



Privacy compliant health data as a service for AI development

Grant Agreement Number: 101095384

D3.8 Data Harmonization for Daas, Maas and Specifications plan

Deliverable Identifier:	D3.8
Deliverable Version:	v.1.0
Status	Final (F)
Work Package:	WP3-4-5
Task:	Task 3.1, 4.1, 5.1
Author(s) and Organisation:	Kassiani Zafeirouli (AIN), Thanos Vidakis (AIN), Christos Chatzichristos (KU Leuven), Roger Marí Molas (EUT), Andy Burton (NTU)
Peer Reviewer(s):	Rafael Redondo Tejedor (EUT), Christos Chatzichristos (KU Leuven)
Deliverable Due Date:	2024/10/31
Deliverable Submission Date:	2024/10/31
Dissemination Level:	PU: Public
Funding Authority:	European Commission
Funding Program:	Horizon Europe Health Work Programme 2021 – 2022
Topic:	HORIZON-HLTH-2022-IND-13-02
Rights:	PHASE-IV-AI Consortium

Document Control History

Version	Date	Edited by	Modification reason
v.0.1	2024/mm/dd	AIN	TOC and 1 st draft
v.0.2	2024/09/19	KUL	Input for UC2 data harmonization
v.0.3	2024/09/25	NTU, EUT	Input for UC1 data harmonization
v.0.4	2024/09/27	AIN	Version ready for peer review
v.1.0	2024/10/28	AIN	Version ready for submission

Executive Summary

The deliverable presents the outcomes of T3.1, T4.1, T51 related to data harmonization strategies and documentation process. Within these tasks the data harmonization requirements of the available data were identified based on the needs of the data owners. Moreover, the available datasets and the selected requirements were examined to identify the data harmonization specifications for the strategies and methods that will be developed. The selected intra- and inter-partner harmonization approaches were reported, including comprehensive analysis of the pipelines. These harmonization pipelines implemented on data provided by the data owners and a whole process is presented.

Disclaimer

The content of the publication herein is the sole responsibility of the publishers and it does not necessarily represent the views expressed by the European Commission or its services.

While the information contained in the documents is believed to be accurate, the authors(s) or any other participant in the PHASE-IV-AI consortium make no warranty of any kind with regard to this material including, but not limited to the implied warranties of merchantability and fitness for a particular purpose.

Neither the PHASE-IV-AI Consortium nor any of its members, their officers, employees or agents shall be responsible or liable in negligence or otherwise howsoever in respect of any inaccuracy or omission herein.

Without derogating from the generality of the foregoing neither the PHASE-IV-AI Consortium nor any of its members, their officers, employees or agents shall be liable for any direct or indirect or consequential loss or damage caused by or arising from any information advice or inaccuracy or omission herein.

Copyright message

©PHASE-IV-AI Consortium, 2023-2026. This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both. Reproduction is authorised provided the source is acknowledged.

Contents

1. Introduction	7
1.1 Document Scope.....	7
1.2 Document Structure.....	7
1.3 List of Acronyms.....	8
2. PHASE IV AI datasets harmonization needs	9
2.1 UC1 Lung Cancer Data	9
2.1.1 General Information about the data.....	9
2.1.2 Dataset details	9
2.1.3 Harmonization requirements	10
2.2 UC2 Prostate Cancer Data.....	10
2.2.1 General Information about the data.....	10
2.2.2 Dataset details	11
2.2.3 Harmonization requirements	11
2.3 UC3 Ischemic Stroke Data	11
2.3.1 General Information about the data.....	11
2.3.2 Dataset details	12
2.3.3 Harmonization requirements	12
3. Data Harmonization and Documentation Plan.....	13
3.1 OMOP CDM Harmonization Methodology for EMR Data	13
3.2 Harmonization Methodology for Medical Imaging Data	13
3.3 Data Harmonization Documentation Plan.....	13
4. Data as a Service: Definition of data harmonization strategy	15
4.1 DaaS UC1 Lung Cancer Harmonization	15
4.1.1 Standardized mapping of DICOM CT images and annotations.....	15
4.2 DaaS UC2 Prostate Cancer Harmonization.....	16
4.2.1 Feature extraction from free text.....	16
5. Model as a Service: Definition of data harmonization strategy	17
5.1 OMOP CDM for PHASE IV AI healthcare data.....	17
5.1.1 Selection of data harmonization tools for OMOP CDM	17
5.1.2 Generic Description OMOP Pipeline	19
5.2 MaaS Lung Cancer Harmonization	20
5.2.1 OMOP CDM for Lung Cancer data	20
5.2.2 Retrospective thoracic CT scans harmonization	21

5.3	MaaS UC2 Prostate Cancer Harmonization	22
5.3.1	OMOP CDM for Prostate Cancer data.....	22
5.3.2	Feature Alignment.....	28
6.	Conclusions	29
7.	References	30

Table of Figures

Figure 1: A part of a DICOM file header	15
Figure 2: Data Flow Diagram illustrating the transformation of source data using Rabbit-In-A-Hat.....	23
Figure 3: Flowchart of the semantic separation of concepts and how they are mapped into the OMOP CDM	25
Figure 4: Diagram showing the process where raw clinical data is mapped through a Concept Mapper to the Stem_Table.....	26
Figure 5: Diagram showing the post-processing procedure where data from the Stem_Table is transformed and stored in the standardized tables of OMOP CDM.....	27
Figure 6: Data Quality Assessment results of the KUL Prostate Cancer subset	27

1. Introduction

This deliverable is reporting the outcome of T3.1 - Data Harmonization for DaaS, T4.1 - Data Harmonization for MaaS and T5.1 - Data Harmonization, Meta Data and Documentation Features for the Health Data Hub tasks. As it is described in Description of Action (DoA) “ T3.1 addresses the diversity within the dataset, stemming from variations in medical imaging techniques and expert annotations, all under the same data proprietor.”, “T1.2 addresses the diversity and inconsistencies among various data custodians, stemming from disparate Electronic Medical Records (EMRs), languages, and data recording procedures. The aim of this task is to collaborate with the local data manager in each data contributing site for harmonization of the data and provision of the routines to accommodate inter-partner variability.” and “T5.1 consists of the documentation (OMOP CDM, semantic mismatches, metadata of demographic characteristics) of both processes performed in T3.1 and T4.1, that will be available within the health data hub.”.

Healthcare organizations hold vast amounts of data, from Electronic Health Records (EHR) to laboratory results and medical imaging, originating from diverse sources. These multi-source datasets could often present discrepancies in data formats, labelling conventions, annotation strategy, screening and storage protocols that pose challenges to further analysis. Data harmonization aims to address these challenges by standardizing and harmonizing healthcare data, ensuring that data from various sources are uniformly structured, ready for use by project’s analytical tools. The developed data harmonization strategy will improve data interoperability and quality, ensure seamless cross-institutional collaboration and enable real-world data analytics.

1.1 Document Scope

The goal of this deliverable is to provide a comprehensive overview of the data harmonization strategy will be implemented within the scope of the project. Specifically, it reports and details harmonization requirements of each use case and its associated datasets, focusing on both the unique internal needs of data owners (DaaS approach) and the needs arising from inter-partner collaboration (MaaS approach). A data and metadata analysis process are presented to assess the available datasets and define the specifications for data harmonization. Additionally, the document explores and describes the available data harmonization methods and techniques for healthcare data (tabular and imaging) that are suitable for addressing the identified requirements. Finally, the deliverable presents the results achieved thus far of implementing the selected harmonization strategy and pipeline on the data provided by the respective data owners.

The strategy and preparation process for producing the homogenised datasets and the documentation activities of the harmonization process of T3.1, T4.1, T5.1 will be presented in D3.8 and the finalised homogenised datasets, documented process and activities will be delivered in D5.2.

