tim A Majchrzak, Tobias Jansen y Herbert Kuchen. «Efficiency evaluation of open source ETL tools». En: *Proceedings of the 2011 ACM symposium on applied computing*. 2011, págs. 287-294. DOI: 10.1145/1982185.1982251 (vid. págs. 36, 39).
- [105] Abbas Raza Ali. «Real-time big data warehousing and analysis framework». En: *2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA)*. IEEE. 2018, págs. 43-49. DOI: 10.1109/ICBDA.2018.8367649 (vid. pág. 36).

- [106] Mohammed K Hassan, Ali I El Desouky, Sally M Elghamrawy y Amany M Sarhan. «Big Data Challenges and Opportunities in Healthcare Informatics and Smart Hospitals». En: *Security in smart cities: Models, applications, and challenges*. Springer, 2019, págs. 3-26. DOI: 10.1007/978-3-030-01560-2_1 (vid. pág. 36).
- [107] Neepa Biswas, Anamitra Sarkar y Kartick Chandra Mondal. «Efficient incremental loading in ETL processing for real-time data integration». En: *Innovations in Systems and Software Engineering* 16.1 (2020), págs. 53-61. DOI: 10.1007/s11334-019-00344-4 (vid. págs. 36, 39, 56).
- [108] Bonginkosi Gina y Adheesh Budree. «A review of literature on critical factors that drive the selection of business intelligence tools». En: *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*. IEEE. 2020, págs. 1-7. DOI: 10.1109/icABCD49160.2020.9183852 (vid. págs. 36, 38).
- [109] ISO/IEC. *ISO/IEC 9126. Software engineering – Product quality*. ISO/IEC, 2001 (vid. pág. 36).
- [110] Ralph Kimball. *The data warehouse toolkit: practical techniques for building dimensional data warehouses*. John Wiley & Sons, Inc., 1996 (vid. pág. 36).
- [111] Vaishali A Kherdekar y Pravin S Metkewar. «A technical comprehensive survey of ETL tools». En: *International Journal of Applied Engineering Research* 11.4 (2016), págs. 2557-2559. DOI: 10.37622/IJAER/11.4.2016.2557-2559 (vid. págs. 38, 39, 56).
- [112] Hong Sun, Kristof Depraetere, Jos De Roo, Giovanni Mels, Boris De Vloed, Marc Twagirimukiza y Dirk Colaert. «Semantic processing of EHR data for clinical research». En: *Journal of biomedical informatics* 58 (2015), págs. 247-259. DOI: 10.1016/j.jbi.2015.10.009 (vid. pág. 40).
- [113] Udayan Khurana y Sainyam Galhotra. «Semantic Concept Annotation for Tabular Data». En: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2021, págs. 844-853. DOI: 10.1145/3459637.3482295 (vid. pág. 40).
- [114] Zhenbang Wu, Cao Xiao, Lucas M Glass, David M Liebovitz y Jimeng Sun. «AutoMap: Automatic Medical Code Mapping for Clinical Prediction Model Deployment». En: *arXiv preprint arXiv:2203.02446* (2022). DOI: 10.48550/arXiv.2203.02446 (vid. pág. 40).
- [115] Rudra Pratap Deb Nath, Katja Hose y Torben Bach Pedersen. «Towards a programmable semantic extract-transform-load framework for semantic data warehouses». En: *Proceedings of the ACM Eighteenth International Workshop on Data Warehousing and OLAP*. ACM. 2015, págs. 15-24. DOI: 10.1145/2811222.2811229 (vid. pág. 44).
- [116] Philip T James. «Obesity: the worldwide epidemic». En: *Clinics in dermatology* 22.4 (2004), págs. 276-280. DOI: 10.1016/j.clindermatol.2004.01.010 (vid. pág. 44).
- [117] Tania Tudorache, Csongor Nyulas, Natalya F Noy y Mark A Musen. «WebProtégé: A collaborative ontology editor and knowledge acquisition tool for the web». En: vol. 4. 1. IOS Press, 2013, págs. 89-99. DOI: 10.3233/SW-2012-0057 (vid. págs. 44, 83).

- [118] João Rafael Almeida y José Luís Oliveira. «Multi-language Concept Normalisation of Clinical Cohorts». En: *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*. 2020, págs. 261-264. DOI: 10.1109/CBMS49503.2020.00056 (vid. pág. 47).
- [119] Shengjun Hong, Dmitry Prokopenko, Valerija Dobricic, Fabian Kilpert, Isabelle Bos, Stephanie JB Vos, Betty M Tijms, Ulf Andreasson, Kaj Blennow, Rik Vandenberghe y col. «Genome-wide association study of Alzheimer's disease CSF biomarkers in the EMIF-AD Multimodal Biomarker Discovery dataset». En: *Translational psychiatry* 10.1 (2020), págs. 1-12. DOI: 10.1038/s41398-020-01074-z (vid. págs. 49, 149).
- [120] Aurore Delvenne, Johan Gobom, Betty M Tijms, Isabelle Bos, Frans RJ Verhey, Inez HGB Ramakers, Philip Scheltens, Charlotte E Teunissen, Rik Vandenberghe, Silvy Gabel y col. «CSF proteomic profiling of mild cognitive impairment individuals with suspected non-Alzheimer's disease pathophysiology: Developing topics». En: *Alzheimer's & Dementia* 16 (2020), e047247. DOI: 10.1002/alz.047247 (vid. págs. 49, 149).
- [121] Matt Casters, Roland Bouman y Jos Van Dongen. *Pentaho Kettle solutions: building open source ETL solutions with Pentaho Data Integration*. John Wiley & Sons, 2010 (vid. pág. 51).
- [122] Stephanie JB Vos, Frans Verhey, Lutz Frölich, Johannes Kornhuber, Jens Wiltfang, Wolfgang Maier, Oliver Peters, Eckart Rütther, Flavio Nobili, Silvia Morbelli y col. «Prevalence and prognosis of Alzheimer's disease at the mild cognitive impairment stage». En: *Brain* 138.5 (2015), págs. 1327-1338. DOI: 10.1093/brain/awv029 (vid. pág. 58).
- [123] Willemijn J Jansen, Rik Ossenkoppele, Dirk L Knol, Betty M Tijms, Philip Scheltens, Frans RJ Verhey, Pieter Jelle Visser, Pauline Aalten, Dag Aarsland, Daniel Alcolea y col. «Prevalence of cerebral amyloid pathology in persons without dementia: a meta-analysis». En: *Jama* 313.19 (2015), págs. 1924-1938. DOI: 10.1001/jama.2015.4668 (vid. pág. 58).
- [124] Athena - OHDSI Vocabularies Repository. <https://athena.ohdsi.org/search-terms/terms/terms/45768723>. Accessed: 2022-04-23. 2022 (vid. págs. 59, 85).
- [125] Sharyl J Nass, Laura A Levit, Lawrence O Gostin y col. «The value, importance, and oversight of health research». En: *National Academies Press (US)* (2009) (vid. pág. 69).
- [126] Hui G Cheng y Michael R Phillips. «Secondary analysis of existing data: opportunities and implementation». En: *Shanghai archives of psychiatry* 26.6 (2014), pág. 371. DOI: 10.11919/j.issn.1002-0829.214171 (vid. pág. 69).
- [127] Dimitrios G. Katehakis y Manolis Tsiknakis. *Electronic Health Record*. John Wiley & Sons, 2006. DOI: 10.1002/9780471740360.ebs1440 (vid. pág. 69).
- [128] Heather A Piwowar y Wendy W Chapman. «Public sharing of research datasets: a pilot study of associations». En: *Journal of informetrics* 4.2 (2010), págs. 148-156. DOI: 10.1016/j.joi.2009.11.010 (vid. pág. 69).
- [129] Kasper Jensen, Cristina Soguero-Ruiz, Karl Oyvind Mikalsen, Rolv-Ole Lindsetmo, Irene Kouskoumvekaki, Mark Girolami, Stein Olav Skrovseth y Knut Magne Augestad. «Analysis of free text in electronic health records for identification of cancer patient trajectories». En: *Scientific reports* 7.1 (2017), págs. 1-12. DOI: 10.1038/srep46226 (vid. pág. 70).

- [130] Stuart J Nelson, Kelly Zeng, John Kilbourne, Tammy Powell y Robin Moore. «Normalized names for clinical drugs: RxNorm at 6 years». En: *Journal of the American Medical Informatics Association* 18.4 (2011), págs. 441-448 (vid. pág. 70).
- [131] David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang y Jennifer Woolsey. «DrugBank: a comprehensive resource for in silico drug discovery and exploration». En: *Nucleic acids research* 34.suppl_1 (2006), págs. D668-D672. DOI: 10.1093/nar/gkj067 (vid. pág. 70).
- [132] Jennifer Liang, Ching-Huei Tsou y Ananya Poddar. «A novel system for extractive clinical note summarization using EHR data». En: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. 2019, págs. 46-54. DOI: 10.18653/v1/W19-1906 (vid. pág. 70).
- [133] Sunyang Fu, David Chen, Huan He, Sijia Liu, Sungrim Moon, Kevin J Peterson, Feichen Shen, Liwei Wang, Yanshan Wang, Andrew Wen y col. «Clinical concept extraction: a methodology review». En: *Journal of biomedical informatics* 109 (2020), pág. 103526. DOI: 10.1016/j.jbi.2020.103526 (vid. pág. 72).
- [134] Abhyuday Jagannatha, Feifan Liu, Weisong Liu y Hong Yu. «Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0)». En: *Drug safety* 42.1 (2019), págs. 99-111. DOI: 10.1007/s40264-018-0762-z (vid. pág. 72).
- [135] Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs y Ozlem Uzuner. «2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records». En: *Journal of the American Medical Informatics Association* 27.1 (oct. de 2019), págs. 3-12. DOI: 10.1093/jamia/ocz166 (vid. págs. 72, 92).
- [136] Sunghwan Sohn, Cheryl Clark, Scott R Halgrim, Sean P Murphy, Christopher G Chute y Hongfang Liu. «MedXN: an open source medication extraction and normalization tool for clinical text». En: *Journal of the American Medical Informatics Association* 21.5 (2014), págs. 858-865. DOI: 10.1136/amiajnl-2013-002190 (vid. págs. 73, 77).
- [137] Hannah L Weeks, Cole Beck, Elizabeth McNeer, Michael L Williams, Cosmin A Bejan, Joshua C Denny y Leena Choi. «medExtractR: A targeted, customizable approach to medication extraction from electronic health records». En: *Journal of the American Medical Informatics Association* 27.3 (2020), págs. 407-418. DOI: 10.1093/jamia/ocz207 (vid. pág. 73).
- [138] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler y Christopher G Chute. «Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications». En: *Journal of the American Medical Informatics Association* 17.5 (2010), págs. 507-513 (vid. pág. 73).
- [139] Sérgio Matos. «Configurable web-services for biomedical document annotation». En: *Journal of cheminformatics* 10.1 (2018), pág. 68. DOI: 10.1186/s13321-018-0317-4 (vid. págs. 73, 76, 88, 177).
- [140] Cássia Trojahn, Bo Fu, Ondřej Zamazal y Dominique Ritze. «State-of-the-art in multi-lingual and cross-lingual ontology matching». En: *Towards the Multilingual Semantic Web*. Springer, 2014, págs. 119-135. DOI: 10.1007/978-3-662-43585-4_8 (vid. pág. 73).

