Data harmonization and federated learning for multi-cohort dementia research using the OMOP CDM

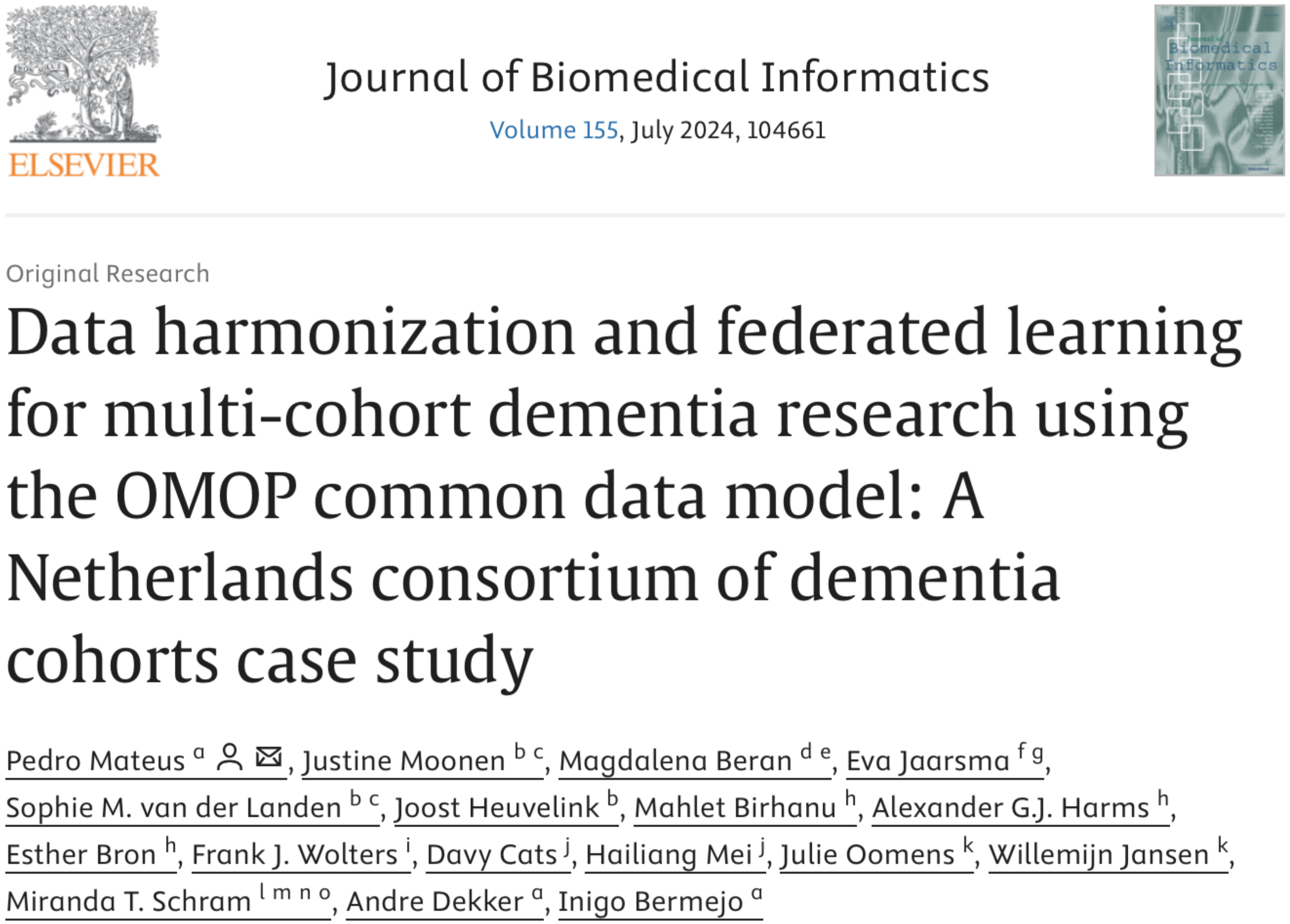
A Netherlands Consortium of Dementia Cohorts case study