1.2 Document Structure

The document includes the following sections:

- **Section 1:** Provides an overview of this document, setting its scope and objectives.
- **Section 2:** Reports the harmonization needs and requirements for each use case and its associated datasets.
- **Section 3:** Provides an overview of the data harmonization specifications and the metadata documentation.
- **Section 4:** Describes the suggested intra-partner harmonization strategy into the DaaS.
- **Section 5:** Describes the inter-partner harmonization strategy into the MaaS and provides details on its implementation across specific data.

- **Section 6:** Summarizes the document and discusses potential future work.

1.3 List of Acronyms

List of Acronyms	
AI	Artificial Intelligence
CDM	Common Data Model
CPRD	Clinical Practice Research Datalink
CSV	Comma-Separated Values
CT	Computed Tomography
DaaS	Data as a Service
DICOM	Digital Imaging and Communications in Medicine
DQD	Data Quality Dashboard
EHR	Electronic Health Record
ETL	Extract, Transform, Load
IDC-P	Intraductal carcinoma of prostate
JSON	JavaScript Object Notation
LIDC-IDRI	Lung Image Database Consortium - Image Database Resource Initiative
Maas	Data as a Service
MRI	Magnetic Resonance Imaging
NifTI	Neuroimaging Informatics Technology Initiative
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
PCa	Prostate Cancer
PoC	Proof of Concept
PSA	Prostate Specific Antigen
SW	Software
WP	Work Package

2. PHASE IV AI datasets harmonization needs

Section 2 outlines the different harmonization requirements for each partners' datasets, based on either their specific internal needs or needs arising from inter-partner collaboration. Intra-partner harmonization includes the processes required to transform a partner's data in a form suitable for further analysis (e.g., medical images format, entity extraction from unstructured text). Conversely, inter-partner harmonization involves actions to ensure that datasets from multiple sources are standardized in a common manner for data interoperability and seamless collaboration (e.g., adopting a common data model for tabular data or standardised annotation format for medical images).

2.1 UC1 Lung Cancer Data

For the Lung Cancer use case both tabular and medical imaging data are used to build AI-driven methods and tools for lung cancer risk prediction, nodule growth prediction and synthetic image generation. Detailed descriptions for the available lung cancer datasets are reported in D6.1 "User stories, usage scenarios and use case validation v1". Below is a brief description of the datasets that have been used and analysed in this phase, focusing mainly on their intra- and inter-partner harmonization needs.

2.1.1 General Information about the data

The tabular datasets include:

- EHRs
- Laboratory tests

The medical imaging datasets include:

- Thoracic CT scans
- Nodule annotations (segmentation mask, class)
- Image metadata

These data will be used by the researchers to build tools to achieve the below objectives:

1. Predict the risk of lung cancer to invite patients for screening (tabular data)
2. Generate synthetic lung CT scans, including pulmonary nodules (medical imaging data)
3. Build methodology for synthetic data evaluation (medical imaging data)

2.1.2 Dataset details

Clinical Practice Research Datalink (CPRD) (NTU, NUH) [1]:

- EHRs from primary care practices in UK: CSV files accompanied with descriptions of fields (~19 million samples) (objective 1)

VARHA EHRs dataset [2]:

- EHRs from primary care practices in SW Finland: Relational Database (objective 1)

Lung Image Database Consortium (LIDC-IDRI) [3]:

- Images from a clinical thoracic CT scan, in DICOM format, and an associated XML file that records the results of a two-phase nodule annotation process (1018 cases) (objective 2, 3)

VHIR CT scans dataset:

- Low-dose CT scans and annotations in CSV file (147 cases) (objective 2, 3)

2.1.3 Harmonization requirements

It is important to highlight that the CPRD, VARHA and VHIR datasets are private, and access requires a specific protocol, licence and a time-consuming approval process. At this stage, AIN team, the data harmonization technical partner, does not have access to these datasets. Given that, we identified the harmonization needs through alignment meetings with the relevant partners, based on the information provided by the data owners and partners with access.

1. A part of the CPRD dataset will be used for UC1, containing only the relevant lung cancer related records. This subset of the CPRD dataset is not yet harmonised.
2. VARHA is in the process of mapping and harmonizing its dataset into OMOP CDM (v5.3). Once more information about the harmonization process for the two datasets is available, we will assess if a common approach has been followed and provide additional guidelines for potential modifications.
3. Regarding the medical imaging datasets, we are currently working with the publicly available LIDC-IDRI dataset to simulate the VHIR private dataset and develop the necessary harmonization processes and guidelines related to formatting, resolution, etc. If anonymized sample data from VHIR dataset are provided, we will adapt the methodology to ensure it aligns with dataset's specific requirements.

2.2 UC2 Prostate Cancer Data

The purpose of this use case is to establish predictive analytics for personalized medicine in the field of PCa and evaluate its suitability for potential replication in other oncology areas. For this objective, tabular data is utilised to develop and evaluate the relevant techniques and tools. Detailed descriptions of the available datasets are provided in D6.1, while a summary of the type of information considered and its harmonisation needs is summarised below.

2.2.1 General Information about the data

The tabular datasets include:

- EHRs
- Patient Report Outcomes
- Biopsy/Pathology reports
- Lab results

These data will be used by the researchers to build tools to achieve the below objectives:

1. Predict lymph node invasion prior to any surgery [4]
2. Predict biopsy outcome [5]
3. Reproduction of results of PIONEER study [6]
4. Gain new insights into the disease mechanism of PCa and its progression from a localized to a metastatic state, and into patient risk stratification. These insights will help identify new clinically

relevant endpoints for this unmet medical need and help identify which patients are resistant to current treatment options.

2.2.2 Dataset details

KUL dataset:

- Undergone prostatectomy (1110 cases) (objective 1)
- Underwent prostate cancer biopsy (~3200 cases) (objective 2)
- EHRs: CSV files accompanied with HTML file with description of the fields (objective 1, 3, 4)
- Patient report outcomes: CSV files accompanied with HTML file with description of the fields
- Biopsy/pathology reports: CSV/text files with free text (objective 2, 4)

UTU dataset:

- Undergone prostatectomy (1700 cases) (objective 1)
- Underwent prostate cancer biopsy (~4300 cases) (objective 2)

2.2.3 Harmonization requirements

Based on the different objectives of the study and the alignment meetings between the data providers we have identified the main harmonization requirements as following:

1. The dataset of UTU is harmonized in OMOP CDM while the KUL dataset is stored in formamas. We will need to align the two datasets by mapping and harmonizing the KUL dataset into OMOP. The version of OMOP that will be used is v5.4 [7]. The initial mapping of UTU dataset did not include the oncology extension of OMOP but this will also be implemented.
2. According to main objective 1, biopsy results are essential for the development of the risk stratification tool. We will need to extract relevant information (i.e. presence of cribriform and intraductal carcinoma) from free text, and then map them to OMOP.
3. According to main objective 2, the variable ARI-5, which indicates whether the patient has received any medication which is categorised as 5a-reductase inhibitors is essential since it might result in a drop in PSA (Prostate Specific Antigen).

The mentioned harmonization needs will be tackled in detail in the following parts of the deliverable.

2.3 UC3 Ischemic Stroke Data

This use case focuses on AI supported technology solutions for ischemic stroke detection and treatment planning. Medical imaging data (MRI, CT) is utilized to train, validate and evaluate the relevant solutions. Detailed descriptions for the available ischemic stroke datasets are reported in D6.1. Below is a brief description of the datasets considered in this phase and their harmonization needs.

2.3.1 General Information about the data

The medical imaging datasets include:

- CT, MRI scans
- Brain lesion annotations (segmentation mask)

- Image metadata

The researchers and developers will use these data to build tools to achieve the below objectives:

1. Detect ischemic stroke from CT and/or MRI images
2. Generate synthetic brain CT and/or MRI images

2.3.2 Dataset details

Anatomical Tracings of Lesions After Stroke (ATLAS) R2.0 dataset [8][9]:

- T1-weighted MRI scans, in NIfTI format, with manually segmented lesion masks
 - 955 cases, split to Training (N=655) and Testing datasets (N=300).
 - The intensity of images is normalized, and images registered to MNI-152 1 mm³ template.
- One .csv file containing scanner metadata.
- One .xlsx file containing patient lesion and days-post-stroke information.