- [141] Gábor Bella, Fausto Giunchiglia y Fiona McNeill. «Language and domain aware lightweight ontology matching». En: *Journal of Web Semantics* 43 (2017), págs. 1 -17. ISSN: 1570-8268. DOI: 10.2139/ssrn.3199131 (vid. pág. 73).
- [142] David Campos, Sérgio Matos y José Luís Oliveira. «A modular framework for biomedical concept recognition». En: *BMC bioinformatics* 14.1 (2013), pág. 281. DOI: 10.1186/1471-2105-14-281 (vid. pág. 73).
- [143] Tiago Nunes, David Campos, Sérgio Matos y José Luís Oliveira. «BeCAS: biomedical concept recognition services and visualization». En: *Bioinformatics* 29.15 (2013), págs. 1915-1916. DOI: 10.1093/bioinformatics/btt317 (vid. pág. 73).
- [144] João Figueira Silva, Rui Antunes, João Rafael Almeida y Sérgio Matos. «Clinical concept normalization on medical records using word embeddings and heuristics». En: *30th Medical Informatics Europe conference, MIE*. 2020. DOI: 10.3233/SHTI200129 (vid. pág. 74).
- [145] Sergey Goryachev, Hyeoneui Kim y Qing Zeng-Treitler. «Identification and extraction of family history information from clinical reports». En: *AMIA Annual Symposium Proceedings*. Vol. 2008. American Medical Informatics Association. 2008, pág. 247 (vid. pág. 74).
- [146] Jeff Friedlin y Clement J McDonald. «Using a natural language processing system to extract and code family history data from admission reports». En: *AMIA Annual Symposium Proceedings*. Vol. 2006. American Medical Informatics Association. 2006, pág. 925 (vid. pág. 75).
- [147] Robert Bill, Serguei Pakhomov, Elizabeth S Chen, Tamara J Winden, Elizabeth W Carter y Genevieve B Melton. «Automated extraction of family history information from clinical notes». En: *AMIA Annual Symposium Proceedings*. Vol. 2014. American Medical Informatics Association. 2014, pág. 1709 (vid. pág. 75).
- [148] Olivier Bodenreider. «The Unified Medical Language System (UMLS): integrating biomedical terminology». En: *Nucleic Acids Research* 32.suppl_1 (ene. de 2004), págs. D267-D270. DOI: 10.1093/nar/gkh061 (vid. págs. 76, 177, 191).
- [149] *n2c2 Shared-Task and Workshop Track 2: n2c2/OHNL Track on Family History Extraction*. <https://n2c2.dbmi.hms.harvard.edu/track2>. Accessed: 2022-04-26. 2019 (vid. págs. 87, 97).
- [150] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard y David McClosky. «The Stanford CoreNLP natural language processing toolkit». En: *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 2014, págs. 55-60. DOI: 10.3115/v1/P14-5010 (vid. pág. 87).
- [151] Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar y Guergana Savova. «Semeval-2014 task 7: Analysis of clinical text». En: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. 2014, págs. 54-62. DOI: 10.3115/v1/S14-2007 (vid. pág. 88).
- [152] Özlem Uzuner, Imre Solti y Eithon Cadag. «Extracting medication information from clinical text». En: *Journal of the American Medical Informatics Association* 17.5 (sep. de 2010), págs. 514-518. DOI: 10.1136/jamia.2010.003947 (vid. pág. 92).

- [153] Lars Bertram, Anke Böckenhoff, Ilja Demuth, Sandra Düzel, Rahel Eckardt, Shu-Chen Li, Ulman Lindenberger, Graham Pawelec, Thomas Siedler, Gert G Wagner y col. «Cohort profile: the Berlin aging study II (BASE-II)». En: *International journal of epidemiology* 43.3 (2014), págs. 703-712 (vid. pág. 96).
- [154] Miranda T Schram, Simone JS Sep, Carla J van der Kallen, Pieter C Dagnelie, An-nemarie Koster, Nicolaas Schaper, Ronald MA Henry y Coen DA Stehouwer. «The Maastricht Study: an extensive phenotyping study on determinants of type 2 diabetes, its complications and its comorbidities». En: *European journal of epidemiology* 29.6 (2014), págs. 439-451 (vid. pág. 96).
- [155] Sijia Liu, Yanshan Wang, Andrew Wen, Liwei Wang, Na Hong, Feichen Shen, Steven Bedrick, William Hersh y Hongfang Liu. «Implementation of a cohort retrieval system for clinical data repositories using the observational medical outcomes partnership common data model: Proof-of-concept system validation». En: *JMIR medical informatics* 8.10 (2020), e17376. DOI: 10.2196/17376 (vid. págs. 102, 179, 192).
- [156] Jimyung Park, Seng Chan You, Eugene Jeong, Chunhua Weng, Dongsu Park, Jin Roh, Dong Yun Lee, Jae Youn Cheong, Jin Wook Choi, Mira Kang y col. «A Framework (SOCRAteX) for Hierarchical Annotation of Unstructured Electronic Health Records and Integration Into a Standardized Medical Database: Development and Usability Study». En: *JMIR Medical Informatics* 9.3 (2021), e23983. DOI: 10.2196/23983 (vid. págs. 103, 179, 192).
- [157] Simon Lovestone y EMIF Consortium. «The European medical information framework: A novel ecosystem for sharing healthcare data across Europe». En: *Learning Health Systems* 4.2 (2020), e10214. DOI: 10.1002/lrh2.10214 (vid. págs. 107, 179, 193).
- [158] José Luís Oliveira, Alina Trifan y Luís A Bastião Silva. «EMIF Catalogue: a collaborative platform for sharing and reusing biomedical data». En: *International journal of medical informatics* 126 (2019), págs. 35-45. DOI: 10.1016/j.ijmedinf.2019.02.006 (vid. págs. 107, 113, 179, 193).
- [159] Xiaoling Chen, Anupama E Gururaj, Burak Ozyurt, Ruiling Liu, Ergin Soysal, Trevor Cohen, Firat Tiryaki, Yueling Li, Nansu Zong, Min Jiang y col. «DataMed—an open source discovery index for finding biomedical datasets». En: *Journal of the American Medical Informatics Association* 25.3 (2018), págs. 300-308. DOI: 10.1093/jamia/ocx121 (vid. pág. 110).
- [160] Susanna-Assunta Sansone, Alejandra Gonzalez-Beltran, Philippe Rocca-Serra, George Alter, Jeffrey S Grethe, Hua Xu, Ian M Fore, Jared Lyle, Anupama E Gururaj, Xiaoling Chen y col. «DATS, the data tag suite to enable discoverability of datasets». En: *Scientific data* 4.1 (2017), págs. 1-8. DOI: 10.1038/sdata.2017.59 (vid. pág. 110).
- [161] Alejandra N Gonzalez-Beltran, John Campbell, Patrick Dunn, Diana Guijarro, Sanda Ionescu, Hyeoneui Kim, Jared Lyle, Jeffrey Wiser, Susanna-Assunta Sansone y Philippe Rocca-Serra. «Data discovery with DATS: exemplar adoptions and lessons learned». En: *Journal of the American Medical Informatics Association* 25.1 (2018), págs. 13-16. DOI: 10.1093/jamia/ocx119 (vid. pág. 110).
- [162] Tyler J Skluzacek. «Dredging a data lake: decentralized metadata extraction». En: *Proceedings of the 20th International Middleware Conference Doctoral Symposium*. 2019, págs. 51-53. DOI: 10.1145/3366624.3368170 (vid. pág. 110).

- [163] Tyler J Skluzacek, Rohan Kumar, Ryan Chard, Galen Harrison, Paul Beckman, Kyle Chard y Ian T Foster. «Skluma: An extensible metadata extraction pipeline for disorganized data». En: *2018 IEEE 14th International Conference on e-Science (e-Science)*. IEEE. 2018, págs. 256-266. DOI: 10.1109/escience.2018.00040 (vid. pág. 111).
- [164] Alina Trifan y José Luís Oliveira. «Patient data discovery platforms as enablers of biomedical and translational research: A systematic review». En: *Journal of Biomedical Informatics* 93 (2019), pág. 103154. DOI: 10.1016/j.jbi.2019.103154 (vid. pág. 112).
- [165] Owen Lancaster, Tim Beck, David Atlan, Morris Swertz, Dhiwagaran Thangavelu, Colin Veal, Raymond Dagleish y Anthony J Brookes. «Cafe Variome: General-purpose software for making genotype–phenotype data discoverable in restricted or open access contexts». En: *Human mutation* 36.10 (2015), págs. 957-964. DOI: 10.1002/humu.22841 (vid. pág. 112).
- [166] Peter McQuilton, Alejandra Gonzalez-Beltran, Philippe Rocca-Serra, Milo Thurston, Allyson Lister, Eamonn Maguire y Susanna-Assunta Sansone. «BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences». En: *Database* 2016 (2016). DOI: 10.1093/database/baw075 (vid. pág. 112).
- [167] Susanna-Assunta Sansone, Peter McQuilton, Philippe Rocca-Serra, Alejandra Gonzalez-Beltran, Massimiliano Izzo, Allyson L Lister y Milo Thurston. «FAIRsharing as a community approach to standards, repositories and policies». En: *Nature biotechnology* 37.4 (2019), págs. 358-367. DOI: 10.1038/s41587-019-0080-8 (vid. pág. 112).
- [168] Vineet Jamwal y Simran Kaur. «Global presence of open-source research data management platform for libraries: the Dataverse project». En: *Library Hi Tech News* (2021). DOI: 10.1108/LHTN-10-2021-0066 (vid. pág. 112).
- [169] Luís Bastião Silva, Alina Trifan y José Luís Oliveira. «MONTRA: An agile architecture for data publishing and discovery». En: *Computer methods and programs in biomedicine* 160 (2018), págs. 33-42. DOI: 10.1016/j.cmpb.2018.03.024 (vid. pág. 113).
- [170] John PA Ioannidis, Sander Greenland, Mark A Hlatky, Muin J Khoury, Malcolm R Macleod, David Moher, Kenneth F Schulz y Robert Tibshirani. «Increasing value and reducing waste in research design, conduct, and analysis». En: *The Lancet* 383.9912 (2014), págs. 166-175. DOI: 10.1016/S0140-6736(13)62227-8 (vid. pág. 113).
- [171] Ning Shang, Chunhua Weng y George Hripcsak. «A conceptual framework for evaluating data suitability for observational studies». En: *Journal of the American Medical Informatics Association* (2017). DOI: 10.1093/jamia/ocx095 (vid. pág. 113).
- [172] Pascal Coorevits, M Sundgren, Gunnar O Klein, A Bahr, B Claerhout, C Daniel, M Dugas, D Dupont, A Schmidt, P Singleton y col. «Electronic health records: new opportunities for clinical research». En: *Journal of internal medicine* 274.6 (2013), págs. 547-560. DOI: 10.1111/joim.12119 (vid. pág. 113).
- [173] Marco Brandizi, Olga Melnichuk, Raffael Bild, Florian Kohlmayer, Benedicto Rodriguez-Castro, Helmut Spengler, Klaus A Kuhn, Wolfgang Kuchinke, Christian Ohmann, Timo Mustonen y col. «Orchestrating differential data access for translational research: a pilot implementation». En: *BMC medical informatics and decision making* 17.1 (2017), pág. 30. DOI: 10.1186/s12911-017-0424-6 (vid. pág. 113).

