Journal of Biomedical Informatics 155 (2024) 104661

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

Data harmonization and federated learning for multi-cohort dementia research using the OMOP common data model: A Netherlands consortium of dementia cohorts case study

Pedro Mateus [a,*], Justine Moonen [b,c], Magdalena Beran [d,e], Eva Jaarsma [f,g], Sophie M. van der Landen [b,c], Joost Heuvelink [b], Mahlet Birhanu [h], Alexander G.J. Harms [h], Esther Bron [h], Frank J. Wolters [i], Davy Cats [j], Hailiang Mei [j], Julie Oomens [k], Willemijn Jansen [k], Miranda T. Schram [l,m,n,o], Andre Dekker [a], Inigo Bermejo [a]

[a] *Department of Radiation Oncology (Maastro), GROW School for Oncology and Reproduction, Maastricht University Medical Centre+, Maastricht, Netherlands*
[b] *Alzheimer Center Amsterdam, Neurology, Vrije Universiteit Amsterdam, Amsterdam UMC location VUmc, Amsterdam, Netherlands*
[c] *Amsterdam Neuroscience, Neurodegeneration, Amsterdam, Netherlands*
[d] *Department of Internal Medicine, School for Cardiovascular Diseases (CARIM), Maastricht University, Maastricht, Netherlands*
[e] *Department of Epidemiology and Global Health, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, Netherlands*
[f] *Center for Nutrition, Prevention, and Health Services, National Institute for Public Health and the Environment (RIVM), Bilthoven, Netherlands*
[g] *Amsterdam UMC location Vrije Universiteit Amsterdam, Epidemiology and Data Science, Amsterdam, Netherlands*
[h] *Biomedical Imaging Group Rotterdam, Dept. Radiology & Nuclear Medicine, Erasmus MC - University Medical Center Rotterdam, Rotterdam, Netherlands*
[i] *Erasmus MC – University Medical Centre Rotterdam, Departments of Epidemiology and Radiology & Nuclear Medicine, Netherlands*
[j] *Sequencing Analysis Support Core, Department of Biomedical Data Sciences, Leiden University Medical Center, Netherlands*
[k] *Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Alzheimer Center Limburg, Maastricht University, Netherlands*
[l] *Cardiovascular Research Institute Maastricht (CARIM), Maastricht University, Maastricht, Netherlands*
[m] *Department of Internal Medicine, Maastricht University Medical Centre, Maastricht, Netherlands*
[n] *MHeNS School for Mental Health and Neuroscience, Maastricht University, Maastricht, Netherlands*
[o] *Heart and Vascular Center, Maastricht University Medical Center+, Maastricht, Netherlands*

ARTICLE INFO

Keywords:
Data harmonization
Cohort studies
ETL
OMOP
CDM
Federated learning

ABSTRACT

Background: Establishing collaborations between cohort studies has been fundamental for progress in health research. However, such collaborations are hampered by heterogeneous data representations across cohorts and legal constraints to data sharing. The first arises from a lack of consensus in standards of data collection and representation across cohort studies and is usually tackled by applying data harmonization processes. The second is increasingly important due to raised awareness for privacy protection and stricter regulations, such as the GDPR. Federated learning has emerged as a privacy-preserving alternative to transferring data between institutions through analyzing data in a decentralized manner.

Methods: In this study, we set up a federated learning infrastructure for a consortium of nine Dutch cohorts with appropriate data available to the etiology of dementia, including an extract, transform, and load (ETL) pipeline for data harmonization. Additionally, we assessed the challenges of transforming and standardizing cohort data using the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) and evaluated our tool in one of the cohorts employing federated algorithms.

Results: We successfully applied our ETL tool and observed a complete coverage of the cohorts' data by the OMOP CDM. The OMOP CDM facilitated the data representation and standardization, but we identified limitations for cohort-specific data fields and in the scope of the vocabularies available. Specific challenges arise in a multi-cohort federated collaboration due to technical constraints in local environments, data heterogeneity, and lack of direct access to the data.

Abbreviations: NCDC, Netherlands Consortium of Dementia Cohorts; OMOP, Observational Medical Outcomes Partnership; CDM, Common Data Model; OHDSI, Observational Health Data Sciences and Informatics; ETL, Extract, Transform and Load; FAIR, Findable, Accessible, Interoperable, Reusable.

[*] Corresponding author.
E-mail address: pedro.mateus@maastro.nl (P. Mateus).

https://doi.org/10.1016/j.jbi.2024.104661
Received 29 January 2024; Received in revised form 22 May 2024; Accepted 23 May 2024
Available online 26 May 2024

Conclusion: In this article, we describe the solutions to these challenges and limitations encountered in our study. Our study shows the potential of federated learning as a privacy-preserving solution for multi-cohort studies that enhance reproducibility and reuse of both data and analyses.

1. Introduction

Cohort studies have been fundamental in the understanding of disease etiology, progression, and the effect of exposure to numerous risk factors. As in many fields, the design of these studies evolved to take advantage of more comprehensive measurements of the participant's condition. This resulted in larger and more complex data systems [1]. Cohort studies often engage in national or international collaborations to enhance external validity and statistical power. Although these collaborations improve the evaluation of research questions in larger subgroups, they face numerous barriers. These include the inability to share data (to protect privacy) and data ownership as well as challenges in the harmonization and standardization of the datasets [2].

In the current environment, techniques that enable data sharing in a privacy-preserving manner have gathered increasing interest. Federated learning is a prominent example in this field [3]. Federated learning protects data by analyzing individual-level data at each location and aggregating the local analyses' results (often iteratively). This avoids the need to transfer the individual-level data between locations. It also ensures that potentially identifiable clinical information remains safely within the data-owner's network. For this purpose, federated learning requires algorithms that are adapted to perform analyses in a decentralized way. In addition, it requires a distributed computing infrastructure connected through a secure network. Setting up such a federated infrastructure demands efforts from a multi-disciplinary team to guarantee efficient privacy-preserving communication and synchronization of multiple organizations, which requires specific solutions [4]. Since the introduction of federated learning, the development of software resources [5] that assist in addressing these difficulties has increased. A growing number of studies provide examples of applications that demonstrate the feasibility and usefulness of this approach [3,4].

In the process of integrating data from different sources, one fundamental step that poses challenges and demands considerable effort is guaranteeing the homogeneity of the data [1]. The representation of cohort data varies significantly between centers due to a lack of consensus on a data model and different methodologies. This results in a heterogeneous landscape that semantically and structurally impedes integration. Defining and implementing standards for structuring and standardizing data can be challenging. However, efforts have been made to create common data models (CDM) to represent clinical data from particular or multiple domains [6]. The Observational Medical Outcomes Partnership (OMOP) CDM [7], developed by the Observational Health Data Sciences and Informatics (OHDSI) program, initially tackled the challenge of integrating data from multiple sources in the field of medical products. Nevertheless, it has rapidly evolved to support additional domains and integrate common representations. These facilitate the standardization of medical terms to guarantee structural and semantical compatibility between data sources. Moreover, by providing mechanisms to minimize the loss of information, the support of a large community, and the development of open-source applications for data exploration and transformation, OMOP has been widely adopted by diverse clinical databases [8]. Among the applications targeted for data transformation, WhiteRabbit and Rabbit-In-A-Hat [9] provide software to scan the source data and generate documentation for the mapping. In addition, OHDSI provides Achilles [1], a tool that characterizes OMOP databases through descriptive statistics within a data quality dashboard.