Ischemic Stroke Lesion Segmentation 2022 (ISLES22) dataset [10]:

- MRI scans (FLAIR, DWI and ADC), in NIfTI format, with manually segmented lesion masks (400 cases).
- Two .csv files containing lesion and scanner metadata.

Ischemic Stroke Lesion Segmentation 2024 (ISLES24) dataset [11]:

- Dataset is not available at the time of writing as official challenge is ended and long-term challenge is not yet published
- Imaging data (N=250) would include CT images (including non-contrast CT, perfusion CT, and CT angiography) as well as MRI (follow-up DWI with delineated infarct labels)
- Tabular data includes clinical and demographic data from patients.

Acute Ischemic Stroke Dataset (AISD) dataset [12]

- MRI scans, in Nifti format, with manually segmented lesion masks (397 cases).

2.3.3 Harmonization requirements

It is important to mention that the initial approach to collect the necessary data for the relevant tools was to obtain access to real-world private data. However, since this process is time consuming and requires legal approval, in parallel the developers and researchers started working with publicly available datasets. These datasets follow the same format and specifications and their needs for harmonization are limited.

- The CT/MRI data used is in standard NIfTI format, and they are named and stored in accordance with the Brain Imaging Data Structure (BIDS), a standard for organizing, annotating, and describing data collected during neuroimaging experiments.

When additional datasets become available, their harmonization needs will be assessed, and the necessary harmonization steps will be applied. The harmonization strategy developed for the medical imaging data of UC1 will be modified and adapted to meet the requirements of UC3 data.

3. Data Harmonization and Documentation Plan

This section is presenting an overview of the data harmonization process in general, and the initial documentation plan will be implemented within the scope of the PHASE IV AI project. The goal of these procedures is to ensure interoperability, enhance data quality, and facilitate usability for federated learning and AI model development by transforming diverse datasets into standardized formats, aligning structural and semantic mappings, and providing robust documentation and metadata to support efficient data usage.

3.1 OMOP CDM Harmonization Methodology for EMR Data

The OMOP CDM is a commonly used framework for data harmonization to transform and standardize EMH data from various sources into a common structure. This transformation process includes:

- Local EHR data are mapped into the OMOP CDM using structural mappings and standardized vocabularies like SNOMED, ICD10, and LOINC. The Athena tool is utilized to search and identify the most suitable vocabulary for the case.
- A well-defined ETL pipeline is followed to analyze the source data structures, generate scan reports, and assist in mapping these structures to the OMOP CDM. The harmonized data is loaded into a repository for further use.
- Quality control assessment is conducted to ensure the integrity of the transformed datasets. Automated data quality tools check for consistency in data transformation, mapping accuracy, compliance with OMOP CDM standards and completeness to ensure no critical data is lost during the harmonization process.

3.2 Harmonization Methodology for Medical Imaging Data

Medical imaging data standardization should address harmonization requirements related to different capturing methods, formats (e.g., DICOM, NIfTI), resolutions, and annotation approaches used. The harmonization framework for imaging data includes:

- Preprocessing methods are applied to address any inconsistencies in resolution, orientation, density across imaging datasets.
- Imaging data from different devices are transformed into a standard format suitable for further analysis, such as 3D matrices or NIfTI. This ensures consistent representation across different data sources.
- Relevant annotations (e.g., segmentation masks, labels) are standardized in specific format (e.g., JSON), ensuring that consistent labels are used across different datasets.

3.3 Data Harmonization Documentation Plan

Comprehensive metadata documentation is crucial for understanding and reusing data efficiently. A metadata documentation plan is developed to ensure that both structural and semantic information about the datasets is captured. This includes:

- Detailed documentation of structural mappings, indicating how each dataset has been transformed to align with the OMOP CDM or the standard imaging format. This includes information on any transformations or data modifications applied during the harmonization process.
- To support transparency and data usability, structured reports on demographic characteristics, their distributions, dataset completion, and missing entries will be provided. The results and metadata from

the quality control assessments will also be included to the reports. These metrics offer a clear view of each dataset's structure and quality allowing users to evaluate data representativeness and integrity.

The information will be forwarded to Health Data Hub (see D5.3), which includes tools for search and filtering based on:

- **Demographic Characteristics:** Age, gender, and other relevant variables.
- **Coding Information:** Diagnosis codes (ICD10, SNOMED)
- **Data Quality:** Indicators such as completeness, missing values, and format consistency.
- **Condition-Based Filters:** Select patients by diagnosis or conditions using clinical codes.

4. Data as a Service: Definition of data harmonization strategy

DaaS platform incorporates the data harmonization strategy and tools designed to address the unique harmonization requirements for each partner and its data. Data discrepancies may arise due to the use of varied medical imaging equipment, differences in labelling practices, inconsistent storage conventions, and the presence of incomplete or inaccurate records. Below the specific harmonization needs and the suggested approaches for each use case and dataset are described in detail.

4.1 DaaS UC1 Lung Cancer Harmonization

4.1.1 Standardized mapping of DICOM CT images and annotations

Modern medical imaging devices typically store data in the DICOM (Digital Imaging and Communications in Medicine) format. DICOM is a tag-based format where each object within the file is encapsulated in a tag that defines the purpose and size of that data segment. Earlier objects in the file typically store information about the patient, the device, the imaging sequence, and image specifics (e.g., image dimensions), while the final object contains the actual image data.

```

| (0029,1044) OB          18 [private]                [binary 8-bit data]
| (FFFE,E00D)          Item Delimitation Item
| (FFFE,E000)          -1 Item
| (0029,0010) IO        22 [private]                SIEMENS MEDCOM HEADER
| (0029,1041) CS         10 [private]                SOM 5 TPOS
| (0029,1042) LO        18 [private]                SOM 5 NULLPOSITION
| (0029,1043) LO        14 [private]                VB10A 20030626
| (0029,1044) OB        12 [private]                [binary 8-bit data]
| (FFFE,E00D)          Item Delimitation Item
| (FFFE,E0DD)          Sequence Delimitation Item
| (0032,1060) LO         8 Requested Procedure Description  CT HEAD
| (0032,1064) SQ        -1 Requested Procedure Code Sequence [sequence]
| (FFFE,E000)          -1 Item
| (0008,0100) SH         6 Code Value                CSKUH
| (0008,0102) SH         2 Coding Scheme Designator    MH
| (0008,0104) LO         8 Code Meaning              CT HEAD
| (FFFE,E00D)          Item Delimitation Item
| (FFFE,E0DD)          Sequence Delimitation Item
| (0040,0275) SQ        -1 Request Attributes Sequence [sequence]
| (FFFE,E000)          -1 Item
| (0040,0007) LO         8 Scheduled Procedure Step Description  CT HEAD
| (0040,0009) SH         4 Scheduled Procedure Step ID    5788
| (0040,1001) SH         8 Requested Procedure ID    19310917
| (FFFE,E00D)          Item Delimitation Item
| (FFFE,E0DD)          Sequence Delimitation Item
| (3711,0010) LO        26 [private]                A.L.I. Technologies, Inc.
| (3711,100C) UI        34 [private]                1.2.840.10008.5.1.4.31.32243680813
| (7FE0,0010) OB        -1 Pixel Data                [binary 8-bit data]

```

Figure 1: A part of a DICOM file header

The CT scans datasets that are used in the lung cancer use case (LIDC-IDRI, VHIR) are stored in the DICOM format. It is crucial to transform this information in a more usable format, as DICOM, while being a widely used standard to transfer, store and printing medical data, aligns poorly with the needs of image analysis tools. Furthermore, different medical imaging equipment manufacturers implement the DICOM format differently. Therefore, a standardised mapping method is essential to transform DICOM CT images and annotations to a format suitable for analysis (e.g., JPEG or NIFTI for the images and JSON for the metadata). Follow a common transformation method ensure the harmonization DICOM data from different devices, which may be used by the same partner, thereby preventing misalignments in the results.