- [174] Chee Sun Liew, Malcolm P Atkinson, Michelle Galea, Tan Fong Ang, Paul Martin y Jano I Van Hemert. «Scientific workflows: moving across paradigms». En: *ACM Computing Surveys (CSUR)* 49.4 (2017), pág. 66. DOI: 10.1145/3012429 (vid. pág. 113).
- [175] Sonja Holl, Olav Zimmermann, Magnus Palmblad, Yassene Mohammed y Martin Hofmann-Apitius. «A new optimization phase for scientific workflow management systems». En: *Future generation computer systems* 36 (2014), págs. 352-362. DOI: 10.1016/j.future.2013.09.005 (vid. pág. 113).
- [176] Katherine Wolstencroft, Robert Haines, Donal Fellows, Alan Williams, David Withers, Stuart Owen, Stian Soiland-Reyes, Ian Dunlop, Aleksandra Nenadic, Paul Fisher y col. «The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud». En: *Nucleic acids research* 41.W1 (2013), W557-W561. DOI: 10.1093/nar/gkt328 (vid. pág. 114).
- [177] Jeremy Goecks, Anton Nekrutenko y James Taylor. «Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences». En: *Genome biology* 11.8 (2010), R86. DOI: 10.1186/gb-2010-11-8-r86 (vid. pág. 114).
- [178] Pedro Lopes y José Luís Oliveira. «An automated real-time integration and interoperability framework for bioinformatics». En: *BMC bioinformatics* 16 (2015), págs. 328-328. DOI: 10.1186/s12859-015-0761-3 (vid. pág. 114).
- [179] Anubhav Jain, Shyue Ping Ong, Wei Chen, Bharat Medasani, Xiaohui Qu, Michael Kocher, Miriam Brafman, Guido Petretto, Gian-Marco Rignanes, Geoffroy Hautier y col. «FireWorks: A dynamic workflow system designed for high-throughput applications». En: *Concurrency and Computation: Practice and Experience* 27.17 (2015), págs. 5037-5059. DOI: 10.1002/cpe.3505 (vid. pág. 114).
- [180] Han Bing y Xia Dan-Mei. «Research and design of document flow model based on JBPM workflow engine». En: *Computer Science-Technology and Applications, 2009. IFCSTA'09. International Forum on*. Vol. 1. IEEE. 2009, págs. 336-339. DOI: 10.1109/IFCSTA.2009.88 (vid. pág. 115).
- [181] James Martin. *Rapid application development*. Macmillan Publishing Co., Inc., 1991 (vid. pág. 117).
- [182] Luis Bastiao Silva, Rafael C Jimenez, Niklas Blomberg y José Luis Oliveira. «General guidelines for biomedical software development». En: *F1000Research* 6 (2017) (vid. pág. 117).
- [183] Arnaldo Pereira, Rui Pedro Lopes y José Luís Oliveira. «SCALEUS-FD: A fair data tool for biomedical applications». En: *BioMed Research International* 2020 (2020). DOI: 10.1155/2020/3041498 (vid. pág. 125).
- [184] Scott Cantor, Jahan Moreh, Rob Philpott y Eve Maler. *Metadata for the OASIS security assertion markup language (SAML) V2. 0*. 2005 (vid. pág. 125).
- [185] Dick Hardt y col. *The OAuth 2.0 authorization framework*. 2012 (vid. pág. 126).
- [186] Natsuhiko Sakimura, John Bradley, Mike Jones, Breno De Medeiros y Chuck Mortimore. «Openid connect core 1.0». En: *The OpenID Foundation* (2014), S3 (vid. pág. 126).

- [187] Nitin Naik y Paul Jenkins. «Securing digital identities in the cloud by selecting an apposite Federated Identity Management from SAML, OAuth and OpenID Connect». En: *2017 11th International Conference on Research Challenges in Information Science (RCIS)*. IEEE. 2017, págs. 163-174. DOI: 10.1109/RCIS.2017.7956534 (vid. pág. 126).
- [188] Alvaro Alonso, Alejandro Pozo, Johnny Choque, Gloria Bueno, Joaquín Salvachúa, Luis Diez, Jorge Marín y Pedro Luis Chas Alonso. «An identity framework for providing access to FIWARE OAuth 2.0-based services according to the eIDAS European regulation». En: *IEEE Access* 7 (2019), págs. 88435-88449. DOI: 10.1109/ACCESS.2019.2926556 (vid. pág. 126).
- [189] Dave Beckett y Brian McBride. «RDF/XML syntax specification (revised)». En: *W3C recommendation* 10.2.3 (2004) (vid. pág. 128).
- [190] Emanuel Lacic, Dominik Kowald, Denis Parra, Martin Kahr y Christoph Trattner. «Towards a scalable social recommender engine for online marketplaces: The case of apache solr». En: *Proceedings of the 23rd International Conference on World Wide Web*. ACM. 2014, págs. 817-822. DOI: 10.1145/2567948.2579245 (vid. pág. 128).
- [191] Bernard Fortz, Olga Oliveira y Cristina Requejo. «Compact mixed integer linear programming models to the minimum weighted tree reconstruction problem». En: *European journal of operational research* 256.1 (2017), págs. 242-251. DOI: 10.1016/j.ejor.2016.06.014 (vid. pág. 128).
- [192] Yehuda Koren, Steffen Rendle y Robert Bell. «Advances in collaborative filtering». En: *Recommender systems handbook* (2022), págs. 91-142. DOI: 10.1007/978-1-0716-2197-4_3 (vid. págs. 128, 181, 195).
- [193] David Goldberg, David Nichols, Brian M Oki y Douglas Terry. «Using collaborative filtering to weave an information tapestry». En: *Communications of the ACM* 35.12 (1992), págs. 61-71. DOI: 10.1145/138859.138867 (vid. pág. 129).
- [194] Badrul Munir Sarwar, George Karypis, Joseph A Konstan, John Riedl y col. «Item-based collaborative filtering recommendation algorithms». En: *Www* 1 (2001), págs. 285-295. DOI: 10.1145/371920.372071 (vid. pág. 129).
- [195] Michael J Pazzani y Daniel Billsus. «Content-based recommendation systems». En: *The adaptive web*. Springer, 2007, págs. 325-341. DOI: 10.1007/978-3-540-72079-9_10 (vid. págs. 129, 181, 195).
- [196] Zohreh Dehghani Champiri, Seyed Reza Shahamiri y Siti Salwah Binti Salim. «A systematic review of scholar context-aware recommender systems». En: *Expert Systems with Applications* 42.3 (2015), págs. 1743-1758. DOI: 10.1016/j.eswa.2014.09.017 (vid. pág. 130).
- [197] Michael J Pazzani, Jack Muramatsu, Daniel Billsus y col. «Syskill & Webert: Identifying interesting web sites». En: *AAAI/IAAI, Vol. 1*. 1996, págs. 54-61 (vid. pág. 130).
- [198] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne y col. «The FAIR Guiding Principles for scientific data management and stewardship». En: *Scientific data* 3.1 (2016), págs. 1-9 (vid. págs. 142, 151).

- [199] J Ben Schafer, Joseph A Konstan y John Riedl. «E-commerce recommendation applications». En: *Data mining and knowledge discovery* 5.1-2 (2001), págs. 115-153. DOI: 10.1023/A:1009804230409 (vid. pág. 143).
- [200] Igor Kulev, Elena Vlahu-Gjorgievska, Vladimir Trajkovic y Saso Koceski. «Development of a novel recommendation algorithm for collaborative health: Care system model». En: *Computer Science and Information Systems* 10.3 (2013), págs. 1455-1471. DOI: 10.2298/CSIS120921057K (vid. pág. 143).

Sinopsis in Spanish

A.1 Introducción

La continua demanda de mejores diagnósticos y tratamientos de salud ha motivado muchos estudios de investigación clínica, como estudios observacionales y ensayos clínicos. En los ensayos clínicos, los pacientes se dividen comúnmente en dos o más grupos (*p. ej.* activo y placebo), para estudiar la efectividad del tratamiento para una condición clínica particular [1]. En este caso, hay intervención directa con los pacientes, *p. ej.*, administración de un fármaco o procedimientos terapéuticos. Sin embargo, este enfoque no siempre es el más apropiado, *p. ej.* abordar preguntas de investigación en cirugía plástica a través de ensayos controlados aleatorios a menudo está sujeto a restricciones éticas [2]. En los estudios observacionales, los investigadores no realizan ninguna intervención activa con los pacientes y la exposición se produce de forma natural o a través de otros factores. Aquí, los investigadores médicos se limitan a documentar la relación entre la exposición y el resultado del estudio [1].

Los estudios observacionales siguen diferentes estrategias y se establecen definiendo un conjunto de criterios de inclusión y exclusión para los sujetos involucrados, así como varias características que se identifican y observan a lo largo del tiempo [3]. Algunas iniciativas reutilizan los datos ya recopilados en instituciones médicas para realizar estudios observacionales. Esta práctica ahorra tiempo y permite la verificación previa del número de sujetos antes de iniciar el análisis [4]. Sin embargo, en enfermedades específicas, es necesario recopilar información sobre los sujetos seleccionados. En estas situaciones, los datos se registran según las pautas del estudio y se pueden usar diferentes soluciones para almacenar los datos, por ejemplo, el EHR system [5] institucional.