Transforming a dataset into a specific CDM can be a cumbersome process that requires careful planning. It usually incorporates an extract, transform, and load (ETL) process to harmonize the data. As reported by previous studies [10–12], in the first stage, the choice of a CDM should entail factors such as stakeholders' and participants' expectations, available tools, maintenance, and scalability. These are relevant considerations that can impact the success of the transformation. Furthermore, an ETL process focuses on providing a framework that assists in extracting the data from the source, transforming the information to the correct format (e.g., standardization), and loading the data in the CDM according to the required conditions. There has been an increasing number of studies [13–22] focusing on the development of ETL processes to transform clinical data, such as electronic health records (EHR), and biobank data, into the OMOP CDM. Overall, these studies identified challenges that can hinder the success of an ETL process to OMOP, mainly the incomplete coverage of medical terms, possible loss of information, and steep learning curve [12]. Notwithstanding these aspects, the transformation and practical use of the OMOP database highlighted the benefits of these efforts. It did so by promoting integration with other databases, access to open-source tools, and enhancing reproducible observational research [6,11].

Implementing a federated learning infrastructure will commonly come across the problem of data heterogeneity. Thus, it requires solutions to harmonize that data, facilitating the analysis across centers. Studies focused on the challenges of data harmonization to OMOP for cohort study data are limited and predominantly focused on the choice of a CDM [23] or applications that successfully employ OMOP as a data source [24]. In this work, we develop an ETL process and apply it across a group of nine cohorts to establish an interoperable federated network. We do it while guaranteeing that each center provides its data in a homogeneous manner, semantically and structurally, by following the OMOP CDM. Furthermore, we assess the main aspects experienced in the project that impact the development of an ETL tool and harmonizing the data. Finally, we evaluate our ETL tool in a comprehensive dataset by simulating the federated infrastructure and using federated algorithms to validate the correct transformation of the data.

1.1. Motivation

The Netherlands Consortium of Dementia Cohorts (NCDC) comprises nine prospective cohort studies from the Netherlands that collected data on cognitive decline and dementia. Together, these cohorts provide a population of over 40,000 participants for the study of dementia, covering population-based samples as well as (memory) clinic patients. Traditionally, in this setting, a team conducting an analysis can perform pooled data-analysis or create custom scripts for each cohort, perform the analysis locally, and combine the results through *meta*-analysis. However, these approaches can be ineffective due to the current growth in privacy protection regulations and the potential to hinder the reproducibility of the analysis and results. A collaboration of this nature combines heterogeneous data sources from cohort studies with different purposes, guidelines, and data representations. Considering these characteristics, the consortium planned efforts to improve the interoperability between cohorts, enhance data reusability by adhering to the FAIR principles, and develop a federated infrastructure to facilitate decentralized data analysis.

1.2. Statement of significance

1.2.1. Problem

Multi-cohort studies are hindered by inconsistent data representations, lack of standards, and privacy constraints to data sharing.

1.2.2. What is already known

Federated learning is a privacy-preserving alternative to traditional analysis methods that require data harmonization to enable interoperability between cohorts. However, manual harmonization or ETL tools available can lack generalizability, standardization, or suitability for cohort data.

1.2.3. What this paper adds

This study proposes an ETL process for multi-cohort studies developed and evaluated with a federated collaboration of 9 cohorts. In this process, data is standardized and transformed to the OMOP CDM. The challenges identified point to the need for community consensus for data models and standard vocabularies.

2. Methods

2.1. Data

The nine cohorts participating in the NCDC, described in Table 1, comprise more than 40,000 participants and collect information relevant to the study of dementia, such as clinical, lifestyle, and cognitive data, blood and plasma biomarkers, and magnetic resonance imaging scans. Although there is substantial overlap in the data collected across cohorts, data models and practices to store and represent the data are often heterogeneous. As a result, the data harmonization developments performed and described in this study were planned to facilitate the integration of these cohorts into a federated learning network.

In this article, we describe the process of data harmonization of all cohorts but place a special focus on the harmonization of one of the cohorts as an exemplary case, the Maastricht Study [25] (9188 participants aged between 40 and 75 years), an observational prospective population-based study focusing on the etiology of type 2 diabetes. This cohort takes part in NCDC, aiming to understand the acceleration in cognitive decline seen in individuals with type 2 diabetes. The dataset for this cohort was made available as a single long-format SPSS (Statistical Package for the Social Sciences) file. Additionally, we created a synthetic dataset using an open-source tool based on conditional generative adversarial networks [26]. We generated 250 examples based on the source data and used it to perform integration testing and experiment with the developed tool.

Table 1
Description of the NCDC cohorts participating in the federated network.

Cohort	Description	Number of participants	Frequency of follow-up	Start
Amsterdam dementia cohort [27] (ADC)	Neurodegenerative dementias (clinical)	6,000	1 year	2004
Doetinchem Cohort Study [28] (DCS)	Lifestyle risk factors	4,300	5 years	1987
EMIF-AD 90 + study [29]	Cognitive impairment	129	1 year	2016
EMIF-AD PreclinAD study [30]	Amyloid pathology and cognitive decline in monozygotic twins	204	2 years	2014
Leiden longevity study [31] (LLS)	Ageing and longevity	3,359	10 years	2003
Longitudinal Aging Study Amsterdam [32] (LASA)	Ageing and longevity	4,000	3–4 years	1992
The Maastricht Study [25]	Type 2 diabetes mellitus, dementia, depression, and other chronic conditions	9,188	> 10 years	2010
Rotterdam Study [33]	Chronic diseases in middle-aged and elderly persons	14,926	3–4 years	1990
SMART[34]	Cardiovascular disease progression (clinical)	1,309	5 years	2001

2.2. Common data model

We employed the OMOP CDM [35] (version 6.0) to represent the information across the organizations with a homogeneous syntax and semantics. This person-centric model encompasses 39 tables that can assimilate clinical data, standard vocabularies, and additional metadata. We mapped each organization's data to the OMOP CDM using 12 of the 39 tables while standardizing the variable names, operators, units, and values using SNOMED [36], OMOP Gender, and UCUM [37] vocabularies.

In OMOP, representing a medical term requires a concept that uniquely defines it using codes and can be derived from international vocabularies. As observed in previous studies [10,13,15,17], obtaining complete coverage by standardized vocabularies of the medical terms used in different domains is challenging. To address this problem, when failing to find a suitable standard concept, we followed the OHDSI recommendations by creating a new concept in the OMOP standardized vocabulary tables, with an ID above 2 billion to avoid conflicts with the existing terminology. In addition, we linked these concepts to the EMIF-AD ontology [38] by using the optional field "concept_code" ("Concept" table), designed to represent the concept identifier from the source vocabulary. When creating a new concept, we established a short-term based on the description provided by the consortium and identified the OMOP domain using the definitions from the OHDSI documentation. We then stored the concept code from the ontology in the "concept_code" field, providing a link to a definition that currently cannot be specified in the OMOP vocabulary tables.

Although a common data model generally provides a comprehensive structure, challenges can arise when adapting the data. In particular, with applications outside the core target of the OMOP CDM, such as cohort data representation. In our work, we encountered three central difficulties concerning the representation of 1) the absence of a medical condition, 2) the observation period, and 3) non-patient data (illustrated in Supplementary Figs. 1 and 2).