4.2 DaaS UC2 Prostate Cancer Harmonization

4.2.1 Feature extraction from free text

Readily accessible, comprehensible, and standardized biopsy information is crucial for timely patient-centric care without subjecting patients to unnecessary procedures or delays. The purpose of this intra-partner harmonization task is to extract key features, i.e. terms previously selected by healthcare professionals, of prostate cancer screening performed at the hospital, in an automated manner.

In prostate cancer, the presence of cribriform carcinoma or intraductal carcinoma during a biopsy can have significant implications for predicting lymph node invasion. Cribriform carcinoma is a histological pattern of prostate cancer where the cancer cells form gland-like structures with a cribriform (sieve-like) appearance. It is often associated with more aggressive disease.

Intraductal carcinoma of the prostate is a variant where cancer cells grow within the ducts of the prostate gland, often presenting with high-grade disease. IDC-P is known for its aggressive behaviour and is frequently found alongside high-grade acinar adenocarcinoma. It is often a marker of advanced disease. The presence of IDC-P in a biopsy indicates a higher risk of more extensive disease spread, including lymph node involvement.

Identifying these patterns can help stratify patients into higher-risk categories. This can influence treatment decisions, such as the consideration of more aggressive systemic therapies or closer monitoring for signs of lymph node involvement. Since one of the objectives of the prostate cancer use case is the risk prediction for prostate cancer the inclusion of such features is important. In the structured data of UZ Leuven such information is not yet included. However, in the pathology reports the related information for the presence of intraductal and cribriform carcinoma can be identified via the use of Natural Language Processing (NLP) techniques. Having extracted the features, the information will be mapped to SNOMED-CT10 and then converted to OMOP CDM to ensure EHDS standards are followed.

The main challenges of this tasks include, among others, the unstructured form of documentation concerning the biopsies, lack of standardization of inputted details, incomplete or inaccurate records, overlapping interventions (e.g., random, and targeted biopsies), ambiguity/ interchangeability in the terms used (Gleason grade and score), the different styles and syntax of reporting (e.g., Gleason grade=25% vs 25% of Gleason grade), potential typographical and spelling errors.

The goal of this task is to transform free text data to findable, accessible, interoperable, and reusable (FAIR) data, an essential prerequisite for the effective (re-)use of different types of clinical data for decision support, patient follow-up and clinical research.

5. Model as a Service: Definition of data harmonization strategy

MaaS platform incorporates the data harmonization strategy and tools to address the inter-partner data heterogeneity. This variability may arise from the use of different EHR software, languages, protocols or diverse screening devices and recording procedures. The goal of the inter-partner data harmonization is to ensure data interoperability and seamless collaboration between multiple data owners. To achieve this, after discussion with the relevant partners and comprehensive research, it was decided to adopt the OMOP CDM standard to standardize the structure and content the tabular healthcare data (HER data, other metadata). For the medical imaging data, a review of the existing harmonization approaches suitable for PHASE IV AI use cases was conducted and an initial pipeline was created. The following sections detail the strategy, tools, and methods used, as well as their application within the project to date.

5.1 OMOP CDM for PHASE IV AI healthcare data

5.1.1 Selection of data harmonization tools for OMOP CDM

The OMOP harmonization process is supported by a range of open-source tools developed by OHDSI (Observational Health Data Sciences and Informatics) [13], designed to facilitate the transformation, standardization, and analysis of healthcare data within the OMOP CDM. Key tools include WhiteRabbit and RabbitInAHat, which help analyse and map the source data to the CDM structure, and ETL tools that automate the data transformation process. Additionally, platforms like Atlas enable researchers to explore, visualize, and analyse the harmonized data for tasks such as cohort creation and outcome studies. These tools ensure a transparent, and reproducible approach to data standardization and analysis, making it easier for researchers to conduct large-scale studies across diverse datasets.

The main open-source tools used into the OMOP CDM process are reported below:

WhiteRabbit: WhiteRabbit is a software tool designed to assist in preparing the ETL processes of mapping healthcare databases into the OMOP CDM. It supports source data from comma-separated text files or databases like MySQL, SQL Server, ORACLE, and PostgreSQL. WhiteRabbit's is mainly used to scan the source data and provide detailed information about the tables, fields, and field values. The resulting scan report serves as a reference for designing the ETL process.

The key features of the WhiteRabbit include:

- **Data Scanning:** WhiteRabbit analyzes the source database and generates a comprehensive report on its structure, including tables, columns, data types, and sample records.
- **Data Profiling:** It analyses the data to identify key characteristics, such as the count of unique values, value ranges, null or missing values, and the frequency of common values.
- **Report Generation:** The tool generates a summary report that is used in the ETL process to map source fields to the OMOP CDM.

The main goal of WhiteRabbit is to provide insights into data distribution, check for inconsistencies, and evaluate the overall quality of the source data.

Rabbit-In-A-Hat: Rabbit-In-A-Hat is another key tool in the OHDSI toolkit, specifically designed to facilitate the mapping of source data to the OMOP CMD. While WhiteRabbit provides detailed information about the source data, Rabbit-In-a-Hat uses this information in a graphical user interface to help users map the source data to the tables and columns of the CDM. It generates documentation for the ETL process but does not produce code for creating the ETL itself.

The key features of the Rabbit-In-A-Hat include:

- **Visual ETL Mapping:** Users are presented with a side-by-side view of the source data structure and the OMOP CDM, enabling them to create mappings between source fields and OMOP CDM fields through a drag-and-drop interface.
- **Documentation of Transformations:** Rabbit-In-a-Hat automatically records the mappings and transformations, generating a comprehensive and detailed ETL specification that can be easily provided to developers.
- **Simplifies Complex Mappings:** It is especially valuable for handling complex mapping scenarios, such as transforming diagnosis codes (e.g., ICD-10 to SNOMED), handling multi-value fields, or merging data from different source systems.
- **Comprehensive ETL Documentation:** Produces an ETL specification document that serves as a guide for the transformation process, whether carried out manually or through automated ETL tools.

Athena Vocabulary: Athena is a web tool developed by OHDSI designed to facilitate the management, searching, and use of standardized vocabularies and ontologies used in the OMOP CMD. Athena serves as the primary interface for users to access and explore standard concept vocabularies, such as SNOMED CT, RxNorm, LOINC, and ICD codes, which are essential for the consistent representation of health data across different systems. Athena enables users to search for and download standardized vocabulary concepts, access metadata about these concepts, and explore their relationships within the vocabulary. This tool is essential for ensuring that data mapping and querying processes rely on accurate and up-to-date standardized terms.

The key features of Athena include:

- **Vocabulary Search:** Allows users to search for concepts across multiple vocabularies, using various criteria such as code, description, or concept class.
- **Concept Details:** Provides detailed information about individual concepts, including their definitions, synonyms, and hierarchical relationships.
- **Concept Relationships:** Displays relationships between concepts, including parent-child relationships and mappings between different vocabularies.

Usagi: Usagi is a tool designed to support the manual process of code mapping by suggesting mappings based on the textual similarity of code descriptions. It enables users to search for appropriate target concepts when automated suggestions are inaccurate. Once a correct match is found, users can approve mappings for use in the ETL process. Source codes are loaded into Usagi, and if the codes are not in English, additional translation columns are required.

The key features of Usagi include:

- **Automated Suggestions:** Provides suggested mappings based on text similarity between source codes and standardized concepts.
- **Manual Review:** Allows users to manually review, select, or modify the suggested mappings to ensure they are suitable.
- **Support for Multiple Vocabularies:** Usagi can map to various standardized vocabularies, including SNOMED CT, RxNorm, LOINC, ICD-10, and more.