La dependencia de los equipos técnicos es una gran barrera cuando existe la necesidad de extraer datos para su análisis. Muchas otras cuestiones éticas y técnicas surgen cuando se desea combinar conjuntos de datos de distintas organizaciones. Este es el

caso de los estudios multicéntricos que pretenden aumentar el tamaño de la población, el poder de la evidencia estadística y, por tanto, el impacto del estudio [6].

La integración de múltiples fuentes de datos no es solo un problema tecnológico. Hay algunas herramientas ETL capaces de realizar esta tarea utilizando grandes cantidades de datos. El problema no técnico de esta agregación es el dominio de los datos, es decir, identificar los conceptos en los datos y combinar correctamente la información asociada a ellos que se extrajo de múltiples fuentes. Las bases de datos de salud pertenecen a uno de los dominios en los que esto es un problema preocupante, debido a la variedad de conceptos para representar procedimientos y términos médicos similares. Resolver estos problemas es útil para optimizar los estudios, pero compartir datos a nivel de paciente aún plantea algunos problemas de privacidad, debido a requisitos legales, éticos y reglamentarios [8]. Los datos de los pacientes son muy sensibles y la interrupción de esta privacidad puede tener consecuencias dramáticas para las personas, los proveedores de atención médica y los subgrupos dentro de la sociedad [9]. Además, la legislación puede ser diferente en cada país, lo que dificulta definir un protocolo que se ajuste a todas las instituciones involucradas [10]. Este es otro desafío, que requiere encontrar una solución que permita el análisis de múltiples fuentes de datos, o partes de estas, sin exponer datos confidenciales.

El impacto potencial de los estudios multicéntricos ha motivado a los investigadores a buscar soluciones más sólidas y reutilizables para agregar conocimientos a partir de conjuntos de datos de salud distribuidos. Se establecieron organizaciones y metodologías para explorar bases de datos clínicas mediante la reutilización de datos existentes [4]. Uno de estos esfuerzos tiene como objetivo crear una estrategia para reutilizar las bases de datos EHR utilizando un esquema homogéneo, con el fin de facilitar la interoperabilidad entre las bases de datos. Actualmente, esta integración es posible mediante el uso de marcos de código abierto que ayudan a respaldar todo el proceso [7].

Los objetivos de investigación de este trabajo se pueden parafrasear en varias preguntas enfocadas en abordar problemas específicos. Reconocemos que los estudios médicos multicéntricos pueden plantear problemas adicionales que no serán considerados en este trabajo, principalmente debido a los tipos de datos utilizados en estos. Por ejemplo, estudios de investigación basados en biomarcadores que se correlacionan con información de DNA, o estudios centrados principalmente en el uso de imágenes médicas. Por lo tanto, para lograr un escenario capaz de respaldar estudios

de salud distribuidos utilizando múltiples fuentes de datos de distintas instituciones, limitamos el alcance a las bases de datos EHR. Este objetivo se logrará respondiendo a las siguientes preguntas de investigación:

1. *¿Cómo ejecutar una consulta de base de datos sobre una red de bases de datos de salud heterogéneas?* La falta de interoperabilidad es el principal problema en este escenario. Un ecosistema con bases de datos heterogéneas puede no compartir el mismo esquema de datos, lo que invalida el uso compartido de consultas. Además de este problema, el dominio de la salud contiene una gran cantidad de conceptos médicos, que pueden diferir entre instituciones, a nivel nacional o internacional. Se requiere una metodología para armonizar dichas bases de datos, que las convierta en un formato interoperable. Este proceso puede tener diferentes etapas y componentes, y algunos de ellos serán automatizados con el fin de reducir el costo y tiempo de ejecución de este procedimiento. Esto puede resultar en una nueva solución de software.
2. *¿Cómo seleccionar las bases de datos de salud más adecuadas para un estudio de investigación específico?* Esta pregunta puede abordarse desde diferentes perspectivas. Sin embargo, al correlacionarlo con los enunciados anteriores, identificamos que los principales problemas que enfrentan los investigadores médicos son: i) el descubrimiento de bases de datos de interés; y ii) el acceso a dichas bases de datos sin violar políticas de privacidad y normas éticas. La solución a estos problemas es demasiado compleja para ser resuelta por una sola aplicación de software. Para responder a esta pregunta, es necesario crear un ecosistema de herramientas y metodologías, en el que los propietarios de los datos puedan sentirse seguros al compartir características sobre sus bases de datos, mientras que los investigadores puedan tener suficiente información para seleccionar las bases de datos que mejor se adapten a sus necesidades. Por tanto, la solución propuesta para este problema será un portal que integre: i) un catálogo web de características de la base de datos; ii) herramientas para visualizar y comparar estas características; y iii) herramientas para orquestar estudios distribuidos.

A.2 Traducción semiautomática de fuentes de datos a un esquema común

La falta de flexibilidad en la colaboración de los usuarios durante el diseño y la definición de las canalizaciones ETL es un problema para algunos dominios de aplicación. Por ejemplo, en el escenario médico, cuando los datos clínicos deben armonizarse en un esquema de datos común, esto requiere la colaboración entre los equipos técnicos y los investigadores médicos [73]. Esta colaboración se requiere en diferentes etapas: i) diseño; ii) implementación; y iii) validación. En cada etapa, hay algunos desafíos que abordamos en este trabajo.

A.2.1 Metodología para la armonización de cohortes

Proponemos una metodología basada en los principios ETL. En la etapa de extracción, los datos de origen seleccionados se leen extrayéndolos de una o varias fuentes de datos. El objetivo principal de esta etapa es obtener los datos de los sistemas de origen sin interferir con su rendimiento habitual. En las bases de datos de salud, esta es una tarea delicada porque el EHR no se puede sobrecargar debido al procedimiento de extracción de datos. Sin embargo, en estudios clínicos, la cantidad de datos no es suficiente para colapsar los sistemas durante esta etapa. Además, los estudios clínicos se exportaron a un formato tabular, que no requiere interacción directa con el sistema utilizado para recopilar los datos del paciente.

La etapa de transformación es el componente más complejo de todas las etapas. Esta etapa requiere el mapeo de la base de datos de origen en el esquema de destino, así como la armonización del contenido. Para una fuente de datos, este procedimiento requiere un mapeo completo, lo que lleva mucho tiempo y requiere entidades especializadas para validar los mapeos. La armonización de contenido podría tener operaciones personalizadas sobre los datos según la fuente de los datos. En las bases de datos clínicas, existe una gran variedad de conceptos clínicos que deben armonizarse utilizando vocabularios estándar. Aunque pudimos automatizar partes de esta etapa, aún requerimos la validación manual por parte de un profesional de la salud especializado para garantizar que todos los datos mapeados sean correctos.

Finalmente, la etapa de carga inserta los datos procesados en la base de datos de destino, a la que luego se puede acceder utilizando herramientas analíticas. Las bases

de datos clínicos se rellenan con datos pseudoanónimos, lo que permite realizar estudios clínicos sin violar los derechos de privacidad de los pacientes. Además, cuando los datos se migran a un esquema de datos estándar, los datos originales terminan siendo validados y se pueden encontrar incoherencias en la base de datos de origen. Esto es posible gracias a los mecanismos de calidad que se crearon en el pipeline, los cuales se encargan de verificar si los datos cargados respetan los atributos de la regla para cada concepto estándar.

A.2.2 El conjunto de herramientas del migrador de cohortes

La metodología propuesta se implementó primero en Python utilizando las adaptaciones de las herramientas descritas anteriormente y está disponible públicamente, bajo la licencia MIT, en <https://bioinformatics-ua.github.io/CMTToolkit/>. Esta metodología incluye las etapas de las operaciones ETL, es decir, el flujo de trabajo desde los datos sin procesar de la cohorte hasta la base de datos OMOP CDM se divide en las tres etapas ETL.

La implementación de algunos componentes planteó algunos desafíos debido a la sensibilidad de los datos del caso de uso propuesto. Cuando se trata de datos médicos, se requiere un conocimiento profundo de la fuente de datos, para poder realizar la armonización correctamente. Otra tarea desafiante fueron las operaciones personalizadas en los datos sin procesar de cada cohorte, es decir, cuando se recopilaban los datos, no siguieron ninguna estrategia estándar. Esta falta de interoperabilidad a la hora de registrar los datos complicaba la implementación del flujo de trabajo de migración.

Una ventaja de utilizar este flujo de trabajo es la calidad de los datos. Al final del procedimiento ETL, el sistema pudo proporcionar un informe de migración que incluye información estadística sobre los datos migrados, incluidas las incoherencias en los datos de origen. Esta información fue útil para que los propietarios de la cohorte pudieran corregir estos problemas, ya que los valores se recolectaron manualmente durante las visitas de seguimiento del paciente.

A.2.3 Blcenter y Blcenter-AD

Blcenter es una herramienta ETL basada en la web que cubre algunas limitaciones y problemas que se encuentran actualmente en la creación y administración de tareas

ETL en entornos de múltiples instituciones. Esta herramienta simplifica la descripción de los flujos de trabajo de ETL y ayuda a los usuarios sin experiencia técnica a comprender dichos flujos de trabajo a través de una interfaz gráfica intuitiva. BICenter replica las funciones de Kettle en un navegador HTML5 y simplifica algunos de los procedimientos en Kettle que pueden requerir un conocimiento técnico profundo de esta herramienta.

El uso de BICenter aprovechó la metodología propuesta para nuevas posibilidades, lo que llevó a un entorno colaborativo y multiinstitucional. BICenter se desarrolló inicialmente para tener diferentes roles asignados a diferentes instituciones. Esta estrategia permite el uso de una sola instalación para definir los pipelines de migración de todas las cohortes con la posibilidad de dividir a los usuarios por instituciones o cohortes. Por lo tanto, los mecanismos RBAC existentes mantienen conjuntos de permisos para acceder a las diferentes funciones de la aplicación. Por ejemplo, permite a usuarios específicos visualizar los resultados de cada transformación o escribirlos en la base de datos de destino.

BICenter-AD se propuso como una extensión de BICenter aplicada a conjuntos de datos de enfermedades de Alzheimer. Esta herramienta proporcionaba un entorno colaborativo centralizado en el Editor de tareas ETL. Este espacio de trabajo permite la definición de canalizaciones ETL con todos los componentes necesarios para armonizar las cohortes de enfermedades de Alzheimer. Por lo tanto, los usuarios con permiso para editar una tarea ETL pueden trabajar en colaboración en el mismo espacio de trabajo. Aunque estas herramientas no crean sesiones de trabajo en tiempo real, estos sistemas proporcionan un entorno fácil de usar donde varios usuarios pueden trabajar en colaboración.

A.3 De texto no estructurado a registros basados en ontologías

En la sección anterior, propusimos diferentes estrategias para migrar datos heterogéneos a un esquema de datos común. Siguiendo esta dirección de investigación, identificamos algunas lagunas en estos procedimientos ETL con respecto a la información médica no estructurada.