Pedro Mateus, Maastricht University
24/09/2024

GROW School for Oncology & Developmental Biology

UM Maastricht University

Maastricht UMC+



ELSEVIER

Journal of Biomedical Informatics

Volume 155, July 2024, 104661

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]

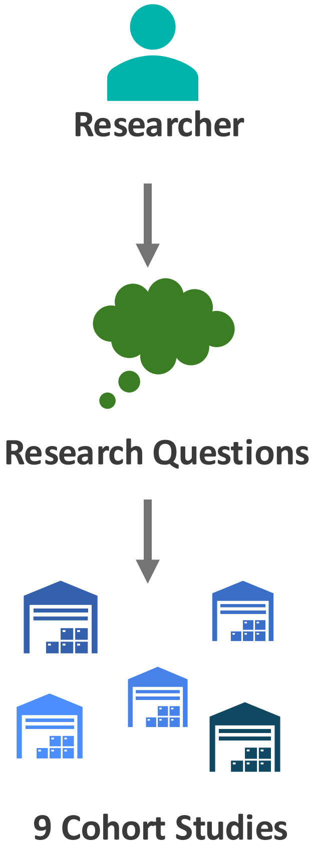
GROW School for Oncology & Developmental Biology

Maastricht University

Maastricht UMC+

Netherlands Consortium of Dementia Cohorts (NCDC)

Goal: “understand dementia in order to find clues for primary prevention by performing analysis of cohorts on aging and dementia.”

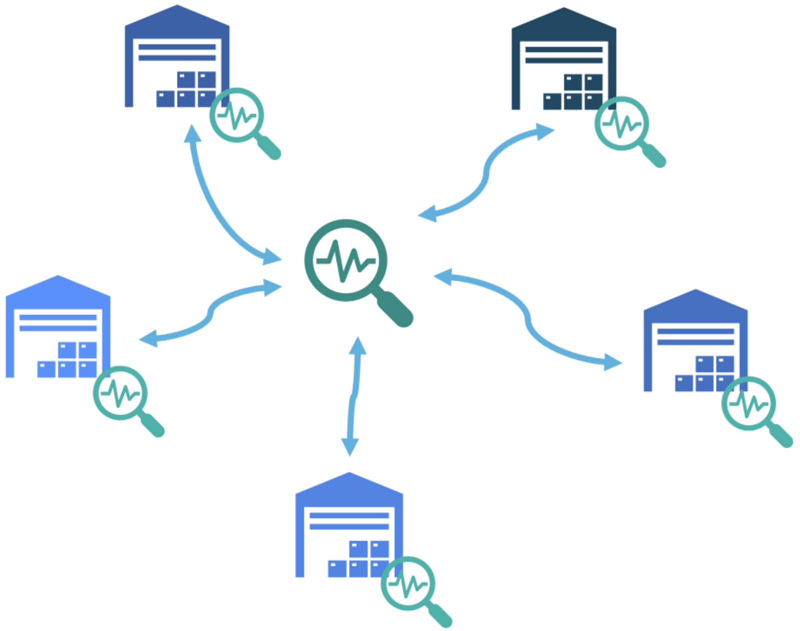


Researcher

Research Questions

9 Cohort Studies

Strategy: Federated Learning



Data remains in each institute. The analyses results are shared with the researcher using a software tool.

Overview

9 cohort studies (± 40,000 participants) from The Netherlands with data on cognitive decline and dementia.

- Population-based studies (cross-sectional and longitudinal) and memory clinic data.
- Tabular data: demographics, mortality, comorbidities, dementia/mci diagnosis, cognitive tests, plasma biomarkers.
- Imaging data: MRI scans.

PERSONAL HEALTH TRAIN

Federated infrastructure

- Installing the software at each cohort.
- Connecting the database.
- Preparing the algorithms for analysis.