1) The OMOP CDM includes the Condition table to represent the presence of medical conditions. However, it does not provide a straightforward method to include information on the absence of a medical condition. We did not find a consensual solution proposed in the OMOP documentation or literature. Moreover, based on the OHDSI public forum (https://forums.ohdsi.org/t/negative-information-in-omop-cdm/4923), the current position of the developers is to maintain such a design. Since this data is essential in epidemiology research, we stored these cases in the Observation table with the respective condition concept ID as the observation concept ID and a concept ID representing "Absence of" (SNOMED concept ID 4132135) as its value.
2) In the OMOP CDM, the "Person" and "Observation_period" tables are mandatory to identify the patients and the period for the events registered. However, as previously recognized elsewhere [13,15,19,39], it can be challenging to assign data into observation periods, as defined in the OMOP CDM. For example, data from longitudinal cohort studies usually characterizes the periods of time not individually but at a cohort level, as a representation of the interval taken to perform the necessary measurements and observations for all participants ('waves'). The resulting data does not guarantee an accurate observation period for each participant, making it difficult to determine the start and end date. In order to address this

limitation, we employed the "Observation_period" table to store the overall interval of each 'wave' and the "datetime" field of each observation, measurement, or condition to store, when available, the individual dates. Additionally, we duplicated the observation interval in the "Visit Occurrence" table by creating a visit for each participant and linking it with the clinical observations, measurements, or conditions occurring in that period.

3) Representing the cohort's data may entail additional information that is not direct observations or measurements of a participant, such as specifications on imaging equipment or cognitive assessments. These data items are provided at a cohort level and not per participant, creating a challenge in the OMOP CDM since this model follows a person-centric model and lacks a solution to represent this data accordingly. Therefore, we included the option to use the additional columns available for each record in the database ("observation_source_value", "measurement_source_value", and "condition_source_value") to link the information. This decision does not interfere with storing the verbatim values from the cohort data in the OMOP database.

Setting up a federated infrastructure may require the representation of multiple cohorts in one institution. Although the OMOP CDM provides a table to define cohorts, its purpose is to identify subsets of patients according to particular criteria, such as being diagnosed with a specific condition. Notwithstanding, the model allows linking each patient with a care site, describing the principal healthcare provider. Based on these characteristics, we represented the cohort metadata (name, responsible institution, and description) using the "Care_Site" table and linked each participant with the respective cohort through the "care_site_id". Since multiple care sites can be represented, this decision allows institutions to represent multiple cohorts in a single OMOP database. Additionally, we employed the "CDM_SOURCE" table, designed to store information on the data harmonization process, to register the ETL GitHub repository, tool version, and vocabulary version. Another aspect of cohort studies is the information on the reasons for missing data. Such details are generally recorded in longitudinal cohort studies. However, the OMOP CDM does not have a solution to store this information. Due to this limitation and the lack of a straightforward strategy, we did not include such details when harmonizing the data.

2.3. OHDSI software tools

The tools developed by OHDSI for data exploration, conversion, and standardization mainly target ETL processes concerning one central dataset. In a federated network, these tools are hindered by potential software installation restrictions at each site, the constraint of accessing the dataset only locally, and requiring training for the researcher supporting the ETL process in the cohort.

In our work, we employed two OHDSI tools for data standardization, Athena [40] and Usagi [41], to identify the concepts that correctly describe the medical terms. The first provides a web-search portal to explore matching concepts and download the standard vocabularies. The second assists in creating a mapping to standard concepts and encountering matching concepts by taking advantage of text similarities. To standardize the variables in our project, we first created a table with the variable names and a description provided by a researcher on the field. We then used this table as input for Usagi and manually selected the best match for our variable based on the similarity score and the metadata provided in Athena. In case of no satisfactory options, we created a new concept as described in the previous section.

2.4. ETL

Ensuring data harmonization between the cohorts is crucial for a federated learning network. For this purpose, we developed an ETL process to harmonize each cohort's data into a common schema using standard representations. Although in our approach, the OMOP CDM provides the schema followed in the transformation, we believe the steps described can be agnostic and applicable to most common data models.

2.4.1. Metadata collection

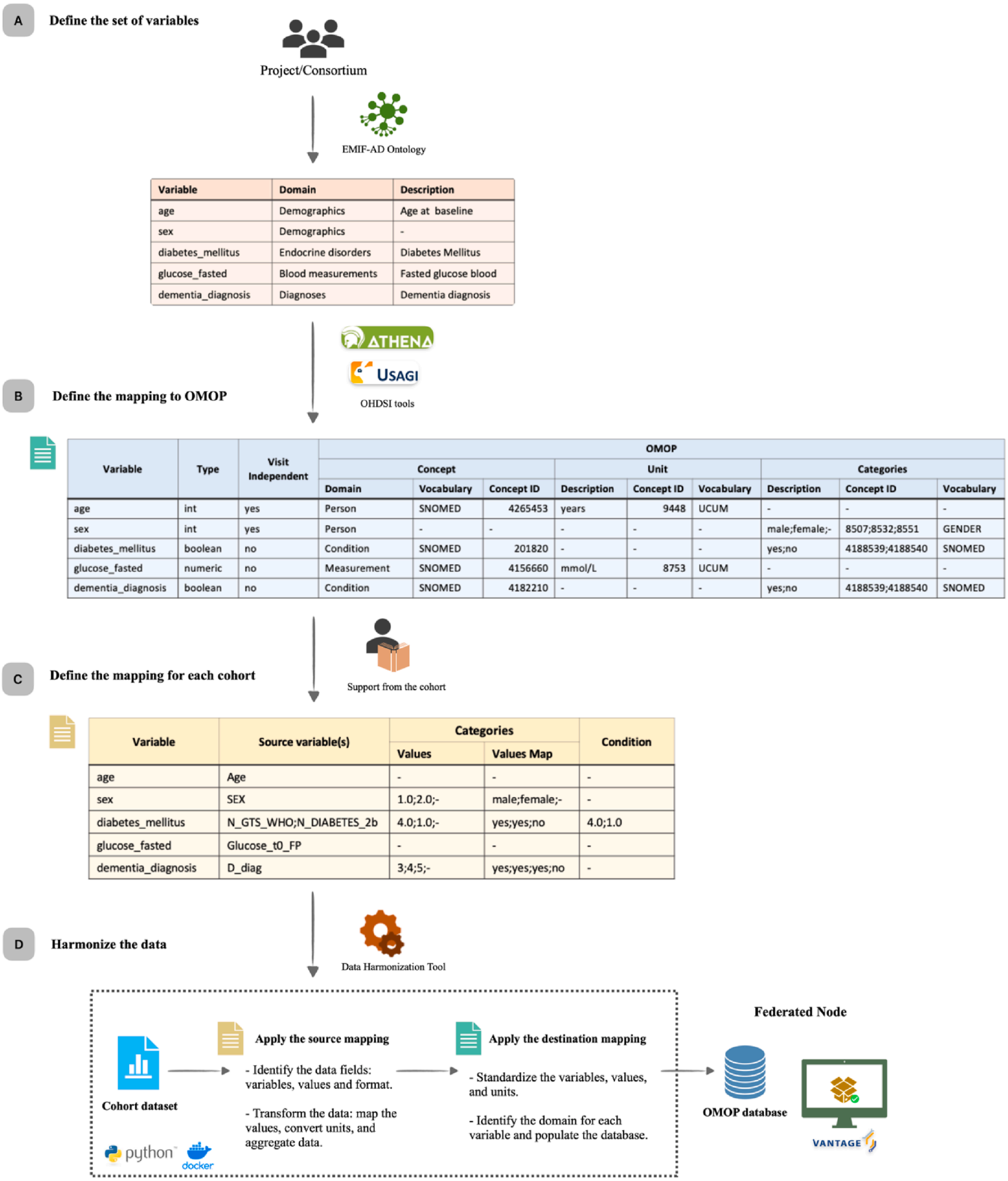
At the initial stage, the consortium defines relevant research questions and the set of variables needed to answer these questions. To characterize this selection, we start by collecting information that defines each variable, such as the type of data (i.e., numeric or categorical), range of values or categories, and units. Additionally, if available, each variable is linked to the EMIF-AD ontology [38] concept for complete characterization. Based on these attributes, we create a consortium-level mapping that defines the correspondence between the relevant variables and their representation within the OMOP CDM (the "destination mapping"). This information is configured with a CSV (comma-separated values) file containing one row per variable with 11 fields (described in Table 1 of the Supplementary Material). The mapping entails identifying the domain (corresponding OMOP table) and standardizing the variable name, range of values or categories, and units using the concepts from the OMOP vocabularies (Fig. 1A–B). Overall, this mapping characterizes the data representation that will become available in all institutions and employed when performing federated analyses.

