SNOMED International

Delivering SNOMED CT

2019 April Business Meeting

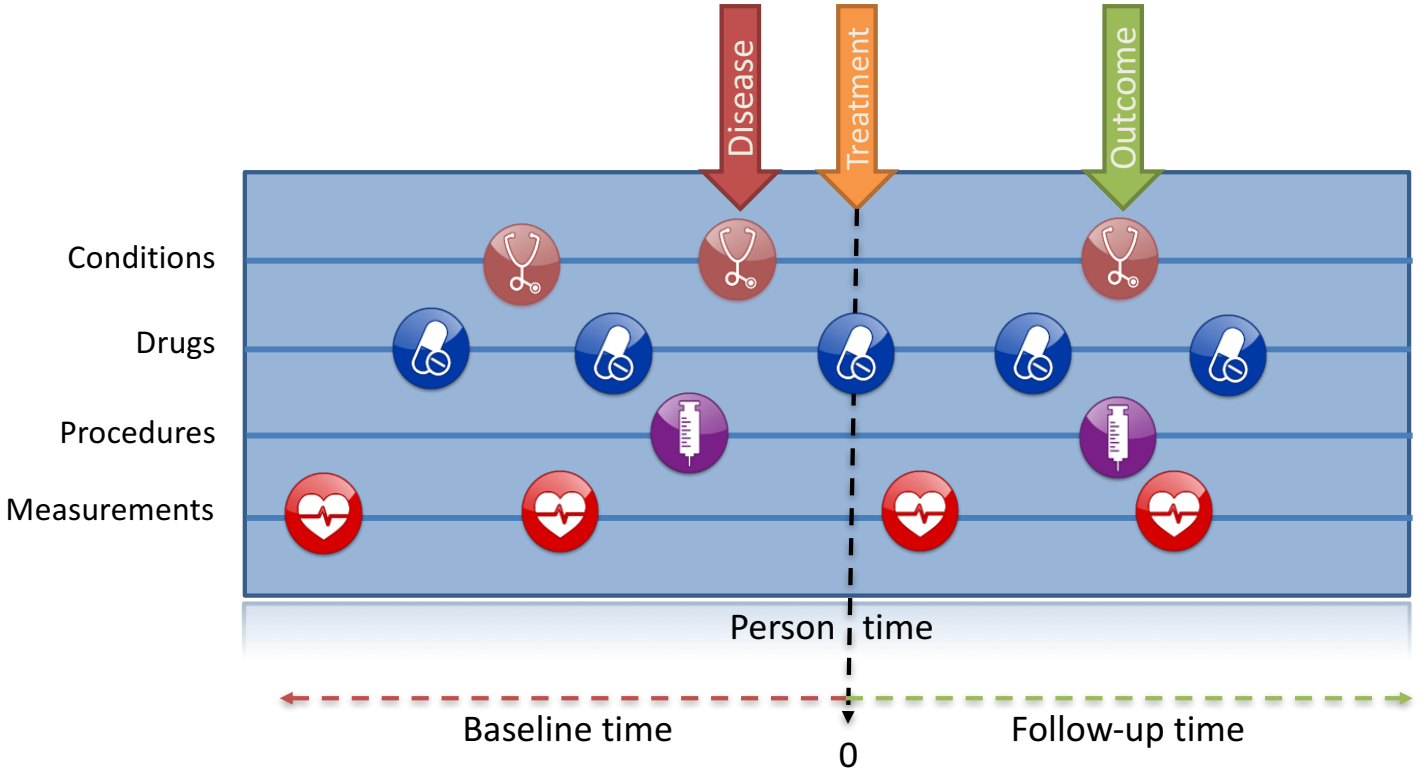
Enabling Large-Scale Analysis of Electronic Health Records by Standardizing to a Common Data Model

Peter R. Rijnbeek, PhD
Associate Professor Health Data Science
Department of Medical Informatics
Erasmus MC, The Netherlands