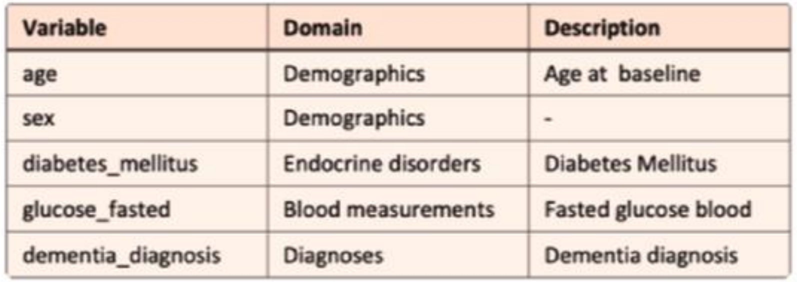
Local data extraction and harmonization

- What data model is suitable for cohort data?
- Standardize the data?
- ETL tools available?

Strategy

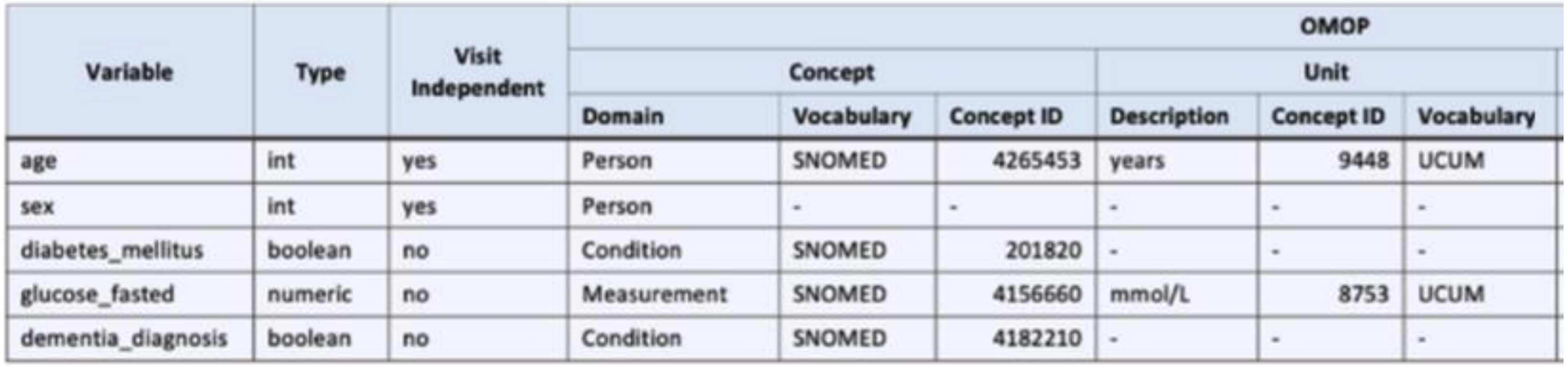
Consortium

Selection the set of variables necessary for the analysis.
Choosing the standard vocabulary and concepts.



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

Consortium OMOP mapping



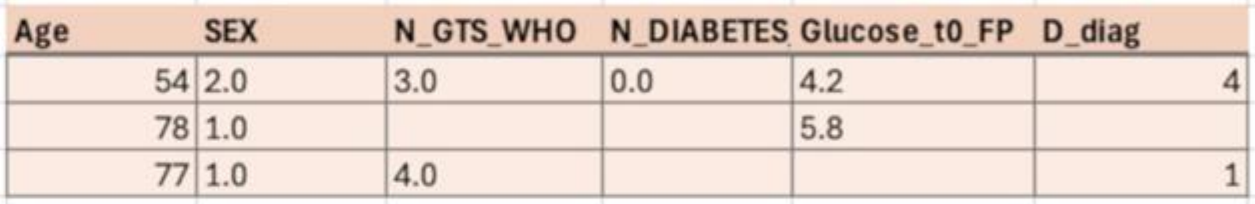
Variable	Type	Visit Independent	OMOP					
			Concept			Unit		
			Domain	Vocabulary	Concept ID	Description	Concept ID	Vocabulary
age	int	yes	Person	SNOMED	4265453	years	9448	UCUM
sex	int	yes	Person	-	-	-	-	-
diabetes_mellitus	boolean	no	Condition	SNOMED	201820	-	-	-
glucose_fasted	numeric	no	Measurement	SNOMED	4156660	mmol/L	8753	UCUM
dementia_diagnosis	boolean	no	Condition	SNOMED	4182210	-	-	-