The complete characterization of the selected variables also defines the guidelines for developing the mapping of each cohort (the "source mapping"). In this process (Fig. 1C), we identify the variables and corresponding categories from the source dataset and map them to the equivalent identifiers from the consortium-level mapping. To accomplish this, a researcher from each institution with knowledge of the cohort's data provides the necessary information for the variable mapping with additional support from the available cohort codebooks. The cohort mapping is agnostic to the CDM selected and exclusively characterizes the cohort's data. Additionally, we include the necessary transformations, such as unit conversion, to minimize the pre-processing of the dataset and facilitate the data extraction process. The result is a CSV file with 12 fields (described in Table 2 of the Supplementary Material), where each row represents the mapping and necessary transformations for each variable using metadata.

2.4.2. Data transformation

We developed a command-line interface (CLI) to transform the data based on the information collected in the source and destination mappings (Fig. 1D). This tool sets up a database following the OMOP schema, inserts the vocabularies, and populates the database with the harmonized data. Moreover, it provides summary statistics to assess the correctness of the transformation. In practice, we perform this process locally for each cohort, assisted by a researcher from the institution with access to the dataset. For support, we provided documentation and directly guided the data harmonization in each cohort via virtual meetings. We did not perform formal training on the OMOP CDM but promoted it through introductory presentations and assisted researchers when implementing their analysis. The ETL tool transforms each row in the dataset sequentially. It starts by obtaining the cohort-specific metadata for each variable from the source mapping. Then, it identifies the correct target table and concept for each variable in the OMOP database using the destination mapping. Lastly, it extracts the data, performs the cohort-specific transformations, and populates the database.

Additionally, in order to facilitate the adoption by researchers who are used to working with a single data table and do not have experience with relational databases, we included the option to generate a single table from the OMOP database. This option does not require additional input and produces a long-format table using the EMIF-AD ontology to designate the columns. The ETL tool accomplishes this by sequentially retrieving the records from OMOP for each patient and employing the destination mapping to identify the column in the new table for each OMOP concept.


A
Define the set of variables
Project/Consortium
EMIF-AD Ontology
B
Define the mapping to OMOP
ATHENA
USAGI
OHDSI tools
C
Define the mapping for each cohort
Support from the cohort
D
Harmonize the data
Data Harmonization Tool
Cohort dataset
python
docker
Apply the source mapping
- Identify the data fields: variables, values and format.
- Transform the data: map the values, convert units, and aggregate data.
Apply the destination mapping
- Standardize the variables, values, and units.
- Identify the domain for each variable and populate the database.
Federated Node
OMOP database
VANTAGE

Variable	Domain	Description
age	Demographics	Age at baseline
sex	Demographics	-
diabetes_mellitus	Endocrine disorders	Diabetes Mellitus
glucose_fasted	Blood measurements	Fasted glucose blood
dementia_diagnosis	Diagnoses	Dementia diagnosis

Variable	Type	Visit Independent	OMOP								
			Concept			Unit			Categories		
			Domain	Vocabulary	Concept ID	Description	Concept ID	Vocabulary	Description	Concept ID	Vocabulary
age	int	yes	Person	SNOMED	4265453	years	9448	UCUM	-	-	-
sex	int	yes	Person	-	-	-	-	-	male;female;-	8507;8532;8551	GENDER
diabetes_mellitus	boolean	no	Condition	SNOMED	201820	-	-	-	yes;no	4188539;4188540	SNOMED
glucose_fasted	numeric	no	Measurement	SNOMED	4156660	mmol/L	8753	UCUM	-	-	-
dementia_diagnosis	boolean	no	Condition	SNOMED	4182210	-	-	-	yes;no	4188539;4188540	SNOMED

Variable	Source variable(s)	Categories		Condition
		Values	Values Map	
age	Age	-	-	-
sex	SEX	1.0;2.0;-	male;female;-	-
diabetes_mellitus	N_GTS_WHO;N_DIABETES_2b	4.0;1.0;-	yes;yes;no	4.0;1.0
glucose_fasted	Glucose_t0_FP	-	-	-
dementia_diagnosis	D_diag	3;4;5;-	yes;yes;yes;no	-

Fig. 1. ETL process plan to harmonize the consortium cohorts' data. The tables presented have been simplified to illustrate the most relevant fields.

The source code is publicly available on GitHub (https://github.com/MaastrichtU-CDS/omop-converter). It was written in Python 3.8.10 and developed to operate with a PostgreSQL database using SQL. It accepts file-based datasets in CSV, SPSS, or SAS (Statistical Analysis Software) formats as input. We used Docker to containerize and version the application, which includes the SQL statements to create the OMOP database schema and the CSV files, downloaded from Athena, representing the selected vocabularies. Additionally, we version the consortium-level and the cohorts' mapping information in our GitHub repository.

2.5. Federated learning infrastructure

In our work, we implemented a federated architecture based on the Personal Health Train [42] (PHT) using Vantage6 [43] version 2.1.0. This Python library facilitates the installation of nodes in each organization that communicate with a central server, responsible for handling the communications between the participants. Moreover, it provides the tools to manage the involved parties, secure communications, and execute analyses at each node. However, Vantage6 does not solve the problems arising from the heterogeneity of the data systems at each organization.

There are no specific requirements for the data source type when using Vantage6 since this needs to be addressed by the algorithm developer. In our work, we opted for a PostgreSQL database to store our data following the OMOP CDM definition. The algorithms developed relied on this database and retrieved the data using SQL queries. To evaluate the quality and functionality of the proposed framework, we implemented a federated summary statistics algorithm based on the OMOP CDM structure, detailed in the Supplementary Material.

The resulting architecture, presented in Fig. 2, allows researchers to analyze the data through authorized algorithms, containerized with Docker, that only provide non-individual level results.

2.6. Experiments

To evaluate the effectiveness of our tool and the quality of the transformation to OMOP, we assessed the coverage of the variables by valid concepts from the available vocabularies by counting how many variables have a valid concept from a standardized vocabulary to represent them. In addition, we used descriptive statistics to compare the source data and the OMOP tables. Additionally, we developed and employed a federated algorithm for summary statistics and cohort selection (detailed in the Supplementary Material) to demonstrate the functionality of the database in a federated architecture.

3. Results

We successfully applied our ETL tool to harmonize the data from the NCDC cohort studies to the OMOP CDM. In this process, we observed that the OMOP CDM can completely represent the data for the selected set of variables. However, the standardized vocabularies lack the terms to describe most variables.