A.3.1 Extraer y armonizar las menciones de medicamentos

La primera propuesta para extraer información médica de datos no estructurados fue un flujo de trabajo de dos etapas, denominado DrAC. La primera etapa del flujo de trabajo extrae las prescripciones presentes en las notas clínicas de los pacientes, mientras que la segunda etapa armoniza la información extraída en su definición estándar y almacena la información resultante en un esquema de base de datos común, a saber, OMOP CDM.

El sistema inicialmente recibe notas clínicas como entrada, lee su contenido y lo almacena de acuerdo a una estructura fija. Este lector se implementa utilizando el patrón de programación de fábrica, por lo que se debe implementar un nuevo lector de conjuntos de datos siempre que se vaya a utilizar un nuevo conjunto de datos de notas clínicas. Después de leer las notas clínicas, se utiliza un anotador para identificar las entidades de medicación en cada nota, y las anotaciones resultantes se almacenan y procesan posteriormente. La herramienta utilizada para esto fue Neji, un marco flexible y modular para el procesamiento y anotación de texto [139].

Para configurar a Neji como anotador de medicamentos, primero extrajimos tres terminologías médicas relacionadas con los medicamentos de UMLS Metathesaurus [148]: RxNorm, DrugBank y AOD. Sin embargo, estas terminologías cubren muchos tipos y grupos semánticos, por lo tanto, para reducir el alcance de los diccionarios, los filtramos conservando solo las entradas del grupo semántico “Chemicals & Drugs”. Los diccionarios resultantes se importaron a Neji y se configuró un servicio de anotación de Neji para la extracción de menciones de medicamentos en el texto clínico. Después de pasar todas las notas clínicas por el lector del sistema, se utilizó el servicio web de Neji para anotar las entidades de medicación en cada nota y se almacenaron las anotaciones resultantes.

Una vez que se completa el proceso de extracción de información, toda la información extraída se almacena en una matriz estructurada por paciente y medicamento, donde cada celda contiene información sobre un medicamento mencionado (potencia, dosis y vía). La razón para almacenar los datos extraídos en este formato particular radica en el hecho de que la estructura resultante es similar a la que ya se usa en los estudios de cohortes, lo que simplifica enormemente el proceso de migración a una base de datos OMOP CDM.

A.3.2 Normalización de conceptos en varios idiomas

Uno de los problemas de los procedimientos ETL es el esfuerzo necesario para mapear los conceptos originales en sus definiciones estándar. Si bien varias soluciones de mapeo automático pueden ayudar en esta tarea, su complejidad aumenta cuando se trata de bases de datos en varios idiomas, lo que lleva a un esfuerzo manual significativo en la traducción y el mapeo. En esta sección, proponemos una estrategia que combina la minería de texto con técnicas de detección de lenguaje, con el objetivo de optimizar estas canalizaciones de migración. Este sistema fue diseñado para integrarse en flujos de trabajo de migración ya existentes, como se propuso anteriormente.

Nuestra propuesta utiliza dos herramientas de código abierto para: i) proporcionar una interfaz de usuario para validar las asignaciones; y ii) proporcionar una plataforma colaborativa web para administrar las ontologías utilizadas en nuestra propuesta. Utilizamos Usagi¹ como interfaz para validar las asignaciones. Proporciona mapeos sugerentes basados en la similitud de palabras a través de una interfaz simple pero intuitiva, en la que los equipos no técnicos pueden validar los mapeos. La sugerencia de Usagi solo compara el concepto con el vocabulario estándar, lo que genera muchas sugerencias incorrectas que terminan siendo modificadas manualmente. Sin embargo, la interfaz de usuario es intuitiva y reutilizable para el enfoque propuesto y actualmente se usa en varios flujos de trabajo de migración, incluso en las metodologías propuestas en la sección anterior.

A.3.3 Extracción de información de historia familiar

A pesar de los esfuerzos por estructurar todos los datos clínicos del paciente, los informes clínicos y las notas contienen información esencial sobre el historial de salud de la familia, que puede ser de gran relevancia para el diagnóstico y pronóstico. En esta sección, proponemos dos metodologías para unificar este conocimiento y extraer información de historia familiar de las notas clínicas usando técnicas basadas en reglas en NLP. Con estos métodos, pretendemos recopilar las informaciones de los miembros de la familia mencionadas en el texto, así como las asociaciones con enfermedades y estado de vida. La implementación de estos métodos resultó en una herramienta denominada PatientFM.

¹<https://github.com/OHDSI/usagi>

En general, los sistemas propuestos mejoran la información presente en las bases de datos observacionales que usan el esquema de datos OMOP CDM. El trabajo de Liu *et al.* [155] es muy útil para recuperar notas clínicas del repositorio en función de las condiciones definidas en una cohorte. Park *et al.* [156] usó la base de datos OMOP CDM para extraer las notas de un esquema estándar en texto libre para luego anotarlas. Aunque ambos trabajos se centraron en usar NLP para aprovechar la información de las bases de datos de OMOP CDM, ninguno de ellos integró los datos resultantes con los datos ya existentes y extraídos del modelo relacional del sistema EHR.

Estas herramientas promueven nuevas estrategias para anotar automáticamente grandes cantidades de datos EHR. También creamos nuevas oportunidades relacionadas principalmente con la exploración de EHR fomentando el descubrimiento de nuevas relaciones y vías entre enfermedades y fenotipos parentales.

A.4 Perfiles de bases de datos escalables para estudios multicéntricos

Uno de los desafíos al reutilizar las bases de datos de salud para la investigación es la correcta selección de las fuentes de datos. Este es un problema complejo ya que requiere estrategias para caracterizar las fuentes de datos sin revelar su contenido y plataformas para difundir las características de las bases de datos [157, 158]. Para el tema de caracterización de bases de datos, ya existen algunas pautas a la hora de tratar con este tipo de datos. Según las políticas del proyecto o de la institución, los propietarios de los datos pueden compartir información agregada sobre sus datos. Esto puede proporcionar un resumen de los pacientes en las bases de datos. También se pueden proporcionar otras características, a saber, políticas de gobierno de datos y datos de contacto. Estas pautas de resumen no son estándar y pueden diferir según el contexto. Por ejemplo, una comunidad centrada en el estudio de la enfermedad de Alzheimer tendría conjuntos de datos con características diferentes en comparación con un dominio más genérico [70].

La creación de perfiles de bases de datos (o huella digital) es la acción de representar una base de datos utilizando un conjunto de características que combinadas pueden crear una concepción singular de la base de datos. La definición de estas características plantea algunas cuestiones que varían según el alcance del proyecto. Si bien estos

problemas tienen soluciones complejas, proponemos una estrategia diferente para ayudar al descubrimiento de bases de datos médicas. Su objetivo es proporcionar suficiente información sobre las bases de datos, que pueda caracterizarlas a un nivel más profundo, sin compartir información sensible.

A.4.1 Marco para crear perfiles de bases de datos

El marco MONTRA 2 se desarrolló como una solución para permitir el intercambio de datos biomédicos mediante la creación de entornos basados en la web con fines de investigación. El catálogo de la base de datos puede considerarse una de las funciones principales de MONTRA 2. En este catálogo se representa cada base de datos a través del concepto de huella dactilar, como ya se ha descrito. Por lo tanto, los propietarios de los datos pueden definir la estructura del catálogo que mejor se adapte a sus necesidades en ese ámbito, y el sistema genera el catálogo web basado en ese archivo. La estructura del esqueleto es flexible y contiene campos (preguntas) que deben completar los propietarios de los datos. Se pueden agregar varias preguntas en un “QuestionSet”, creando una representación de datos jerárquica. Cada pregunta puede almacenar diferentes tipos de datos, por ejemplo, fechas, números, cadenas, valores de opción múltiple, ubicación geográfica, entre otros. Estos campos, que representan los metadatos sobre las bases de datos de salud del catálogo, se utilizan para la búsqueda de texto libre, la búsqueda avanzada, la comparación de conjuntos de datos y otras funciones del catálogo.

MONTRA 2 se implementó para apoyar también la creación de un entorno para integrar distintas herramientas en una plataforma centralizada. El objetivo de este paradigma era proporcionar a los investigadores un lugar de trabajo con todas las herramientas necesarias para: i) comparar e identificar las bases de datos de interés para los estudios clínicos; ii) agilizar un estudio sobre la red; y iii) recuperar los resultados y agregarlos. Todas estas herramientas están protegidas por un mecanismo de inicio de sesión único federado con verificación de perfil. MONTRA 2 se utiliza actualmente para apoyar otros proyectos diferentes. El sistema tiene tres instancias en producción, para soportar diferentes plataformas, a saber, el Portal EHDEN, el Catálogo EMIF y el Portal MSDA.

A.4.2 Recomendar bases de datos de salud

Los investigadores necesitan analizar periódicamente las actualizaciones en las bases de datos disponibles, buscando nuevos conjuntos de datos de interés. El filtrado manual es necesario porque se pueden realizar nuevos estudios siguiendo diferentes prácticas, generando conjuntos de datos no relacionados que se centran en la misma enfermedad. Con el objetivo de simplificar la identificación correcta de nuevas fuentes de datos de interés, propusimos una solución para sugerir conjuntos de datos o publicaciones similares a los usuarios involucrados en un estudio clínico, aumentando la información de interés. Esta solución recomienda nuevas fuentes de datos basadas en perfiles de usuario, manteniendo a los investigadores actualizados sobre estudios similares realizados con datos de MONTRA 2.

El filtrado colaborativo en los sistemas de recomendación produce sugerencias específicas para los usuarios, según patrones de uso o calificaciones. Estas sugerencias se pueden realizar después de recopilar las preferencias de varios usuarios que se consideran con intereses similares [192]. Por otro lado, un sistema de recomendación basado en contenido intenta dar una sugerencia basada en la calificación del usuario y en el contenido del artículo y su similitud. Esto se calcula en función de las características más relevantes [195]. El sistema de recomendación propuesto combina las dos técnicas presentadas para llenar los vacíos de cada metodología aislada. El filtrado colaborativo puede detectar perfiles de usuarios similares y proporcionar recomendaciones cuando la estructura de las fuentes de datos varía significativamente. Por otro lado, la recuperación basada en el contexto puede proporcionar mejores sugerencias, basándose únicamente en la similitud de las fuentes de datos. Por lo tanto, aplicamos métricas para medir primero cada enfoque y luego combinar ambos.