Strategy

Cohort

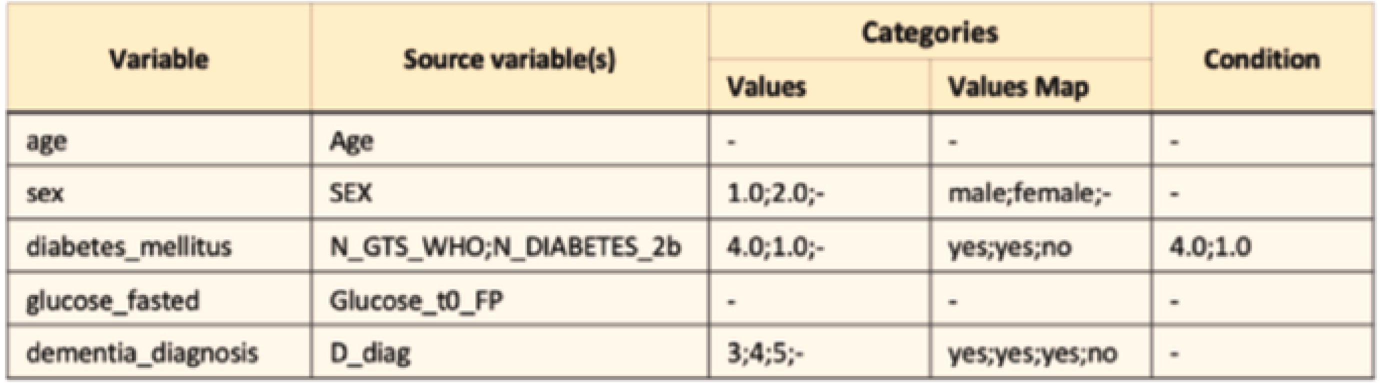
Collect codebook information and experts’ input.
Identify the metadata for the necessary variables.