3.1. Data harmonization challenges

We developed the ETL tool to address the challenges of harmonizing the data of multiple cohorts by structuring the different components in modules. As illustrated in Fig. 3, three main modules provide the necessary information to perform the ETL process: the consortium-level variable definition (destination mapping), the cohort-specific information (source mapping), and the CDM's requirements. We followed an iterative approach to develop and test our tool, initially using a representative synthetic dataset, followed by an initial trial with three cohorts, and ultimately applying the mapping to every cohort. In this process, we identified difficulties associated with the cohort's resources, data access, and the CDM that hindered the successful application of the ETL tool. Table 2 describes these challenges and the solutions implemented to tackle them.

3.2. Common data model evaluation

We have applied our ETL process and tool to the nine cohorts in the NCDC project, comprising approximately 40,000 participants, and successfully harmonized the data into the OMOP CDM. In total, we installed eight PostgreSQL databases (keeping two cohort studies from the same institution in a single database) at eight different locations and the necessary software to establish the intended federated system. The initial selection of relevant data for the project resulted in 201 variables considered for the data harmonization process. From this selection, we matched 62 (31 %) variables to standard concepts from the SNOMED vocabulary and created new concepts for the remaining 139 (69 %) variables. Overall, we represented the cohort data with four OMOP domains: the Person, Observation, Measurement, and Condition Occurrence tables. As shown in Fig. 4, the "Measurement" domain presented the highest percentage of unmapped variables (117 out of 140), where we found a particular difficulty with the cognitive screening tests domain, from which 71 out of the 72 variables remained unmapped. In the "Condition" domain, which had the lowest percentage of unmapped variables, four of the five unmapped variables were related to medical imaging assessments. Additionally, we completely mapped the

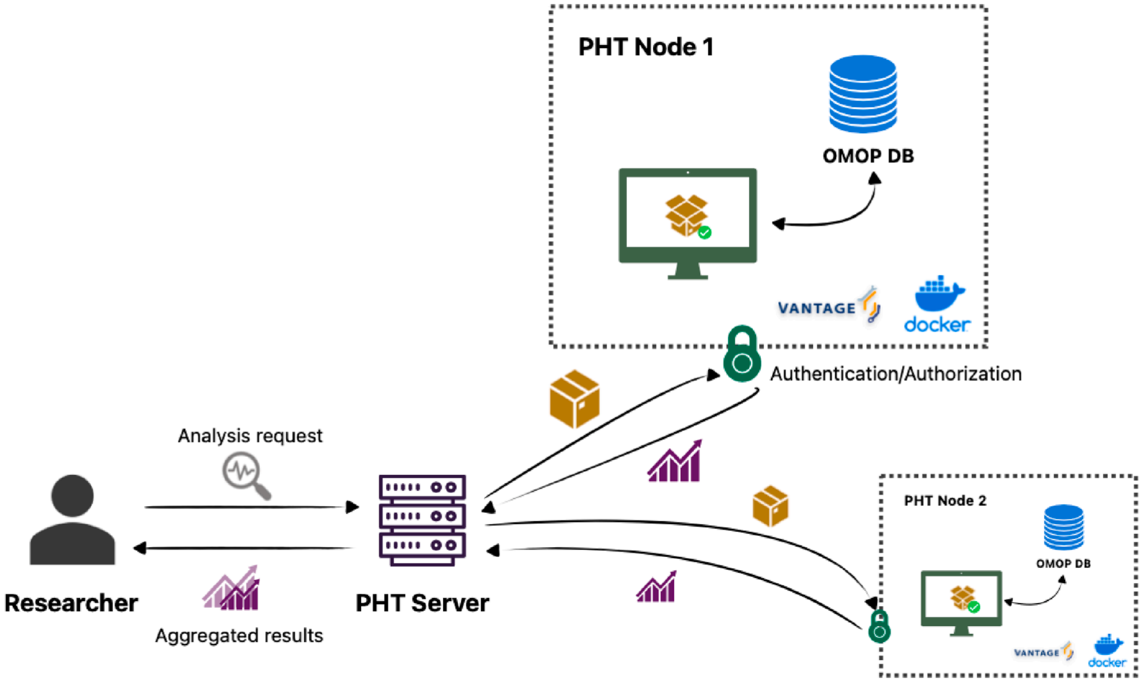


Fig. 2. Representation of the federated architecture based on the Personal Health Train (PHT) [42]. The PHT server mediates the communications between the PHT nodes and handles requests for analysis by authorized researchers.

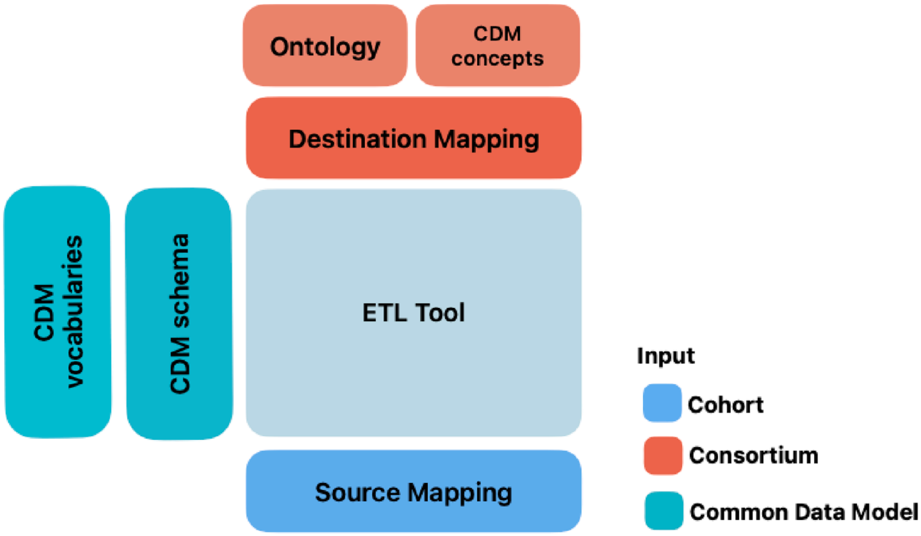


Fig. 3. ETL design illustration: the consortium (destination mapping), cohort (source mapping), and Common Data Model (schema and vocabularies) components are separated. The ETL tool uses these elements as input to harmonize the dataset provided.

Table 2
Description of relevant challenges that influenced the ETL process development.

Challenge	Solution
Estimating the technical resources available for each cohort can be challenging. *Data access methods, security rules, and software tools available vary greatly, in some cases only available behind secure environments, limiting the preparation of the ETL tool.*	Containerizing the application and developing the ETL tool iteratively.
Variability of the cohort data structure. *Commonly, cohorts generate a specific dataset for each research question, containing the necessary information for the analysis. However, the data characteristics and schema can vary between requests.*	Automating the necessary data transformations within the ETL tool and minimizing modifications of the dataset supplied before the ETL process. Documenting and versioning the cohort-specific transformations separately from the ETL tool.
No direct access to the data by the ETL tool developing team. *Data access is often restricted, especially in the federated learning context and completing the data harmonization process may rely on non-technical personnel following instructions.*	Using a synthetic dataset to develop the mappings and test the transformation. Minimizing the knowledge necessary to execute the ETL tool and reducing the interaction with technical frameworks, such as Python scripts. Solutions to simplify the interaction encompass CLIs and visual interfaces.
Complexity of the OMOP CDM relational structure. *The relational structure that characterizes the OMOP CDM is often more complex than the plain format data provided by cohort studies. This complexity can result in a steep learning curve for its users, negatively impacting the successful adoption of CDMs.*	Generate a plain table, in a long or wide format, from the data stored in the OMOP CDM format. Promote the integration of the CDM by adapting analysis algorithms and tools.

units and categorical values to standard concepts from the UCUM and SNOMED vocabularies.