A.4.3 Explorar bases de datos distribuidas a nivel de paciente

La metodología propuesta para agilizar la ejecución de estudios multicéntricos se basa en MONTRA 2. Para lograr esto, desarrollamos una herramienta adicional que se integró en MONTRA 2 como complemento. Tiene como objetivo simplificar la ejecución de los estudios de salud, así como centralizar y coordinar las operaciones entre todas las entidades involucradas. El sistema propuesto, designado como Study Manager, adoptó las mismas tecnologías utilizadas en MONTRA 2, es decir, Django

en su núcleo. Para simplificar la integración entre sistemas, esta herramienta se implementó para cumplir con MONTRA SDK, siguiendo un patrón de software MVC. Este patrón segrega la lógica de la aplicación en tres elementos principales: i) el modelo, responsable de manejar el almacenamiento de datos; ii) la vista, que genera la representación de datos para el cliente; y iii) el controlador, que contiene la capa empresarial.

Con todas las soluciones propuestas en los apartados anteriores, incluido el marco MONTRA 2, el proceso de realización de estudios médicos multicéntricos es actualmente una realidad. Los investigadores de ciencias de la salud y de la vida han identificado varias oportunidades para compartir datos. Estas oportunidades solo se pueden lograr si los investigadores pueden compartir datos entre ellos. La estrategia propuesta en este trabajo los empodera con conjuntos de datos más grandes para cada estudio, lo que aumenta el impacto de sus hallazgos. Sin embargo, con esta idea se plantearon diferentes cuestiones gubernamentales. Por lo tanto, las estrategias propuestas tienen como objetivo facilitar la exploración de bases de datos a nivel de paciente, minimizando el riesgo de violar la privacidad del paciente.

A.5 Conclusiones

Enriquecer las etapas de extracción de información en los sistemas de apoyo a la decisión clínica es un tema de investigación que puede abordarse desde diferentes puntos de vista. En este trabajo intentamos enriquecer estas etapas comenzando por trabajar sobre las bases de los sistemas de apoyo a la decisión clínica. Reconocimos que para aumentar la calidad de los tratamientos, los investigadores deben estudiar el impacto de los nuevos medicamentos o la eficiencia de los tratamientos actuales. Estos hallazgos pueden originar nuevos protocolos de tratamiento que pueden integrarse en los sistemas de apoyo a la decisión de las instituciones de salud. Por lo tanto, en este trabajo, nos enfocamos en crear metodologías y herramientas para ayudar a los investigadores médicos a realizar hallazgos más impactantes, para mejorar la fuente de estos sistemas.

Comenzamos especificando el alcance de este trabajo, en función de los formatos de datos biomédicos que podríamos utilizar. Motivados por el proyecto EHDEN, centramos este trabajo en los datos relacionales de EHR, que intentamos complementar con datos extraídos de narrativas médicas. Luego, en la etapa posterior, después de definir estrategias para tener una red interoperable de fuentes de datos, propusimos

soluciones para apoyar la investigación utilizando estas fuentes de datos. En resumen, presentamos varias soluciones de software para integrar datos biomédicos y el producto final es una plataforma que facilita la exploración de esta información a través de bases de datos.

La primera hipótesis abordó la falta de interoperabilidad entre las bases de datos de salud. Sin embargo, como encontramos durante este trabajo, el problema no fue la falta de soluciones estándar para interconectar estas bases de datos. En cambio, el problema fue el esfuerzo requerido para adoptar uno de estos estándares. Para responder a este problema, propusimos soluciones para simplificar la migración de datos EHR a uno de los esquemas de datos estándar que se utilizan actualmente en estudios médicos. Validamos estas soluciones utilizando cohortes heterogéneas de datos de pacientes que padecen la enfermedad de Alzheimer. La interoperabilidad se aseguró mediante la conversión de fuentes de datos al esquema OMOP CDM.

La segunda hipótesis se refería a enriquecer la información almacenada en las bases de datos, utilizando datos no estructurados presentes en las narrativas clínicas. Para ello propusimos una solución capaz de extraer conceptos médicos y almacenarlos en una base de datos OMOP CDM. Parte de esta solución está respaldada por el trabajo realizado para responder a la primera hipótesis. Validamos las estrategias NLP propuestas utilizando desafíos científicos, concretamente organizados por la organización n2c2.

Finalmente, la tercera hipótesis se centró en encontrar las bases de datos de salud más adecuadas para estudios de investigación específicos. Para responder a esta pregunta, hemos colaborado durante este programa de doctorado con los socios de EHDEN con el objetivo de proponer y ajustar una solución basada en necesidades reales. El resultado fue un marco flexible capaz de ampliarse para admitir herramientas complementarias. Este trabajo fue validado en el contexto del proyecto EHDEN. Además, también reemplazó tecnologías antiguas que han respaldado el proyecto EMIF en el pasado. Esta herramienta ha sido validada con miles de usuarios, con un gran impacto en entornos de la vida real.

Sinopsis in Galician

B.1 Introducción

A continua demanda de mellores diagnósticos e tratamentos de saúde motivou moitos estudos de investigación clínica, como estudos observacionais e ensaios clínicos. Nos ensaios clínicos, os pacientes divídense comunmente en dous ou máis grupos (*p. ex.* activo e placebo), para estudar a efectividade do tratamento para unha condición clínica particular[1]. Neste caso, hai intervención directa cos pacientes, *p.ex.*, administración dun fármaco ou procedementos terapéuticos. Con todo, este enfoque non sempre é o máis apropiado, *p. ex.* abordar preguntas de investigación en cirurxía plástica a través de ensaios controlados aleatorios a miúdo está suxeito a restricións éticas[2]. Nos estudos observacionais, os investigadores non realizan ningunha intervención activa cos pacientes e a exposición prodúcese de forma natural ou a través doutros factores. Aquí, os investigadores médicos límitanse a documentar a relación entre a exposición e o resultado do estudo[1].

Os estudos observacionais seguen diferentes estratexias e establécense definindo un conxunto de criterios de inclusión e exclusión para os suxeitos involucrados, así como varias características que se identifican e observan ao longo do tempo[3]. Algunhas iniciativas reutilizan os datos xa recompilados en institucións médicas para realizar estudos observacionais. Esta práctica aforra tempo e permite a verificación previa do número de suxeitos antes de iniciar a análise[4]. Con todo, en enfermidades específicas, é necesario recompilar información sobre os suxeitos seleccionados. Nestas situacións, os datos rexístranse segundo as pautas do estudo e pódense usar diferentes solucións para almacenar os datos, por exemplo, o EHR system[5] institucional.

A dependencia dos equipos técnicos é unha gran barreira cando existe a necesidade de extraer datos para a súa análise. Moitas outras cuestións éticas e técnicas xorden cando se desexa combinar conxuntos de datos de distintas organizacións. Este é o

caso dos estudos multicéntricos que pretenden aumentar o tamaño da poboación, o poder da evidencia estatística e, por tanto, o impacto do estudo[6].

A integración de múltiples fontes de datos non é só un problema tecnolóxico. Hai algunhas ferramentas ETL capaces de realizar esta tarefa utilizando grandes cantidades de datos. O problema non técnico desta agregación é o dominio dos datos, é dicir, identificar os conceptos nos datos e combinar correctamente a información asociada a eles que se extraeu de múltiples fontes. As bases de datos de saúde pertencen a un dos dominios nos que isto é un problema preocupante, debido á variedade de conceptos para representar procedementos e termos médicos similares. Resolver estes problemas é útil para optimizar os estudos, pero compartir datos a nivel de paciente aínda expón algúns problemas de privacidade, debido a requisitos legais, éticos e regulamentarios[8]. Os datos dos pacientes son moi sensibles e a interrupción desta privacidade pode ter consecuencias dramáticas para as persoas, os provedores de atención médica e os subgrupos dentro da sociedade[9]. Ademais, a lexislación pode ser diferente en cada país, o que dificulta definir un protocolo que se axuste a todas as institucións involucradas[10]. Este é outro desafío, que require atopar unha solución que permita a análise de múltiples fontes de datos, ou partes destas, sen expoñer datos confidenciais.

O impacto potencial dos estudos multicéntricos motivou aos investigadores para buscar solucións máis sólidas e reutilizables para agregar coñecementos a partir de conxuntos de datos de saúde distribuídos. Establecéronse organizacións e metodoloxías para explorar bases de datos clínicas mediante a reutilización de datos existentes[4]. Un destes esforzos ten como obxectivo crear unha estratexia para reutilizar as bases de datos EHR utilizando un esquema homoxéneo, co fin de facilitar a interoperabilidade entre as bases de datos. Actualmente, esta integración é posible mediante o uso de marcos de código aberto que axudan a apoiar todo o proceso[7].

Os obxectivos de investigación deste traballo pódense parafrasear en varias preguntas enfocadas en abordar problemas específicos. Recoñecemos que os estudos médicos multicéntricos poden expor problemas adicionais que non serán considerados neste traballo, principalmente debido aos tipos de datos utilizados nestes. Por exemplo, estudos de investigación baseados en biomarcadores que se correlacionan con información de DNA, ou estudos centrados principalmente no uso de imaxes médicas. Por tanto, para lograr un escenario capaz de apoiar estudos de saúde distribuídos utilizando múltiples fontes de datos de distintas institucións, limitamos o alcance ás

bases de datos EHR. Este obxectivo lograrase respondendo ás seguintes preguntas de investigación:

1. *¿Como executar unha consulta de base de datos sobre unha rede de bases de datos de saúde heteroxéneas?* A falta de interoperabilidade é o principal problema neste escenario. Un ecosistema con bases de datos heteroxéneas pode non compartir o mesmo esquema de datos, o que invalida o uso compartido de consultas. Ademais deste problema, o dominio da saúde contén unha gran cantidade de conceptos médicos, que poden diferir entre institucións, a nivel nacional ou internacional. Requírese unha metodoloxía para harmonizar as devanditas bases de datos, que as converta nun formato interoperable. Este proceso pode ter diferentes etapas e compoñentes, e algúns deles serán automatizados co fin de reducir o custo e tempo de execución deste procedemento. Isto pode resultar nunha nova solución de software.
2. *¿Como seleccionar as bases de datos de saúde máis adecuadas para un estudo de investigación específico?* Esta pregunta pode abordarse desde diferentes perspectivas. Con todo, ao correlacionarlo cos enunciados anteriores, identificamos que os principais problemas que enfrontan os investigadores médicos son: i) o descubrimento de bases de datos de interese; e ii) o acceso ás devanditas bases de datos sen violar políticas de privacidade e normas éticas. A solución a estes problemas é demasiado complexa para ser resolta por unha soa aplicación de software. Para responder a esta pregunta, é necesario crear un ecosistema de ferramentas e metodoloxías, no que os propietarios dos datos poidan sentirse seguros ao compartir características sobre as súas bases de datos, mentres que os investigadores poidan ter suficiente información para seleccionar as bases de datos que mellor se adapten ás súas necesidades. Por tanto, a solución proposta para este problema será un portal que integre: i) un catálogo web de características da base de datos; ii) ferramentas para visualizar e comparar estas características; e iii) ferramentas para orquestrar estudos distribuídos.