Cohort dataset



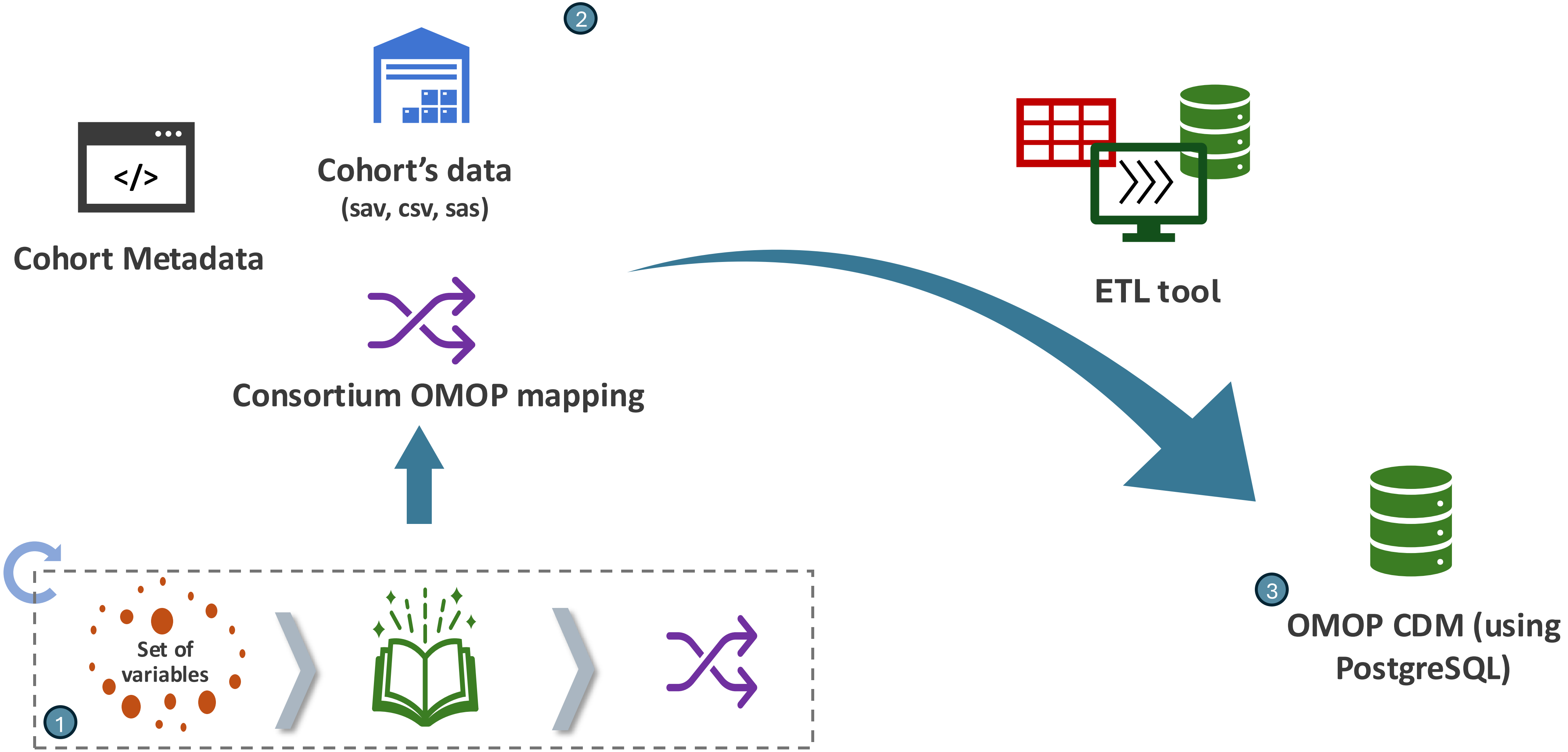
Age	SEX	N_GTS_WHO	N_DIABETES	Glucose_t0_FP	D_diag
54	2.0	3.0	0.0	4.2	4
78	1.0			5.8	
77	1.0	4.0			1

Cohort metadata



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	-

ETL Process



2

Cohort Metadata

Cohort's data
(sav, csv, sas)