Achilles: ACHILLES is a tool within the OHDSI, designed to perform automated data quality assessments and generate summary statistics for observational health data stored in the OMOP CMD. It provides valuable insights into the data, helping users understand dataset characteristics and identify potential quality issues before undertaking more complex analyses. The results are presented in a clear, visual format—often through

interactive dashboards like ACHILLES Web-making it easier for researchers to explore and assess their data. ACHILLES ensures the transformed data meets the expected standards and is ready for analysis.

Key Features of Achilles include:

- **Data Quality Assessment:** ACHILLES identifies data anomalies, including missing values, invalid dates, and inconsistent relationships between tables.
- **Automated Reports:** Produces interactive reports that offer a high-level summary of the data in the CDM, enabling users to explore data quality and identify general trends.
- **Integration with OHDSI Tools:** ACHILLES outputs can be leveraged with other OHDSI tools like Atlas, facilitating more comprehensive studies and cohort analyses.

Atlas: Atlas is a web-based, open-source application developed by OHDSI that provides a graphical user interface for researchers to work with data in the OMOP CDM. It functions as a central platform for designing and conducting various analyses on large-scale observational health data, including cohort definitions, study designs, and data characterizations.

Key features of Atlas include:

- **Cohort Creation:** Defining patient groups by applying specific inclusion and exclusion criteria.
- **Data Exploration:** Offers tools to explore and analyze the structure and content of OMOP CDM datasets without requiring SQL queries.
- **Integration with WebAPI:** Atlas integrates with WebAPI to handle back-end communication with OMOP CDM databases, enabling real-time interaction with healthcare data.

Broadsea: Broadsea is a Docker-based deployment framework that streamlines the installation and setup of key OHDSI tools, including Atlas, WebAPI, and Achilles, for working with the OMOP CDM. By using containerization, Broadsea ensures quick, consistent deployment of the OHDSI ecosystem across different environments, removing the complexities of manual configuration. It combines tools like Atlas (for cohort creation and analysis), WebAPI (for connecting to OMOP CDM databases), and Achilles (for data characterization and quality checks) into pre-configured containers, enabling institutions to easily integrate, scale, and manage their research infrastructure.

5.1.2 Generic Description OMOP Pipeline

The OMOP pipeline process involves the transformations of raw healthcare data into a standardized format that enables efficient, reliable, and comparable research across multiple data sources. This process uses a CMD to ensure consistency and interoperability.

The first step of OMOP CDM process is the extraction of raw data from various healthcare systems such as EHR, Clinical trials, Health surveys, etc. In some cases, the received data should be translated in English for better understanding of the given data as well as more comfortable usage of the tools provided by the OHDSI framework.

The second step in the OMOP process, suggested by OHDSI, is the scan of our source data performed by WhiteRabbit. WhiteRabbit generates a scan report for each table along with an overview of the data. The overview scan report will list all the scanned tables, each variable, the data types and empty values. The most useful information that we can gather from the scan report is the truncated list of variables with the frequency of each variable which is significantly valuable in the next phases.

The next step in the typical OHDSI software workflow involves using Rabbit-in-a-Hat to read and display the WhiteRabbit scan document, which provides detailed information about the source data. Through its graphical interface, Rabbit-in-a-Hat allows users to map source data to the appropriate tables and columns in the CDM. Users can also store concept IDs for each variable within Rabbit-in-a-Hat. Additionally, the tool generates documentation for the ETL process.

One of the most critical steps in this process is the variable mapping of the source concepts. This can be performed manually with the use of the Athena Vocabulary environment. However, this can be rather difficult and time-consuming. The OHDSI community has developed another tool to help map codes from a source system into standard terminologies from OMOP Vocabularies. Usagi suggests a range of matches based on the textual similarity of code descriptions where the user can choose the concept which is most suitable for his/her use case. If source codes are in a foreign language, using Google Translate often provides surprisingly accurate translations of the terms into English, which can aid in the mapping process.

The storage of data in a database is a critical step in the process. Initially, three additional schemas should be created where each one of them has stored the translated data, the data in original format and the vocabularies for the USAGI tool. Developers can then perform the transformation either by executing SQL queries or by developing scripts to automate the process.

Once the data is transformed into the OMOP CMD, it is crucial to conduct comprehensive quality control and data validation to ensure the dataset's accuracy, consistency, and reliability. This process includes verifying that all mandatory fields are populated, ensuring the data conforms to the CDM structure, and confirming the correct application of standardized terminologies such as SNOMED or RxNorm. Tools like Achilles and the DataQualityDashboard are invaluable during this phase, as they automate checks for common data issues such as missing values, incorrect mappings, and out-of-range dates. These tools provide detailed reports to highlight potential inconsistencies, helping to ensure the dataset meets stringent quality standards, which is essential for enabling trustworthy and meaningful research.

In the data analysis and study execution phase, the standardized OMOP data becomes ready for research and analysis. Atlas enables researchers to define patient cohorts, exposures, and outcomes, streamlining the design and execution of studies without requiring complex programming. Analysts can conduct various types of research, including cohort studies and population-level effect estimation, across the available datasets. The OMOP-transformed data can then be used in Atlas, enabling efficient execution of comprehensive analyses.

The OMOP pipeline provides a structured approach to transforming raw healthcare data into a standardized, research-ready format using the OMOP Common Data Model. By utilizing tools such as Usagi, WhiteRabbit, Achilles, Athena, and Atlas, healthcare institutions and researchers can collaborate effectively to produce high-quality, replicable research that enhances patient care and medical outcomes. The integration of automated tools and manual processes ensures a robust and reliable transformation pipeline, facilitating a seamless transition from data conversion to comprehensive analysis.

5.2 MaaS Lung Cancer Harmonization

5.2.1 OMOP CDM for Lung Cancer data

NTU have not yet begun the harmonisation of the lung cancer subset of the CPRD dataset to align it with VARHA's OMOP-based harmonized data. However, an initial harmonization strategy has been developed including the following steps:

1. **Mapping Risk Assessment Factors:**

- A collaborative effort will establish a comprehensive mapping table that links risk assessment factors between the CPRD and VARHA datasets. These factors may include smoking status, family history of cancer, occupational exposures, and other relevant variables.
- NUH/NTU will assess VARHA's proposed variable definitions and suggesting alternatives or refinements where necessary.
- NUH/NTU will provide in-depth descriptions of how each risk assessment factor is captured in their CPRD dataset, including specific codes, algorithms, or data extraction methods applied.

2. Comorbidity Variables:

- VARHA will focus on extracting comorbidity variables from diagnosis codes within their patient records.
- NUH/NTU will review and validate VARHA's proposed mappings between comorbidities and specific diagnosis codes, ensuring accuracy and consistency.
- NUH/NTU will provide detailed information on how comorbidities are identified and recorded within their CPRD dataset, including specific code sets and any time windows or algorithms utilised.

3. Symptom Variables:

- VARHA will primarily leverage text mining techniques to extract symptom variables from free-text patient narrative records.
- NUH will provide the most clinically relevant and impactful symptoms for lung cancer research, guiding VARHA's text mining efforts.
- NUH/NTU will share information on how symptom data is captured and processed in their CPRD dataset, whether through coded data or natural language processing approaches.

5.2.2 Retrospective thoracic CT scans harmonization

Investigating the harmonization needs for the lung cancer CT scans we conclude that a retrospective, after data collection, harmonization process required to mitigate biases from different sites or scanners, thereby ensuring data interoperability and seamless inter-partner collaboration. Below is described an initial harmonization pipeline.