3.3. ETL tool evaluation

We assessed the quality of the ETL tool by comparing the distribution of the harmonized variables between the source data and the OMOP database. We transformed the data from 3,807 participants considering 22 variables, resulting in 3,807 visits and 74,894 events (observations, measurements, and conditions) registered in the OMOP CDM database. The resulting summary statistics for the source data (Table 3), obtained using federated algorithms for summary statistics, matched precisely with the OMOP CDM and the EMIF-AD table. We observed consistency of the data for every variable after performing the harmonization, including the derived variables, such as the year of birth, calculated by the tool from the participant's age. Moreover, in Table 4, we calculated the expected number of rows in the OMOP tables based on the number of variables and missing information from the source dataset. In particular, the "Demographics" and "Risk factors" variables were mapped to the Observation domain, except for "year of birth" and "sex" which belong to the "Person" domain. The "Clinical measurements" and "Cognitive screening tests" were mapped to the Measurement domain and the "Clinical conditions" to the Condition Occurrence domain. Furthermore, the "Clinical conditions" were also included in the Observation domain to account for the number of entries representing the absence of a condition, as specified in the Methods. The results demonstrated the complete coverage of the source data by the OMOP tables.

4. Discussion

In this study, we have set up a federated learning network connecting nine cohort studies using a newly developed ETL process to harmonize and standardize cohort study data into the OMOP CDM. Following this approach allowed us to transform the data from nine cohorts with minimal loss of information, improve compliance with the FAIR data principles [44], and analyse the data in a decentralized setting using federated learning algorithms.

Developing an ETL process for a federated network required decoupling the project, cohort, and CDM-specific components, making the tool generalizable to new cohorts. This approach automates the interaction with the OMOP database, avoids incorporating database-specific metadata and standardization preferences, and improves versioning by allowing separate management of each component. Furthermore, it benefits the partition of tasks and responsibilities within the multidisciplinary team necessary to harmonize the data. Previous studies [13,14,16–22] focusing on ETL frameworks for a single database transformation commonly developed project-specific scripts or required a pre-processing step to a specified format, limiting its reusability. Although with similar constraints, a recent study [8] presented a metadata-driven framework, improving readability and manipulation of the data transformations. Notwithstanding the impact of project specificities in ETL development, efforts for agnostic tools can be fundamental to potentiate interoperability with new databases. Finally, it is relevant to acknowledge that these works focused on transforming

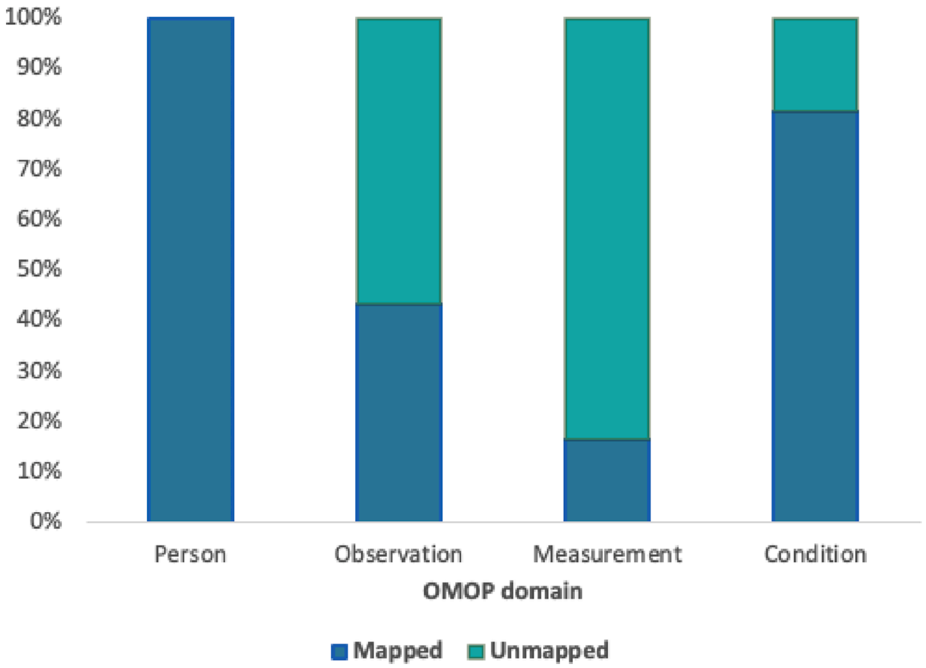


Fig. 4. Mapping of variables from the Netherlands Consortium of Dementia Cohorts to OHDSI standard concepts from the SNOMED vocabulary by OMOP domain.

Table 3
Summary statistics and mapping metadata for the 22 variables harmonized from the Maastricht Study source data. Mean and standard deviation were calculated for the continuous variables and frequency distribution for the categorical variables.

n	**Source Data**		**OMOP DB V6.0**		
	3807		3807		
	Frequency	Missing information	Concept ID[*]	Vocabulary	Domain[**]
Demographics					
Year of birth[***]	–	–	–	–	Person
Sex (female/male)	1857/1950	0	–	–	Person
Age at visit (years)	59.90 (8.34)	0	4265453	SNOMED	Observation
Education level[***] (low/high)	105/3567	135	4171617	SNOMED	Observation
Education level (low/medium/high)	1274/1037/1417	79	–	–	Observation
Risk factors					
Smoking behaviour (current/past/never)	520/1937/1300	50	4275495	SNOMED	Observation
Current smoker[***] (yes/no)	–	–	4298794	SNOMED	Observation
Current alcohol consumption (yes/no)	3044/715	48	4074035	SNOMED	Observation
Physical Activity (hours/week)	14.00 (8.15)	484	2000000417	EMIF-AD	Observation
History of hypertension (yes/no)	2180/1625	2	4058286	SNOMED	Observation
Hypertension medication (yes/no)	1533/2274	0	2000000494	EMIF-AD	Observation
Hypercholesterolemia medication (yes/no)	1387/2420	0	2000000359	EMIF-AD	Observation
History of cardiovascular disease (yes/no)	651/3089	67	4144290	SNOMED	Observation
Clinical measurements					
BMI (kg/m2)	27.09 (4.57)	3	4245997	SNOMED	Measurement
Waist circumference (cm)	95.91 (13.77)	4	4172830	SNOMED	Measurement
SBP (mmHg)	135.04 (18.08)	3	4152194	SNOMED	Measurement
DBP (mmHg)	75.99 (9.84)	3	4154790	SNOMED	Measurement
Cholesterol ratio	3.66 (1.16)	4	4195214	SNOMED	Measurement
Cognitive screening tests					
MMSE score	28.93 (1.27)	143	4169175	SNOMED	Measurement
Clinical conditions					
Depression (yes/no)	136/3500	171	440383	SNOMED	Condition
Diabetes type 1 (yes/no)	39/3768	0	201254	SNOMED	Condition
Diabetes type 2 (yes/no)	1078/2729	0	201826	SNOMED	Condition

[*] The OMOP standard concept ID for concepts from the SNOMED vocabulary; the EMIF-AD ontology concept code for terms without a standard concept from the vocabularies.

[**] Variables from the “Person” domain are core elements of the CDM and do not require a standard concept to represent them.

[***] Derived variables: “year of birth” was calculated from the “age” variable and “current smoker” from the “smoking behaviour”. Recoded variables: “Education level” was recoded from a source variable with eight levels.

clinical relational databases. In most cases, these databases have unique and more complex schemas than cohort data, which typically makes data extraction more challenging.

Addressing the harmonization of cohort study data presented unique challenges, notably the semantic heterogeneity between cohorts, the differences in IT resources available, and security measures to access and

Table 4
Assessment of the cohort's data coverage by the OMOP database.