Consortium OMOP mapping

1

Set of variables

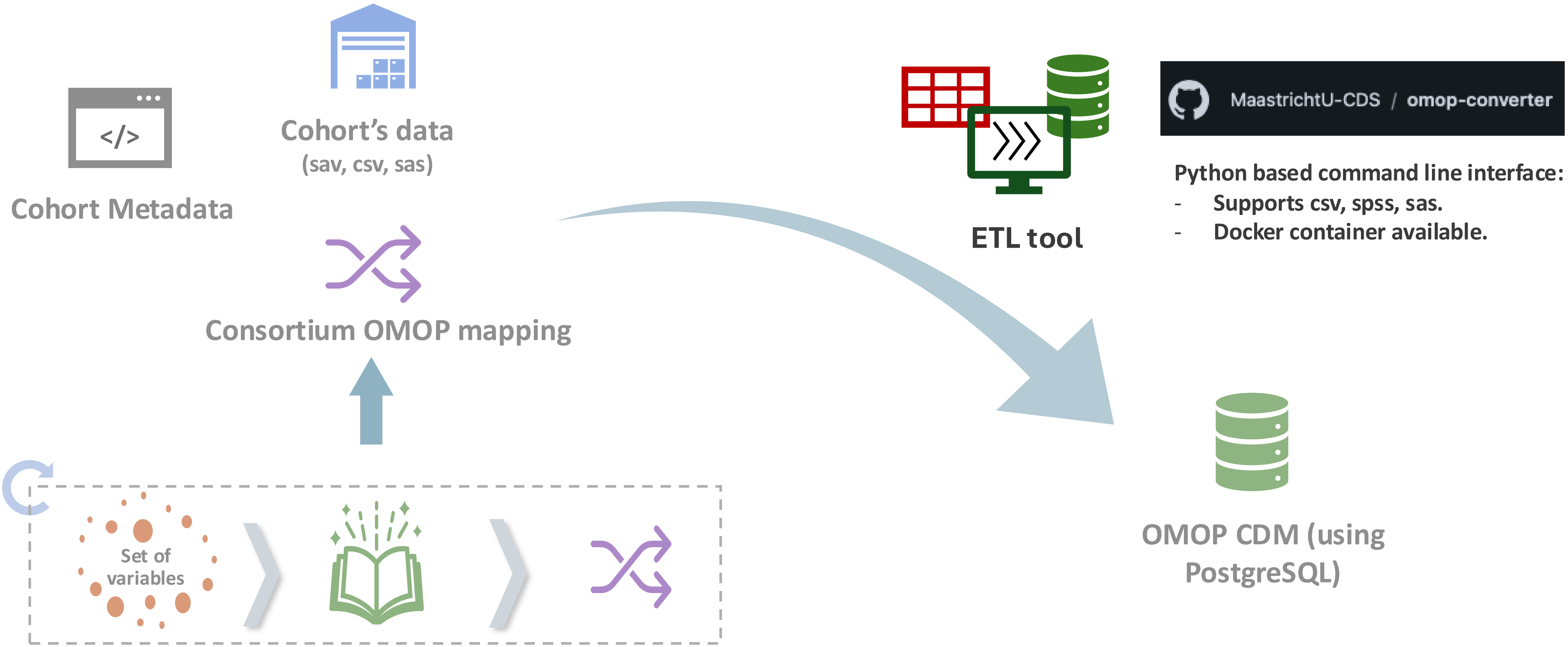
ETL tool

3

OMOP CDM (using PostgreSQL)

Consortium OMOP mapping development (new variables can be included in future iterations)

ETL Process



MaastrichtU-CDS / omop-converter

Python based command line interface:

- Supports csv, spss, sas.
- Docker container available.

Consortium OMOP mapping development (new variables can be included in future iterations)

Achievements and Challenges

Cohort data harmonized to the OMOP CDM for the 9 cohorts.

ETL tool to harmonize cohort data that decouples cohort and consortium metadata.

Federated infrastructure connecting the consortium cohorts.

Successfully performing analysis with the federated infrastructure.

Achievements and Challenges

Cohort data harmonized to the OMOP CDM for the 9 cohorts.

ETL tool to harmonize cohort data that decouples cohort and consortium metadata.

Federated infrastructure connecting the consortium cohorts.

Successfully performing analysis with the federated infrastructure.

Cohort experts support

- ⚠ *Data access methods, security rules, and software tools available*
- ⚠ *Variability of the cohort data structure.*
- ⚠ *Local support may not be available.*

OMOP and Standardization

- ⚠ ***Complexity of the relational structure.***
- ⚠ ***Interoperability depends on the standardization – lack of consensus***
- ⚠ ***Adaptations needed to represent the cohort data***