For CT images:

- **Step 1:** Standardized mapping of DICOM image to a 3D matrix
- **Step 2:** Normalization of CT Hounsfield units (HU) to a specific range
- **Step 3:** Resampling to a target resolution (x mm/pixel)
- **Step 4:** Center cropping to specific dimensions (HxWxD)

For binary segmentation masks:

- **Step 1:** Standardized mapping of DICOM binary image to a 3D matrix
- **Step 2:** Resampling to a target resolution (x mm/pixel)
- **Step 4:** Center cropping to specific dimensions (HxWxD)

For image and annotations metadata:

- **Step 1:** Standardized mapping of DICOM metadata to JSON format (image id, nodule attributes, etc.)
- **Step 2:** Aligning metadata with the image transformations where necessary

As we continue refining this process, additional steps will be added to address potential biases that may arise.

5.3 MaaS UC2 Prostate Cancer Harmonization

5.3.1 OMOP CDM for Prostate Cancer data

In our efforts to harmonize the data in the Prostate Cancer Use Case our team utilized the open-source work performed by The Hyve, specifically their OHDSI-ETL-PRIAS tool [14]. This tool is used to transform our raw data into the OMOP CDM, enabling integration with other datasets and supporting robust observational research. This set of tools not only establishes the structure of the OMOP CDM database but also automates the time-consuming process of mapping raw data concepts to standardized OMOP concepts.

5.3.1.1 Methodology

Data Collection and Translation

Step 1: Data Collection from KU Leuven

The initial phase of our work involved collecting data from KU Leuven, which consisted of four EHR forms known as Formasa. These forms were filled out by healthcare professionals for each prostate cancer patient. Additionally, four Excel files were provided, containing a sample dataset of the EHR forms, each detailing a comprehensive, step-by-step outline of the prostate cancer examination process including patient health information, examination results, medication details, and any other relevant data.

Step 2: Translation of Excel Files

To ensure that all internal project stakeholders could easily understand the data, each of the Excel files was meticulously translated from Dutch to English. This translation was crucial for the subsequent data processing steps, as it allowed for better collaboration and comprehension among us.

Data Preparation and Initial Analysis

Step 3: Extraction of Important Columns

Post-translation, with the contribution of our partners we identified and extracted only the important columns from each Excel file. These columns were selected based on their relevance and potential to provide valuable insights for our partners in the following phases of the project, focusing on critical variables that would be most beneficial for analysis.

Step 4: Data Scanning with WhiteRabbit

The translated Excel files were scanned using WhiteRabbit, a tool designed to scan the source data and provide detailed information on the tables, fields, and values that appear in a field. Finally, a report was generated that helped us prepare and organize for the intricate ETL (Extraction, Transformation, Loading) procedure.

Mapping and Concept Identification

Step 5: Variable Mapping with Rabbit-In-A-Hat

The generated WhiteRabbit scan report was imported into Rabbit-In-A-Hat, an intuitive tool that simplifies the process of mapping variables to the appropriate tables within the OMOP CDM. Rabbit-In-A-Hat streamlined the process of assigning and storing the corresponding concept IDs for each variable, ensuring that the mapping process was both accurate and efficient, significantly reducing the likelihood of errors. The following image is a visual presentation of Rabbit-In-A-Hat and our transformation of the source data to the OMOP CDM:

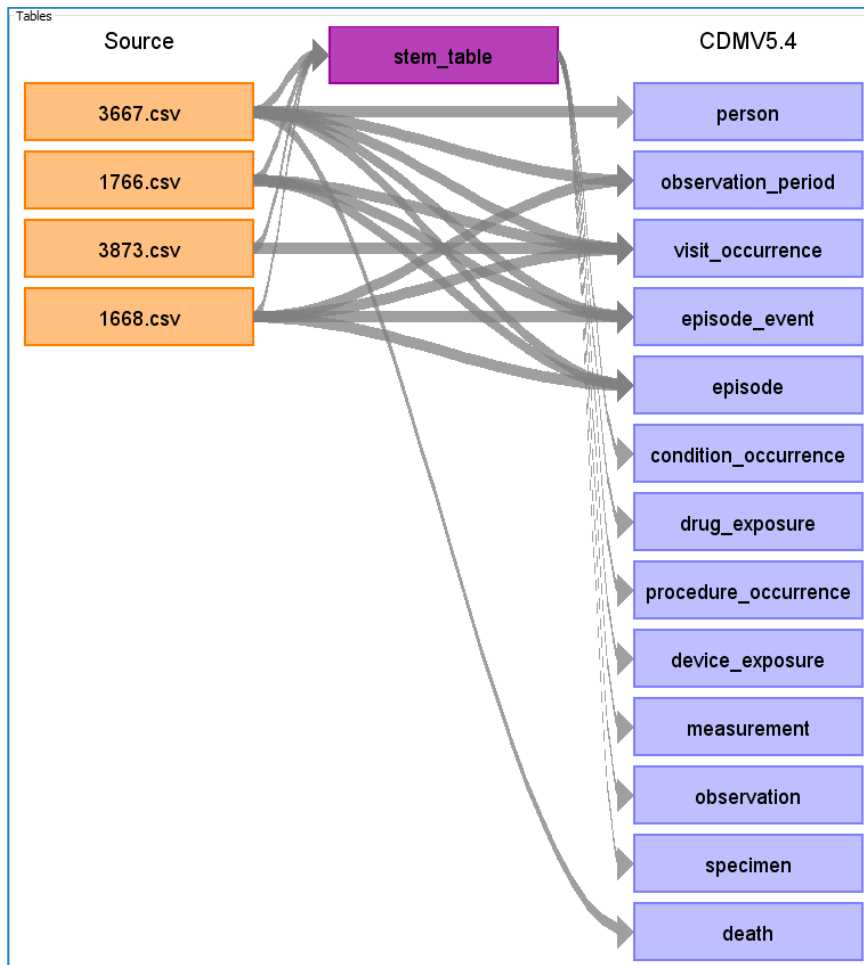


Figure 2: Data Flow Diagram illustrating the transformation of source data using Rabbit-In-A-Hat

Step 6: Concept ID Assignment

To enhance and speed the mapping process, Usagi was utilized to automatically assign concept IDs from the Athena vocabulary. Usagi processed a new CSV file derived from the WhiteRabbit scan report, searching for the most relevant text-matching concept IDs. We meticulously reviewed the results to ensure semantic accuracy. For concepts that did not have a match with an existing concept in Athena, we assigned custom concept IDs. This step streamlined the identification of appropriate concepts, saving time and improving the precision of the mapping process.

Step 7: Semantic Separation of Concepts

For the mapping of source terms to OMOP CDM concepts, we categorized our concepts in four categories. These categories are: variable_mapping, value_mapping, variable_value_mapping, and unit_mapping. The fundamental structure of this approach is based on the concept of pairing variables and values to map them correctly into the OMOP CDM.

The variable_mapping table serves the role of mapping the source field names (i.e., variables) to standardized OMOP concepts. Each row in this table represents a specific source variable, which is mapped to a corresponding field in the OMOP CDM. On the other hand, the value_mapping table handles the values that correspond to the variables defined in the variable_mapping table. The value_mapping table defines how the data values from the source system are translated into standardized OMOP concepts. This variable-value pair

structure is commonly used to map data to OMOP tables like Observation and Measurement, where field names are mapped to `observation_concept_id` or `measurement_concept_id`, and their corresponding values are mapped to the respective `value_as_concept_id` fields

For cases where we need to map specific values without directly mapping the field names, the `variable_value_mapping` table is employed. This table is useful when only the values in the source data require mapping, and the source field names are not necessary. This kind of mapping typically applies to fields such as `observation_concept_id` or `condition_concept_id`, where the values from the source data are directly mapped to OMOP's target concept IDs without needing to consider the field name.

Finally, the `unit_mapping` table is used to standardize units of measurement that accompany the values in the `value_mapping` table. This table contains the mappings for units if present or required in the source data. If a value in the source system comes with a unit, the `unit_mapping` ensures that it is correctly translated into OMOP by mapping it to the relevant `unit_concept_id` field, commonly found in the Measurement table. This allows the data to remain accurate and complete, with both the values and their associated units being standardized appropriately.