B.2 Tradución semiautomática de fontes de datos a un esquema común

A falta de flexibilidade na colaboración dos usuarios durante o deseño e a definición das canalizacións ETL é un problema para algúns dominios de aplicación. Por

exemplo, no escenario médico, cando os datos clínicos deben harmonizarse nun esquema de datos común, isto require a colaboración entre os equipos técnicos e os investigadores médicos[73]. Esta colaboración requírese en diferentes etapas: i) deseño; ii) implementación; e iii) validación. En cada etapa, hai algúns desafíos que abordamos neste traballo.

B.2.1 Metodoloxía para a harmonización de cohortes

Propoñemos unha metodoloxía baseada nos principios ETL. Na etapa de extracción, os datos de orixe seleccionados lense extraéndoo dunha ou varias fontes de datos. O obxectivo principal desta etapa é obter os datos dos sistemas de orixe sen interferir co seu rendemento habitual. Nas bases de datos de saúde, esta é unha tarefa delicada porque o EHR non se pode sobrecargar debido ao procedemento de extracción de datos. Con todo, en estudos clínicos, a cantidade de datos non é suficiente para colapsar os sistemas durante esta etapa. Ademais, os estudos clínicos exportáronse a un formato tabular, que non require interacción directa co sistema utilizado para recompilar os datos do paciente.

A etapa de transformación é o compoñente máis complexo de todas as etapas. Esta etapa require o mapeo da base de datos de orixe no esquema de destino, así como a harmonización do contido. Para unha fonte de datos, este procedemento require un mapeo completo, o que leva moito tempo e require entidades especializadas para validar os mapeos. A harmonización de contido podería ter operacións personalizadas sobre os datos segundo a fonte dos datos. Nas bases de datos clínicas, existe unha gran variedade de conceptos clínicos que deben harmonizarse utilizando vocabularios estándar. Aínda que podemos automatizar partes desta etapa, aínda requirimos a validación manual por parte dun profesional da saúde especializado para garantir que todos os datos mapeados sexan correctos.

Finalmente, a etapa de carga insere os datos procesados na base de datos de destino, á que logo se pode acceder utilizando ferramentas analíticas. As bases de datos clínicas énchense con datos pseudoanónimos, o que permite realizar estudos clínicos sen violar os dereitos de privacidade dos pacientes. Ademais, cando os datos se migran a un esquema de datos estándar, os datos orixinais terminan sendo validados e pódense atopar incoherencias na base de datos de orixe. Isto é posible grazas aos mecanismos de calidade que se crearon nas distintas etapas, os cales se encargan

de verificar se os datos cargados respectan os atributos da regra para cada concepto estándar.

B.2.2 O conxunto de ferramentas do migrador de cohortes

A metodoloxía proposta implementouse primeiro en Python utilizando as adaptacións das ferramentas descritas anteriormente e está dispoñible publicamente, baixo a licenza MIT, en <https://bioinformatics-ua.github.io/CMToolkit/>. Esta metodoloxía inclúe as etapas das operacións ETL, é dicir, o fluxo de traballo desde os datos sen procesar da cohorte ata a base de datos OMOP CDM divídese nas tres etapas ETL.

A implementación dalgúns compoñentes expuxo algúns desafíos debido á sensibilidade dos datos do caso de uso proposto. Cando se trata de datos médicos, requírese un coñecemento profundo da fonte de datos, para poder realizar a harmonización correctamente. Outra tarefa desafiante foron as operacións personalizadas nos datos sen procesar de cada cohorte, é dicir, cando se recompilaron os datos, non seguiron ningunha estratexia estándar. Esta falta de interoperabilidade á hora de rexistrar os datos complicaba a implementación do fluxo de traballo de migración.

Unha vantaxe de utilizar este fluxo de traballo é a calidade dos datos. Ao final do procedemento ETL, o sistema puido proporcionar un informe de migración que inclúe información estatística sobre os datos migrados, incluídas as incoherencias nos datos de orixe. Esta información foi útil para que os propietarios da cohorte puidesen corrixir estes problemas, xa que os valores colleitáronse manualmente durante as visitas de seguimento do paciente.

B.2.3 BICenter e BICenter-AD

BICenter é unha ferramenta ETL baseada na web que cobre algunhas limitacións e problemas que se atopan actualmente na creación e administración de tarefas ETL en contornas de múltiples institucións. Esta ferramenta simplifica a descrición dos fluxos de traballo de ETL e axuda aos usuarios sen experiencia técnica a comprender os devanditos fluxos de traballo a través dunha interface gráfica intuitiva. BICenter replica as funcións de Kettle nun navegador HTML5 e simplifica algúns dos procedementos en Kettle que poden requirir un coñecemento técnico profundo desta ferramenta.

O uso de BCenter aproveitou a metodoloxía proposta para novas posibilidades, o que levou a unha contorna colaborativa e multi-institucional. BCenter desenvolveuse inicialmente para ter diferentes roles asignados a diferentes institucións. Esta estratexia permite o uso dunha soa instalación para definir os pipelines de migración de todas as cohortes coa posibilidade de dividir aos usuarios por institucións ou cohortes. Por tanto, os mecanismos RBAC existentes manteñen conxuntos de permisos para acceder ás diferentes funcións da aplicación. Por exemplo, permite a usuarios específicos visualizar os resultados de cada transformación ou escribilos na base de datos de destino.

BCenter-AD propúxose como unha extensión de BCenter aplicada a conxuntos de datos de enfermidades de Alzheimer. Esta ferramenta proporcionaba unha contorna colaborativa centralizada no Editor de tarefas ETL. Este espazo de traballo permite a definición de canalizacións ETL con todos os compoñentes necesarios para harmonizar as cohortes de enfermidades de Alzheimer. Por tanto, os usuarios con permiso para editar unha tarefa ETL poden traballar en colaboración no mesmo espazo de traballo. Aínda que estas ferramentas non crean sesións de traballo en tempo real, estes sistemas proporcionan unha contorna fácil de usar onde varios usuarios poden traballar en colaboración.

B.3 De texto non estruturado a rexistros baseados en ontoloxías

Na sección anterior, propuxemos diferentes estratexias para migrar datos heteroxéneos a un esquema de datos común. Seguindo esta dirección de investigación, identificamos algunhas lagoas nestes procedementos ETL con respecto á información médica non estruturada.

B.3.1 Extraer e harmonizar as mencións de medicamentos

A primeira proposta para extraer información médica de datos non estruturados foi un fluxo de traballo de dúas etapas, denominado DrAC. A primeira etapa do fluxo de traballo extrae as prescricións presentes nas notas clínicas dos pacientes, mentres que a segunda etapa harmoniza a información extraída na súa definición estándar e almacena a información resultante nun esquema de base de datos común, a saber, OMOP CDM.

O sistema inicialmente recibe notas clínicas como entrada, le o seu contido e almacénalo de acordo a unha estrutura fixa. Este lector impleméntase utilizando o patrón de programación de fábrica, polo que se debe implementar un novo lector de conxuntos de datos sempre que se vaia a utilizar un novo conxunto de datos de notas clínicas. Despois de ler as notas clínicas, utilízase un anotador para identificar as entidades de medicación en cada nota, e as anotacións resultantes almacénanse e procesan posteriormente. A ferramenta utilizada para isto foi Neji, un marco flexible e modular para o procesamento e anotación de texto[matogueiras2018configurable].

Para configurar a Neji como anotador de medicamentos, primeiro extraemos tres terminoloxías médicas relacionadas cos medicamentos de UMLS Metathesaurus[148]: RxNorm, DrugBank e AOD. Con todo, estas terminoloxías cobren moitos tipos e grupos semánticos, por tanto, para reducir o alcance dos dicionarios, filtrámoslos conservando só as entradas do grupo semántico “Chemicals & Drugs”. Os dicionarios resultantes importáronse a Neji e configúrouse un servizo de anotación de Neji para a extracción de mencións de medicamentos no texto clínico. Despois de pasar todas as notas clínicas polo lector do sistema, utilizouse o servizo web de Neji para anotar as entidades de medicación en cada nota e almacenáronse as anotacións resultantes.

Unha vez que se completa o proceso de extracción de información, toda a información extraída almacénase nunha matriz estruturada por paciente e medicamento, onde cada cela contén información sobre un medicamento mencionado (potencia, dose e vía). A razón para almacenar os datos extraídos neste formato particular radica no feito de que a estrutura resultante é similar á que xa se usa nos estudos de cohortes, o que simplifica enormemente o proceso de migración a unha base de datos OMOP CDM.

B.3.2 Normalización de conceptos en varios idiomas

Uno dos problemas dos procedementos ETL é o esforzo necesario para mapear os conceptos orixinais nas súas definicións estándar. Aínda que varias solucións de mapeo automático poden axudar nesta tarefa, a súa complexidade aumenta cando se trata de bases de datos en varios idiomas, o que leva a un esforzo manual significativo na tradución e o mapeo. Nesta sección, propoñemos unha estratexia que combina a minería de texto con técnicas de detección de linguaxe, co obxectivo de optimizar estas canalizacións de migración. Este sistema foi deseñado para integrarse en fluxos de traballo de migración xa existentes, como se propuxo anteriormente.

A nosa proposta utiliza dúas ferramentas de código aberto para: i) proporcionar unha interface de usuario para validar as asignacións; e ii) proporcionar unha plataforma colaborativa web para administrar as ontoloxías utilizadas na nosa proposta. Utilizamos Usagi¹ como interface para validar as asignacións. Proporciona mapeos suxestivos baseados na similitude de palabras a través dunha interface simple pero intuitiva, na que os equipos non técnicos poden validar os mapeos. A suxerencia de Usagi só compara o concepto co vocabulario estándar, o que xera moitas suxerencias incorrectas que terminan sendo modificadas manualmente. Con todo, a interface de usuario é intuitiva e reutilizable para o enfoque proposto e actualmente úsase en varios fluxos de traballo de migración, mesmo nas metodoloxías propostas na sección anterior.

B.3.3 Extracción de información de historia familiar

A pesar dos esforzos por estruturar todos os datos clínicos do paciente, os informes clínicos e as notas conteñen información esencial sobre o historial de saúde da familia, que pode ser de gran relevancia para o diagnóstico e prognóstico. Nesta sección, propoñemos dúas metodoloxías para unificar este coñecemento e extraer información de historia familiar das notas clínicas usando técnicas baseadas en regras en NLP. Con estes métodos, pretendemos recompilar as informacións dos membros da familia mencionadas no texto, así como as asociacións con enfermidades e estado de vida. A implementación destes métodos resultou nunha ferramenta denominada PatientFM.