		Source data	OMOP	Coverage
Observation	Demographics	11,207	–	–
	Risk factors	29,755	–	–
	Clinical conditions	9997	–	–
	Total	50,959	50,959	100 %
Measurement	Clinical measurements	19,018	–	–
	Cognitive screening tests	3664	–	–
	Total	22,682	22,682	100 %
Condition	Clinical conditions	1253	–	–
	Total	1253	1253	100 %

operate the datasets. In light of the planned data and applications, proven stability and scalability of relational databases [11], integration of standard vocabularies, and support from previous studies [24,45], we opted for the OMOP CDM to ensure network interoperability. Concerning the limitations due to diversity in cohort resources and security conventions, containerization proved an efficient method to minimize system dependencies and facilitate setting up and populating the databases. Additionally, it proved effective when incorporated into the federated infrastructure, facilitating the management and development of standardized algorithms due to the straightforward interface. Lastly, ensuring reproducibility can be challenging due to cohort-specific data management procedures and the legal agreements. Nevertheless, establishing versioning of the ETL components and including the possibility for periodic backups can facilitate this procedure and ensure its reproducibility.

The conversion of the cohorts' data into the OMOP CDM proved successful, with the results displaying a complete coverage of the data and correct transformation of the values. Notwithstanding the flexibility of OMOP, the implementation process and development of applications highlighted limitations and drawbacks to consider, such as the additional complexity that emerges with a relational schema and the standardization of terms. Compared to standard file-based datasets provided by cohorts to researchers, employing OMOP can result in a steep learning curve to understand and query the data, as previously reported [12]. To address this, we developed examples of queries and generated a single table from the OMOP database that facilitates the initial interaction with a relational database. Nonetheless, this table does not conform to an international community data standard, representing an instrument to promote familiarity with the federated platform and future transition to the OMOP CDM. Furthermore, we identified three matters that required specific solutions to avoid information loss, specifically the representation of negative medical diagnostics, observation periods, and additional specifications on imaging equipment or cognitive assessments. Although these decisions allowed for a more complete representation of the cohort data, they can potentially impact the interoperability with other OMOP databases. These findings reinforce the need for agreements within the community on how to represent cohort data in OMOP.

As observed in previous studies [10,17,20], mapping the terms into standard concepts proved difficult, mainly due to the lack of granularity and limited scope in the cognitive screening domain. Cohort information, such as cognitive test performance or questionnaire data, can unfold in multiple data elements for each record (e.g., the number of correct answers and the time taken may be necessary to characterize a single test). This level of detail is not available in the standard vocabularies included in the OMOP ecosystem. The solution requires creating project-specific concepts, which may hinder the benefit of future interoperability with external databases employing the OMOP CDM. Utilizing metadata-driven tools (e.g., ontologies) can facilitate future integration by linking and describing the new concepts. Nonetheless, community efforts to reach domain standardization consensuses are fundamental to promote this further. A comprehensive standard vocabulary with widespread acceptance could avoid project-specific solutions and facilitate interoperability between CDMs. In particular, contributing to the OHDSI standardized vocabularies creates an opportunity to enhance its sustainability for the OMOP CDM.

The federated network described has been used in practice to obtain summary statistics on cohort data, develop linear models, and train a deep learning model [46] between cohorts. The ETL process enabled such applications by establishing a homogeneous data representation across cohorts. Although this represents one central problem for establishing such collaborations, additional aspects can impact the success of this strategy. Namely, the diversity in measurements between cohorts (e.g., distinct standard cognitive tests to assess the same domain) affects the interoperability and requires data harmonization solutions beyond employing a CDM. Moreover, setting the legal contracts and attaining consensus proved demanding and time-consuming. Although partially due to a lack of past instances and the novelty of the federated approach, enhancing this phase is crucial to drive this method's effectiveness. Finally, federated learning introduces a challenge for researchers accustomed to directly accessing the cohorts' data. Future developments should consolidate the data quality guarantees, provide mature federated applications for data exploration, and facilitate interaction with the ETL tool (e.g., by developing a user interface).

In conclusion, the efforts for data harmonization greatly benefit a federated infrastructure by extending the interoperability to the analysis, improving reproducibility and future reuse. Although current CDMs allow for an almost complete representation of data, facilitating standardization and future collaborations, community efforts are still necessary to develop standard vocabularies with higher domain coverage. This work demonstrates the necessity of non project-specific ETL processes and the potential of federated learning as the future for multi-cohort studies.

Statement of significance

Problem: Multi-cohort studies are hindered by inconsistent data representations, lack of standards, and privacy constraints to data sharing.

What is already known: Federated learning is a privacy-preserving alternative to traditional analysis methods that require data harmonization to enable interoperability between cohorts. However, manual harmonization or ETL tools available can lack generalizability, standardization, or suitability for cohort data.

What this paper adds: This study proposes an ETL process for multi-cohort studies developed and evaluated with a federated collaboration of 9 cohorts. In this process, data is standardized and transformed to the OMOP CDM. The challenges identified point to the need for community consensus for data models and standard vocabularies.

Code availability

The custom code developed to harmonize the cohorts' data and perform the federated summary statistics is available at GitHub: https://github.com/MaastrichtU-CDS/omop-converter, https://github.com/pedro-cmat/v6-summary-omop-py, https://github.com/pedro-cmat/v6-summary-rdb-py.

CRediT authorship contribution statement

Pedro Mateus: Writing – review & editing, Writing – original draft, Software, Resources, Methodology, Data curation, Conceptualization. **Justine Moonen:** Writing – review & editing, Software, Resources, Data curation. **Magdalena Beran:** Writing – review & editing, Software, Data curation. **Eva Jaarsma:** Writing – review & editing, Software, Data curation. **Sophie M. van der Landen:** Writing – review & editing, Software, Data curation. **Joost Heuvelink:** Writing – review & editing, Software, Data curation. **Mahlet Birhanu:** Writing – review & editing, Software, Data curation. **Alexander G.J. Harms:** Writing – review &

editing, Software, Data curation. **Esther Bron:** Writing – review & editing, Resources. **Frank J. Wolters:** Writing – review & editing, Resources, Data curation. **Davy Cats:** Writing – review & editing, Software, Data curation. **Hailiang Mei:** Writing – review & editing, Resources. **Julie Oomens:** Writing – review & editing, Resources. **Willemijn Jansen:** Writing – review & editing, Resources. **Miranda T. Schram:** Writing – review & editing, Resources. **Andre Dekker:** Writing – review & editing, Supervision. **Inigo Bermejo:** Writing – review & editing, Writing – original draft, Supervision, Software, Resources, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors acknowledge all researchers of the Netherlands Consortium of Dementia Cohorts (NCDC), funded in the context of Deltaplan Dementie from ZonMW Memorabel (projectnr 73305095005) and Alzheimer Nederland.

FJW is supported by The Netherlands Organisation for Health Research and Development (project BIRD-NL, number 10510032120005), the Dutch Research Council (Veni 09150162010108), and the Alzheimer's Association (AARF-22-924982).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2024.104661.

References

[1] V. Ehrenstein, H. Nielsen, A.B. Pedersen, S.P. Johnsen, L. Pedersen, Clinical epidemiology in the era of big data: new opportunities, familiar challenges, CLEP 9 (2017) 245–250.

[2] T. Hulsen, et al., From big data to precision medicine, Front. Med. 6 (2019) 34.

[3] J. Xu, et al., Federated learning for healthcare informatics, J. Healthc. Inform. Res. 5 (2021) 1–19.