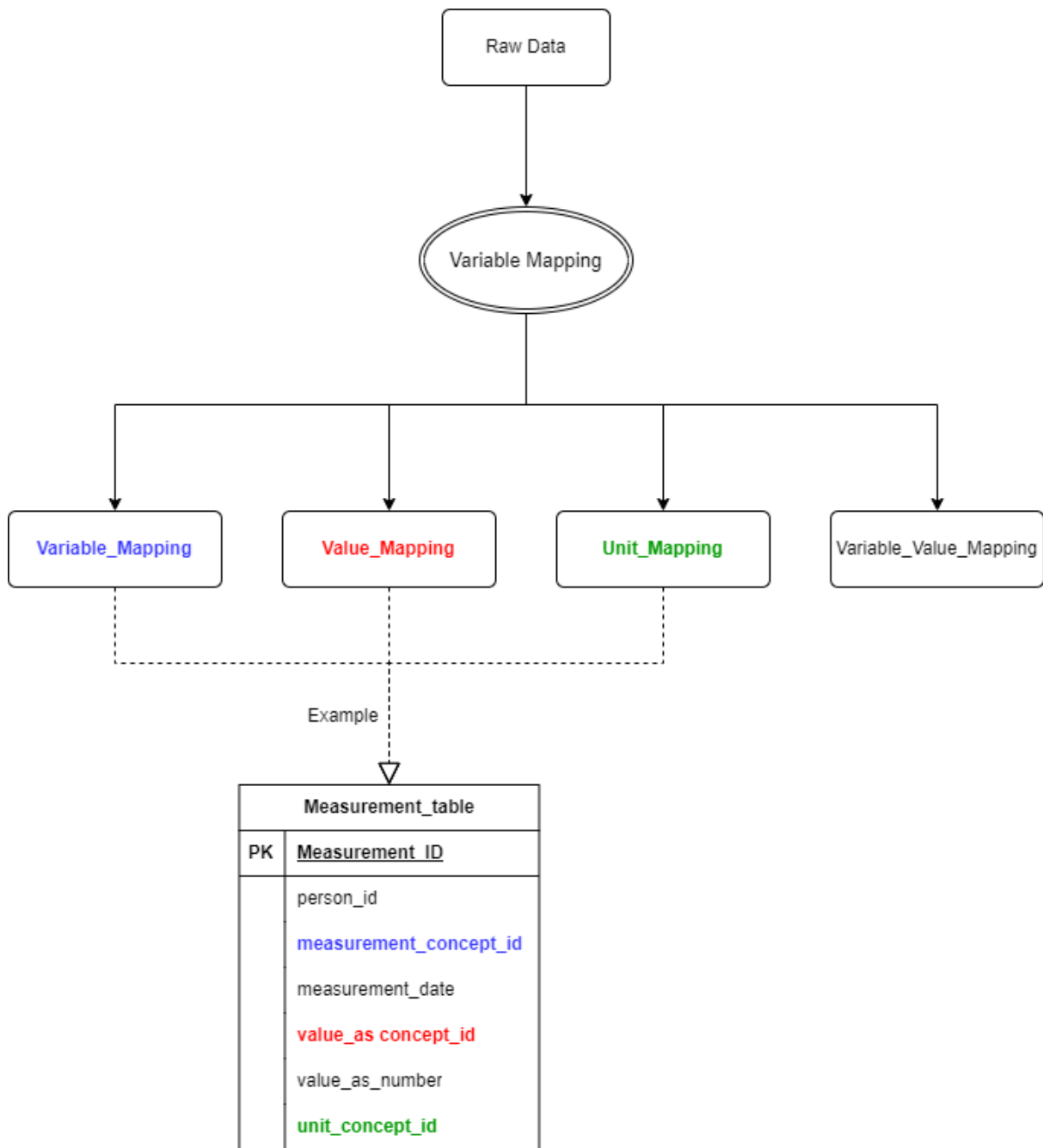


Figure 3: Flowchart of the semantic separation of concepts and how they are mapped into the OMOP CDM

Data Transformation

Step 8: Python Environment Setup

For the Python development, we used Visual Studio Code as the primary IDE, along with Docker Compose to create a consistent environment with Dockerised components. PostgreSQL was employed for the storage of data both before and after its transformation into the OMOP CDM format.

Step 9: Data Cleaning and ETL

Before the ETL process, data cleaning was performed to address inconsistencies, missing values, and incorrect formats in the raw data. This step ensured that all data proceeded to the transformation stage will be consistent and correct, improving the reliability of the final harmonized dataset. After this process, we were ready to begin the ETL (Extract, Transform, Load) procedure.

During the ETL process, the STEM table played a crucial role in bridging the gap between raw data and the OMOP CDM. The STEM table, which is not part of the OMOP CDM, served as an intermediary that captured and organized key variables from the raw data before they were fully mapped to their respective OMOP CDM tables. Mappings of source terms to the correct OMOP CDM table is done automatically, based on the semantic mappings of the source terms and its domain. The following image is a visual description of the process:

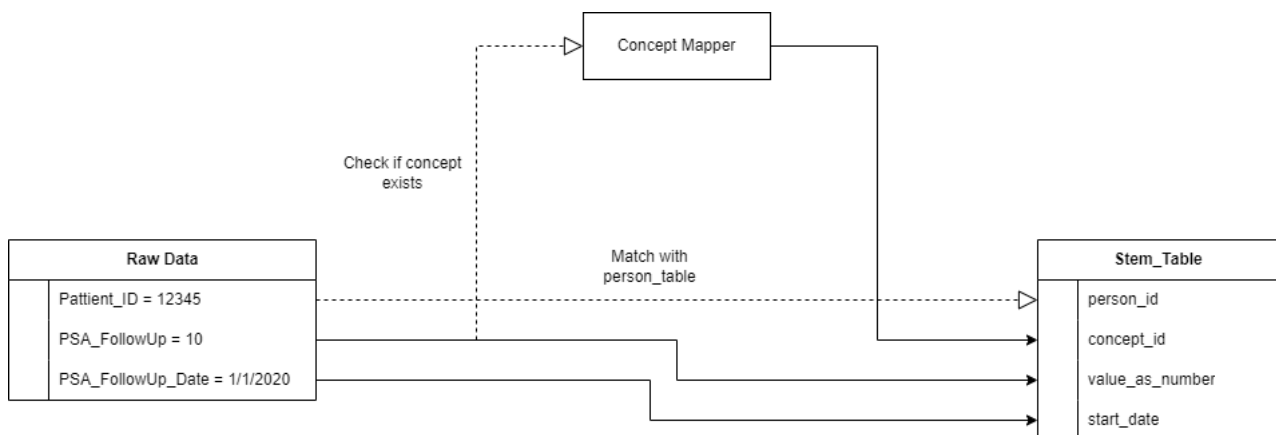


Figure 4: Diagram showing the process where raw clinical data is mapped through a Concept Mapper to the Stem_Table

Additional Steps and Final Transformation

Step 10: Post-Processing and SQL Scripts

In this step, we focus on transforming the data stored in the STEM table into the proper OMOP CDM tables by executing carefully constructed SQL scripts. These scripts are designed to map and migrate the data from the STEM table into its appropriate OMOP tables, ensuring that every record is placed in the correct table of the OMOP schema. To do this, we execute an SQL script for each OMOP table that we want to populate. Each SQL script follows a defined workflow: it reads records from the STEM table and based on the mappings derived from the concept table, assigns the data to its respective OMOP CDM table, such as Observation, Measurement, Condition Occurrence etc. The process begins by identifying the domain of each concept using a join with the Concept Table and the domain_id field. For example, if a record's concept ID has the domain_id of "Measurement", the SQL script will route that record to the Measurement table. Likewise, records with concepts related to observations or conditions are funneled into the appropriate Observation or Condition Occurrence tables, respectively.