En xeral, os sistemas propostos melloran a información presente nas bases de datos observacionais que usan o esquema de datos OMOP CDM. O traballo de Liuet *al.* [155] é moi útil para recuperar notas clínicas do repositorio en función das condicións definidas nunha cohorte. Parket *al.* [156] usou a base de datos OMOP CDM para extraer as notas dun esquema estándar en texto libre para logo anotalas. Aínda que ambos os traballos centráronse en usar NLP para aproveitar a información das bases de datos de OMOP CDM, ningún deles integrou os datos resultantes cos datos xa existentes e extraídos do modelo relacional do sistema EHR.

Estas ferramentas promoven novas estratexias para anotar automaticamente grandes cantidades de datos EHR. Tamén creamos novas oportunidades relacionadas princi-

¹<https://github.com/OHDSI/usagi>

palmente coa exploración de EHR fomentando o descubrimento de novas relacións e vías entre enfermidades e fenotipos parentais.

B.4 Perfís de bases de datos escalables para estudos multicéntricos

Un dos desafíos ao reutilizar as bases de datos de saúde para a investigación é a correcta selección das fontes de datos. Este é un problema complexo xa que require estratexias para caracterizar as fontes de datos sen revelar o seu contido e plataformas para difundir as características das bases de datos[157, 158]. Para o tema de caracterización de bases de datos, xa existen algunhas pautas á hora de tratar con este tipo de datos. Segundo as políticas do proxecto ou da institución, os propietarios dos datos poden compartir información agregada sobre os seus datos. Isto pode proporcionar un resumo dos pacientes nas bases de datos. Tamén se poden proporcionar outras características, a saber, políticas de goberno de datos e datos de contacto. Estas pautas de resumo non son estándar e poden diferir segundo o contexto. Por exemplo, unha comunidade centrada no estudo da enfermidade de Alzheimer tería conxuntos de datos con características diferentes en comparación cun dominio máis xenérico[70].

A creación de perfís de bases de datos (ou pegada dixital) é a acción de representar unha base de datos utilizando un conxunto de características que combinadas poden crear unha concepción singular da base de datos. A definición destas características expón algunhas cuestións que varían segundo o alcance do proxecto. Aínda que estes problemas teñen solucións complexas, propoñemos unha estratexia diferente para axudar ao descubrimento de bases de datos médicas. O seu obxectivo é proporcionar suficiente información sobre as bases de datos, que poida caracterizalas a un nivel máis profundo, sen compartir información sensible.

B.4.1 Marco para crear perfís de bases de datos

O marco MONTRA2 desenvolveuse como unha solución para permitir o intercambio de datos biomédicos mediante a creación de contornas baseadas na web con fins de investigación. O catálogo da base de datos pode considerarse unha das funcións principais de MONTRA2. Neste catálogo represéntase cada base de datos a través do concepto de pegada dactilar, como xa se describiu. Por tanto, os propietarios dos

datos poden definir a estrutura do catálogo que mellor se adapte ás súas necesidades nese ámbito, e o sistema xera o catálogo web baseado nese arquivo. A estrutura do esqueleto é flexible e contén campos (preguntas) que deben completar os propietarios dos datos. Pódense agregar varias preguntas nun “QuestionSet”, creando unha representación de datos xerárquica. Cada pregunta pode almacenar diferentes tipos de datos, por exemplo, datas, números, cadeas, valores de opción múltiple, localización xeográfica, entre outros. Estes campos, que representan os metadatos sobre as bases de datos de saúde do catálogo, utilízanse para a procura de texto libre, a procura avanzada, a comparación de conxuntos de datos e outras funcións do catálogo.

MONTRA2 implementouse para apoiar tamén a creación dunha contorna para integrar distintas ferramentas nunha plataforma centralizada. O obxectivo deste paradigma era proporcionar aos investigadores un lugar de traballo con todas as ferramentas necesarias para: i) comparar e identificar as bases de datos de interese para os estudos clínicos; ii) axilizar un estudo sobre a rede; e iii) recuperar os resultados e agregalos. Todas estas ferramentas están protexidas por un mecanismo de inicio de sesión único federado con verificación de perfil. MONTRA2 utilízase actualmente para apoiar outros proxectos diferentes. O sistema ten tres instancias en produción, para soportar diferentes plataformas, a saber, o Portal EHDEN, o Catálogo EMIF e o Portal MSDA.

B.4.2 Recomendar bases de datos de saúde

Os investigadores necesitan analizar periodicamente as actualizacións nas bases de datos dispoñibles, buscando novos conxuntos de datos de interese. O filtrado manual é necesario porque se poden realizar novos estudos seguindo diferentes prácticas, xerando conxuntos de datos non relacionados que se centran na mesma enfermidade. Co obxectivo de simplificar a identificación correcta de novas fontes de datos de interese, propuxemos unha solución para suxerir conxuntos de datos ou publicacións similares aos usuarios involucrados nun estudo clínico, aumentando a información de interese. Esta solución recomenda novas fontes de datos baseadas en perfís de usuario, mantendo aos investigadores actualizados sobre estudos similares realizados con datos de MONTRA2.

O filtrado colaborativo nos sistemas de recomendación produce suxerencias específicas para os usuarios, segundo patróns de uso ou cualificacións. Estas suxerencias pódense realizar despois de recompilar as preferencias de varios usuarios que se

consideran con intereses similares[192]. Doutra banda, un sistema de recomendación baseado en contido tenta dar unha suxerencia baseada na cualificación do usuario e no contido do artigo e a súa similitude. Isto calcúlase en función das características máis relevantes[195]. O sistema de recomendación proposto combina as dúas técnicas presentadas para encher os baleiros de cada metodoloxía illada. O filtrado colaborativo pode detectar perfís de usuarios similares e proporcionar recomendacións cando a estrutura das fontes de datos varía significativamente. Doutra banda, a recuperación baseada no contexto pode proporcionar mellores suxerencias, baseándose unicamente na similitude das fontes de datos. Por tanto, aplicamos métricas para medir primeiro cada enfoque e logo combinar ambos.

B.4.3 Explorar bases de datos distribuídas a nivel de paciente

A metodoloxía proposta para axilizar a execución de estudos multicéntricos baséase en MONTRA2. Para lograr isto, desenvolvemos unha ferramenta adicional que se integrou en MONTRA2 como complemento. Ten como obxectivo simplificar a execución dos estudos de saúde, así como centralizar e coordinar as operacións entre todas as entidades involucradas. O sistema proposto, designado como Study Manager, adoptou as mesmas tecnoloxías utilizadas en MONTRA2, é dicir, Django no seu núcleo. Para simplificar a integración entre sistemas, esta ferramenta implementouse para cumprir con MONTRA SDK, seguindo un patrón de software MVC. Este patrón segrega a lóxica da aplicación en tres elementos principais: i) o modelo, responsable de manexar o almacenamento de datos; ii) a vista, que xera a representación de datos para o cliente; e iii) o controlador, que contén a capa empresarial.

Con todas as solucións propostas nos apartados anteriores, incluído o marco MONTRA2, o proceso de realización de estudos médicos multicéntricos é actualmente unha realidade. Os investigadores de ciencias da saúde e da vida identificaron varias oportunidades para compartir datos. Estas oportunidades só pódense lograr se os investigadores poden compartir datos entre eles. A estratexia proposta neste traballo os empodera con conxuntos de datos máis grandes para cada estudo, o que aumenta o impacto dos seus achados. Con todo, con esta idea expuxéronse diferentes cuestións gobernamentais. Por tanto, as estratexias propostas teñen como obxectivo facilitar a exploración de bases de datos a nivel de paciente, minimizando o risco de violar a privacidade do paciente.

B.5 Conclusións

Enriquecer as etapas de extracción de información nos sistemas de apoio á decisión clínica é un tema de investigación que pode abordarse desde diferentes puntos de vista. Neste traballo tentamos enriquecer estas etapas comezando por traballar sobre as bases dos sistemas de apoio á decisión clínica. Recoñecemos que para aumentar a calidade dos tratamentos, os investigadores deben estudar o impacto dos novos medicamentos ou a eficiencia dos tratamentos actuais. Estes achados poden orixinar novos protocolos de tratamento que poden integrarse nos sistemas de apoio á decisión das institucións de saúde. Por tanto, neste traballo, enfocámonos en crear metodoloxías e ferramentas para axudar aos investigadores médicos a realizar achados máis impactantes, para mellorar a fonte destes sistemas.

Comezamos especificando o alcance deste traballo, en función dos formatos de datos biomédicos que poderíamos utilizar. Motivados polo proxecto EHDEN, centramos este traballo nos datos relacionais de EHR, que tentamos complementar con datos extraídos de narrativas médicas. Logo, na etapa posterior, despois de definir estratexias para ter unha rede interoperable de fontes de datos, propuxemos solucións para apoiar a investigación utilizando estas fontes de datos. En resumo, presentamos varias solucións de software para integrar datos biomédicos e o produto final é unha plataforma que facilita a exploración desta información a través de bases de datos.

A primeira hipótese abordou a falta de interoperabilidade entre as bases de datos de saúde. Con todo, como atopamos durante este traballo, o problema non foi a falta de solucións estándar para interconectar estas bases de datos. En cambio, o problema foi o esforzo requirido para adoptar un destes estándares. Para responder a este problema, propuxemos solucións para simplificar a migración de datos EHR a un dos esquemas de datos estándar que se utilizan actualmente en estudos médicos. Validamos estas solucións utilizando cohortes heteroxéneas de datos de pacientes que padecen a enfermidade de Alzheimer. A interoperabilidade asegurouse mediante a conversión de fontes de datos ao esquema OMOP CDM.

A segunda hipótese referíase a enriquecer a información almacenada nas bases de datos, utilizando datos non estruturados presentes nas narrativas clínicas. Para iso propuxemos unha solución capaz de extraer conceptos médicos e almacenalos nunha base de datos OMOP CDM. Parte desta solución está apoiada polo traballo

realizado para responder á primeira hipótese. Validamos as estratexias NLP propostas utilizando desafíos científicos, concretamente organizados pola organización n2c2.

Finalmente, a terceira hipótese centrouse en atopar as bases de datos de saúde máis adecuadas para estudos de investigación específicos. Para responder a esta pregunta, colaboramos durante este programa de doutoramento cos socios de EHDEN co obxectivo de propoñer e axustar unha solución baseada en necesidades reais. O resultado foi un marco flexible capaz de ampliarse para admitir ferramentas complementarias. Este traballo foi validado no contexto do proxecto EHDEN. Ademais, tamén substituíu tecnoloxías antigas que apoiaron o proxecto EMIF no pasado. Esta ferramenta foi validada con miles de usuarios, cun gran impacto en contornas da vida real.