[4] S. Banabilah, M. Aloqaily, E. Alsayed, N. Malik, Y. Jararweh, Federated learning review: fundamentals, enabling technologies, and future applications, Inf. Process. Manag. 59 (2022) 103061.

[5] I. Kholod, et al., Open-source federated learning frameworks for IoT: a comparative review and analysis, Sensors 21 (2020) 167.

[6] E.A. Voss, et al., Feasibility and utility of applications of the common data model to multiple, disparate observational health databases, J. Am. Med. Inform. Assoc. 22 (2015) 553–564.

[7] G. Hripcsak, et al., Observational Health Data Sciences and Informatics (OHDSI): opportunities for Observational Researchers, Stud. Health Technol. Inform. 216 (2015) 574–578.

[8] J.C. Quiroz, et al., Extract, transform, load framework for the conversion of health databases to OMOP, PLoS One 17 (2022) e0266911.

[9] WhiteRabbit and Rabbit-In-A-Hat (Version 0.10.8). OHDSI. https://github.com/OHDSI/WhiteRabbit/.

[10] X. Zhou, et al., An evaluation of the THIN database in the OMOP common data model for active drug safety surveillance, Drug Saf. 36 (2013) 119–134.

[11] S.T. Rosenbloom, R.J. Carroll, J.L. Warner, M.E. Matheny, J.C. Denny, Representing knowledge consistently across health systems, Yearb. Med. Inform. 26 (2017) 139–147.

[12] B. Li, R. Tsui, How to improve the reuse of clinical data– openEHR and OMOP CDM, J. Phys. Conf. Ser. 1624 (2020) 032041.

[13] V. Papez, et al., Transforming and evaluating the UK Biobank to the OMOP Common Data Model for COVID-19 research and beyond, J. Am. Med. Inform. Assoc. 30 (2022) 103–111.

[14] Y. Yu, et al., Developing an ETL tool for converting the PCORnet CDM into the OMOP CDM to facilitate the COVID-19 data integration, J. Biomed. Inform. 127 (2022) 104002.

[15] S.M.K. Sathappan, et al., Transformation of electronic health records and questionnaire data to OMOP CDM: a feasibility study using SG_T2DM dataset, Appl. Clin. Inform. 12 (2021) 757–767.

[16] N. Paris, A. Lamer, A. Parrot, Transformation and evaluation of the MIMIC database in the OMOP common data model: development and usability study, JMIR Med. Inform. 9 (2021) e30970.

[17] J.G. Klann, M.A.H. Joss, K. Embree, S.N. Murphy, Data model harmonization for the All Of Us Research Program: Transforming i2b2 data into the OMOP common data model, PLoS One 14 (2019) e0212463.

[18] J.R. Almeida, L.B. Silva, I. Bos, P.J. Visser, J.L. Oliveira, A methodology for cohort harmonisation in multicentre clinical research, Inf. Med. Unlocked 27 (2021) 100760.

[19] A. Matcho, P. Ryan, D. Fife, C. Reich, Fidelity assessment of a clinical practice research datalink conversion to the OMOP common data model, Drug Saf. 37 (2014) 945–959.

[20] M. Oja, et al., Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model: lessons learned, JAMIA Open 6 (2023) ooad100.

[21] P. Biedermann, et al., Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases, BMC Med. Res. Method. 21 (2021) 238.

[22] D. Puttmann, N. De Keizer, R. Cornet, E. Van Der Zwan, F. Bakhshi-Raiez, FAIRifying a Quality Registry Using OMOP CDM: Challenges and Solutions, in: B. Séroussi, et al. (Eds.) Studies in Health Technology and Informatics, IOS Press, 2022. https://doi.org/10.3233/SHTI220476.

[23] F. Cremonesi, et al., The need for multimodal health data modeling: a practical approach for a federated-learning healthcare platform, J. Biomed. Inform. 141 (2023) 104338.

[24] G.H. Lee, et al., Feasibility study of federated learning on the distributed research network of OMOP common data model, Healthc. Inform. Res. 29 (2023) 168–173.

[25] M.T. Schram, et al., The Maastricht Study: an extensive phenotyping study on determinants of type 2 diabetes, its complications and its comorbidities, Eur. J. Epidemiol. 29 (2014) 439–451.

[26] C. Sun, J. Van Soest, M. Dumontier, Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy, J. Biomed. Inform. 143 (2023) 104404.

[27] W.M. Van Der Flier, et al., Optimizing patient care and research: the Amsterdam dementia cohort, JAD 41 (2014) 313–327.

[28] W. Verschuren, A. Blokstra, H. Picavet, H. Smit, Cohort profile: the doetinchem cohort study, Int. J. Epidemiol. 37 (2008) 1236–1241.

[29] N. Legdeur, et al., Resilience to cognitive impairment in the oldest-old: design of the EMIF-AD 90+ study, BMC Geriatr. 18 (2018) 289.

[30] E. Konijnenberg, et al., The EMIF-AD PreclinAD study: study design and baseline cohort overview, Alz Res Therapy 10 (2018) 75.

[31] M. Schoenmaker, et al., Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study, Eur. J. Hum. Genet. 14 (2006) 79–84.

[32] M. Huisman, et al., Cohort profile: the longitudinal aging study Amsterdam, Int. J. Epidemiol. 40 (2011) 868–876.

[33] M.M.B. Breteler, J.J. Claus, D.E. Grobbee, A. Hofman, Cardiovascular disease and distribution of cognitive function in elderly people: the Rotterdam study, BMJ 308 (1994) 1604–1608.

[34] A.P. Appelman, et al., Total cerebral blood flow, white matter lesions and brain atrophy: the SMART-MR study, J. Cereb. Blood Flow Metab. 28 (2008) 633–639.

[35] S.J. Reisinger, et al., Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases, J. Am. Med. Inform. Assoc. 17 (2010) 652–662.

[36] K.A. Spackman, K.E. Campbell, R.A. Côté, SNOMED RT: a reference terminology for health care, Proc AMIA Annu Fall Symp 640–644 (1997).

[37] G. Shadow, C.J. McDonald, The Unified Code for Units of Measure, 2009. https://link.springer.com/chapter/10.1007/978-3-319-98192-5_37.

[38] I. Bermejo, S. Vos European Medical Information Framework's (EMIF) Alzheimer's disease (AD) ontology, 2021.

[39] V. Papez, et al., Transforming and evaluating electronic health record disease phenotyping algorithms using the OMOP common data model: a case study in heart failure, JAMIA Open 4 (2021) ooab001.

[40] Athena (Version 1.11.0). OHDSI. https://github.com/OHDSI/Athena/.

[41] Schuemie, M. Usagi (Version 1.3.0). OHDSI. https://github.com/OHDSI/Usagi/.

[42] O. Beyan, et al., Distributed analytics on sensitive medical data: the personal health train, Data Intellegence 2 (2020) 96–107.

[43] A. Moncada-Torres, F. Martin, M. Sieswerda, J. Van Soest, G. Geleijnse, VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange, AMIA Annu. Symp. Proc. 2020 (2020) 870–877.

[44] M.D. Wilkinson, et al., The FAIR Guiding Principles for scientific data management and stewardship, Sci. Data 3 (2016) 160018.

[45] M. Garza, G. Del Fiol, J. Tenenbaum, A. Walden, M.N. Zozus, Evaluating common data models for use with a longitudinal community registry, J. Biomed. Inform. 64 (2016) 333–341.

[46] P. Mateus, et al., Federated BrainAge estimation from MRI: a proof of concept, Alzheimer's & Dementia 19 (2023) e076747.