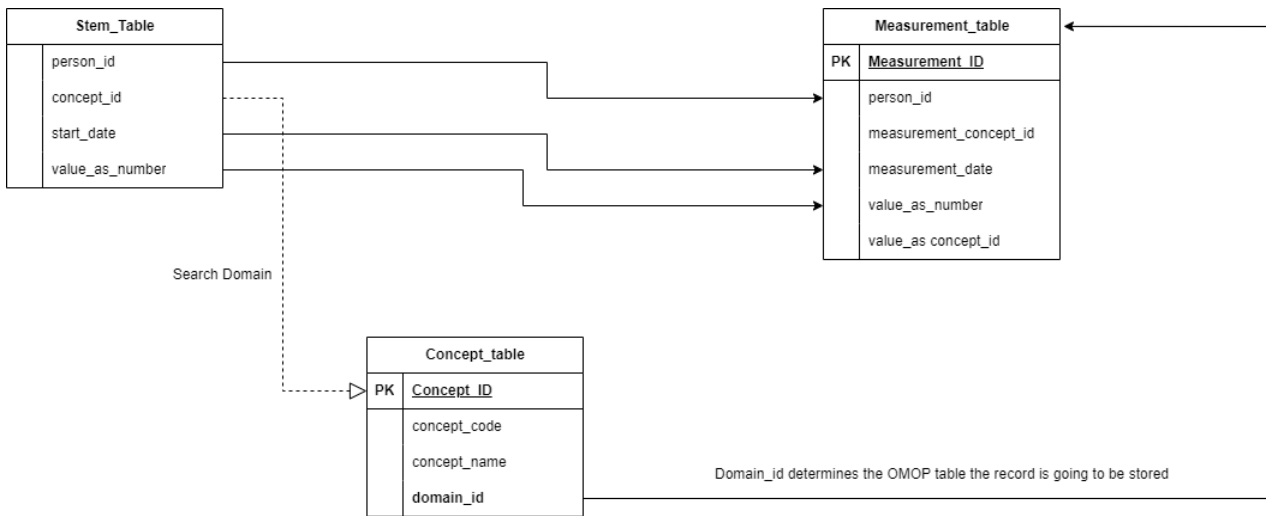


Figure 5: Diagram showing the post-processing procedure where data from the Stem_Table is transformed and stored in the standardized tables of OMOP CDM

Step 11: Data Quality Assessment

The final step in the process is the Data Quality Assessment. For this step we use the Data Quality Dashboard (DQD). This step certifies the integrity and reliability of the transformed data that now reside within the OMOP CDM. The Data Quality Dashboard is a powerful tool that executes a comprehensive series of data quality checks including the existence of required fields, if foreign key relationships between tables are valid, the confirmation that the concept mappings are correct, and that there are no anomalies such as missing or duplicate records. The result is a visual description of both the successful and the failed tests as well as an analytical review of the potential data issues.

DATA QUALITY ASSESSMENT

PHASE-IV, KU LEUVEN PROSTATE CANCER USE CASE

DataQualityDashboard Version: 2.6.1
 Results generated at 2024-10-14 11:49:58 in 3 mins

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	755	83	838	90%	375	0	375	100%	1130	83	1213	93%
Conformance	1150	226	1376	84%	179	31	210	85%	1329	257	1586	84%
Completeness	583	110	693	84%	22	1	23	96%	605	111	716	84%
Total	2488	419	2907	86%	576	32	608	95%	3064	451	3515	87%

Figure 6: Data Quality Assessment results of the KUL Prostate Cancer subset

5.3.2 Feature Alignment

During the architecture design and the harmonization needs phase in the alignment phase of the two datasets coming from UTU and KUL and the protocol submission it was identified the lack of one variable of interest that is related to the presence of 5-ari medication.

According to the objective of predicting biopsy outcomes, the variable ARI-5, which indicates whether the patient has taken 5 α -reductase inhibitors (5-ARIs), is essential because these medications can lead to a drop in PSA (Prostate Specific Antigen) levels. 5-ARIs, such as finasteride and dutasteride, work by reducing the production of dihydrotestosterone (DHT), which in turn reduces prostate size. Since PSA is produced by the prostate, the use of these medications typically causes a significant reduction in PSA levels, often by as much as 50%.

When predicting biopsy outcomes, PSA levels are a key factor, but if a patient is taking a 5-ARI, the interpretation of PSA must be adjusted. Without accounting for ARI-5, one might incorrectly assume that a low PSA level suggests a reduced risk of prostate cancer or benign prostatic hyperplasia (BPH), when it could simply be the effect of the medication. Therefore, ARI-5 is critical for ensuring that PSA levels are interpreted accurately in the context of the biopsy outcome prediction.

The dataset from KUL does not include a separate ARI-5 variable and it shall be extracted from the medication log of the patients looking for medications such as finasteride and dutasteride.

6. Conclusions

The document outlines the data harmonization framework established for the PHASE IV AI project. The data harmonization requirements for the available datasets were identified and reported. Based on these requirements the data harmonization specifications were defined both for tabular and imaging data. Moreover, the suitable strategies and pipelines were designed and developed to address both intra- (DaaS) and inter-partner (MaaS) harmonization needs. Preliminary results from the harmonization implementation were documented, while an initial documentation plan was developed to capture all the necessary information and metadata related to the harmonization process.

7. References

- [1] <https://www.cprd.com/>
- [2] <https://www.auria.fi/tietopalvelu/en/tutkijalle/>
- [3] <https://www.cancerimagingarchive.net/collection/lidc-idri/>
- [4] Draulans C, Everaerts W, Isebaert S, Van Bruwaene S, Gevaert T, Oyen R, Joniau S, Lerut E, De Wever L, Laenen A, Weynand B, Defraene G, Vanhoutte E, De Meerleer G, Haustermans K. Development and External Validation of a Multiparametric Magnetic Resonance Imaging and International Society of Urological Pathology Based Add-On Prediction Tool to Identify Prostate Cancer Candidates for Pelvic Lymph Node Dissection. *J Urol*. 2020 Apr;203(4):713-718. doi: 10.1097/JU.0000000000000652. Epub 2019 Nov 13. PMID: 31718396.
- [5] Ettala, Otto, Ivan Jambor, Ileana Montoya Perez, Marjo Seppänen, Antti Kaipia, Heikki Seikkula, Kari T. Syvänen et al. "Individualised non-contrast MRI-based risk estimation and shared decision-making in men with a suspicion of prostate cancer: protocol for multicentre randomised controlled trial (multi-IMPROD V. 2.0)." *BMJ open* 12, no. 4 (2022): e053118.
- [6] Giorgio Gandaglia, et al "Clinical Characterization of Patients Diagnosed with Prostate Cancer and Undergoing Conservative Management: A PIONEER Analysis Based on Big Data," *European Urology*, 2023.
- [7] [OMOP CDM v5.4 \(ohdsi.github.io\)](https://github.com/ohdsi/omop-cdm-v5.4)
- [8] [Anatomical Tracings of Lesions After Stroke \(ATLAS\)](#)
- [9] Liew, Sook-Lei, et al. "A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms." *Scientific data* 9.1 (2022): 320.
- [10] ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset
- [11] Riedel, Evamaria O., et al. "ISLES 2024: The first longitudinal multimodal multi-center real-world dataset in (sub-) acute stroke." *arXiv preprint arXiv:2408.11142* (2024).
- [12] Liang, Kongming, et al. "Symmetry-enhanced attention network for acute ischemic infarct segmentation with non-contrast CT images." *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24*. Springer International Publishing, 2021.
- [13] [OHDSI – Observational Health Data Sciences and Informatics](#)
- [14] <https://github.com/thehyve/ohdsi-etl-prias>