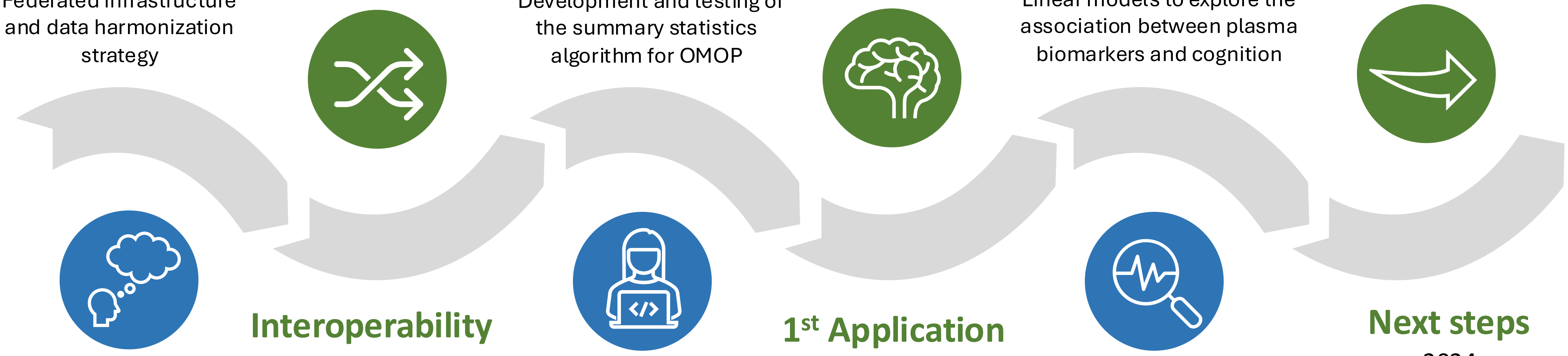
Legal agreements

- ⚠ *Defining standard agreements for new methods of analysis.*

Software tools

- ⚠ *No direct access to the data by the ETL tool developing team.*

Applications



Planning

2021

Federated infrastructure and data harmonization strategy

Interoperability

2021-2022

Data harmonization of the 9 cohorts. Federated infrastructure installation

Trial

2022-2023

Development and testing of the summary statistics algorithm for OMOP

1st Application

2023-2024

Training a convolutional neural network across 3 cohorts to predict brain age

2nd Application

2023-2024

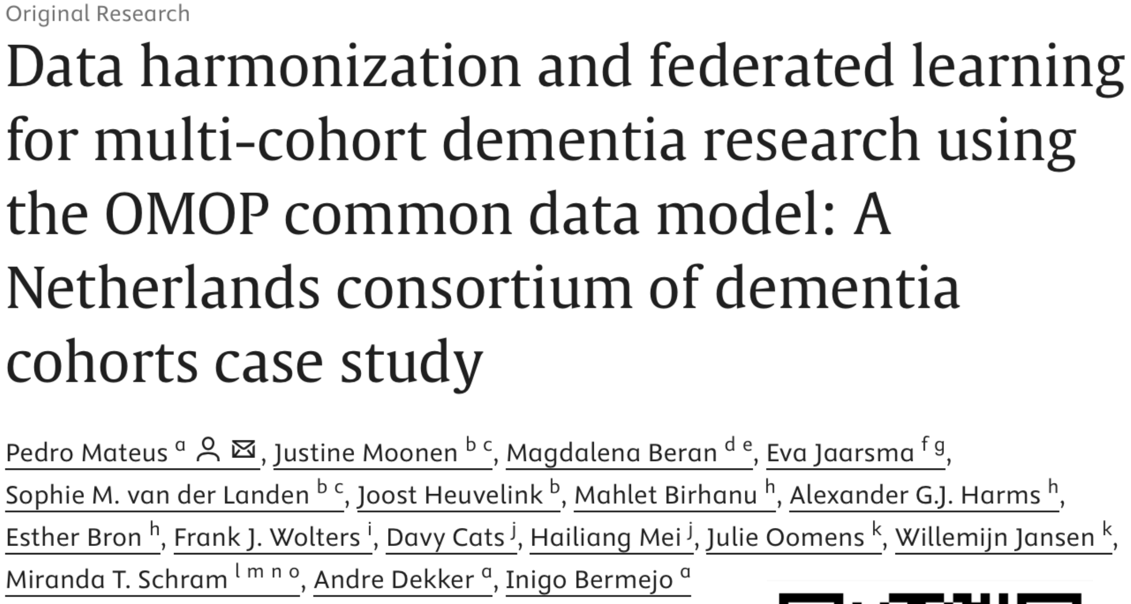
Linear models to explore the association between plasma biomarkers and cognition

Next steps

2024-...

New applications, additional data, maintenance, and continuous improvement of the infrastructure and ETL tool.

Questions



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]

OMOP converter for cohort studies

https://github.com/MaastrichtU-CDS/omop-converter

Feel free to contact us

pedro.mateus@maastro.nl, inigo.bermejo@maastro.nl