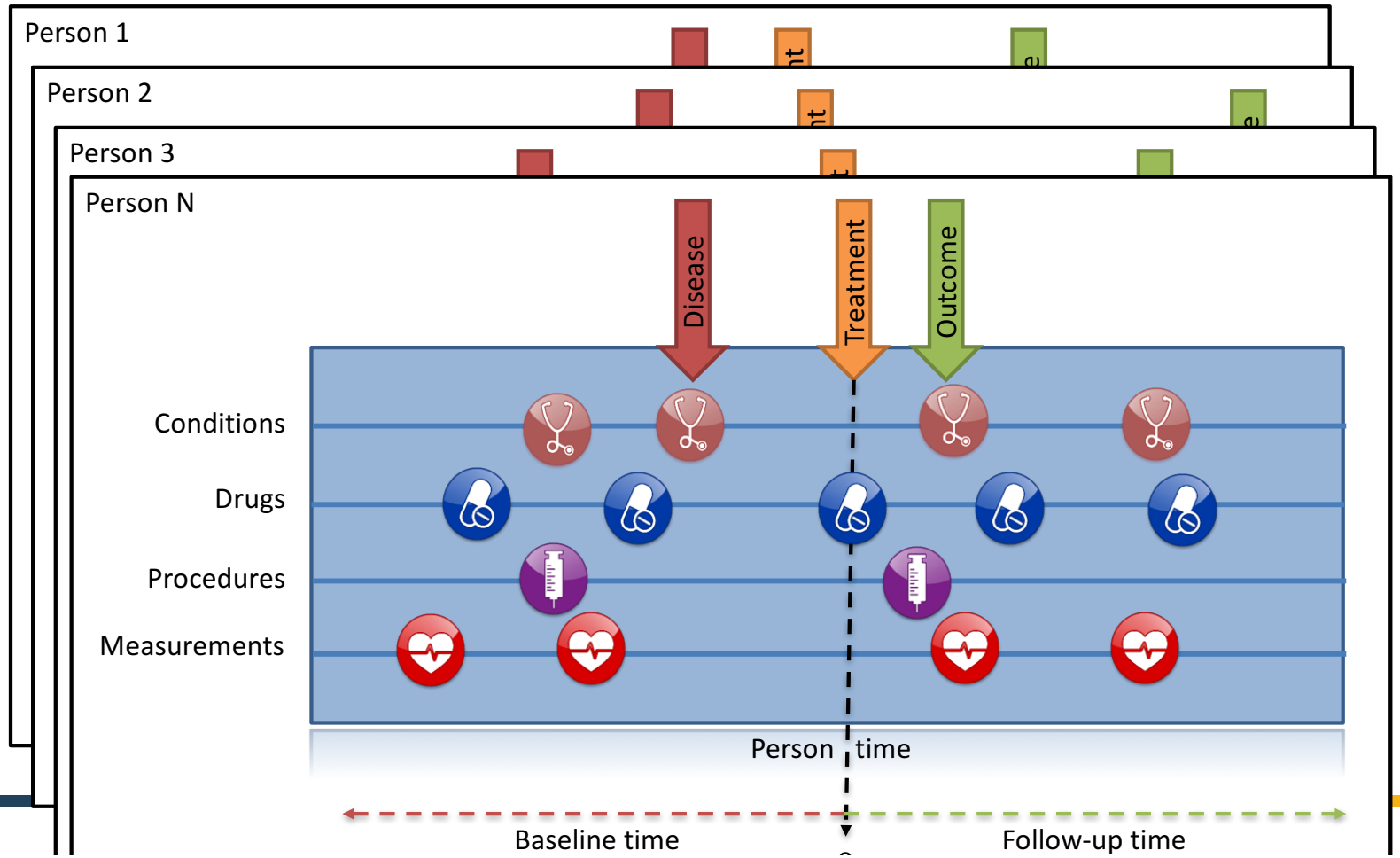
The Journey to large-scale analytics

- Introduction to the use of a Common Data Model and Standardized Vocabularies
- Example: Large-scale Patient-Level Prediction
- The European OHDSI Symposium
- The European Health Data and Evidence (EHDEN) Project

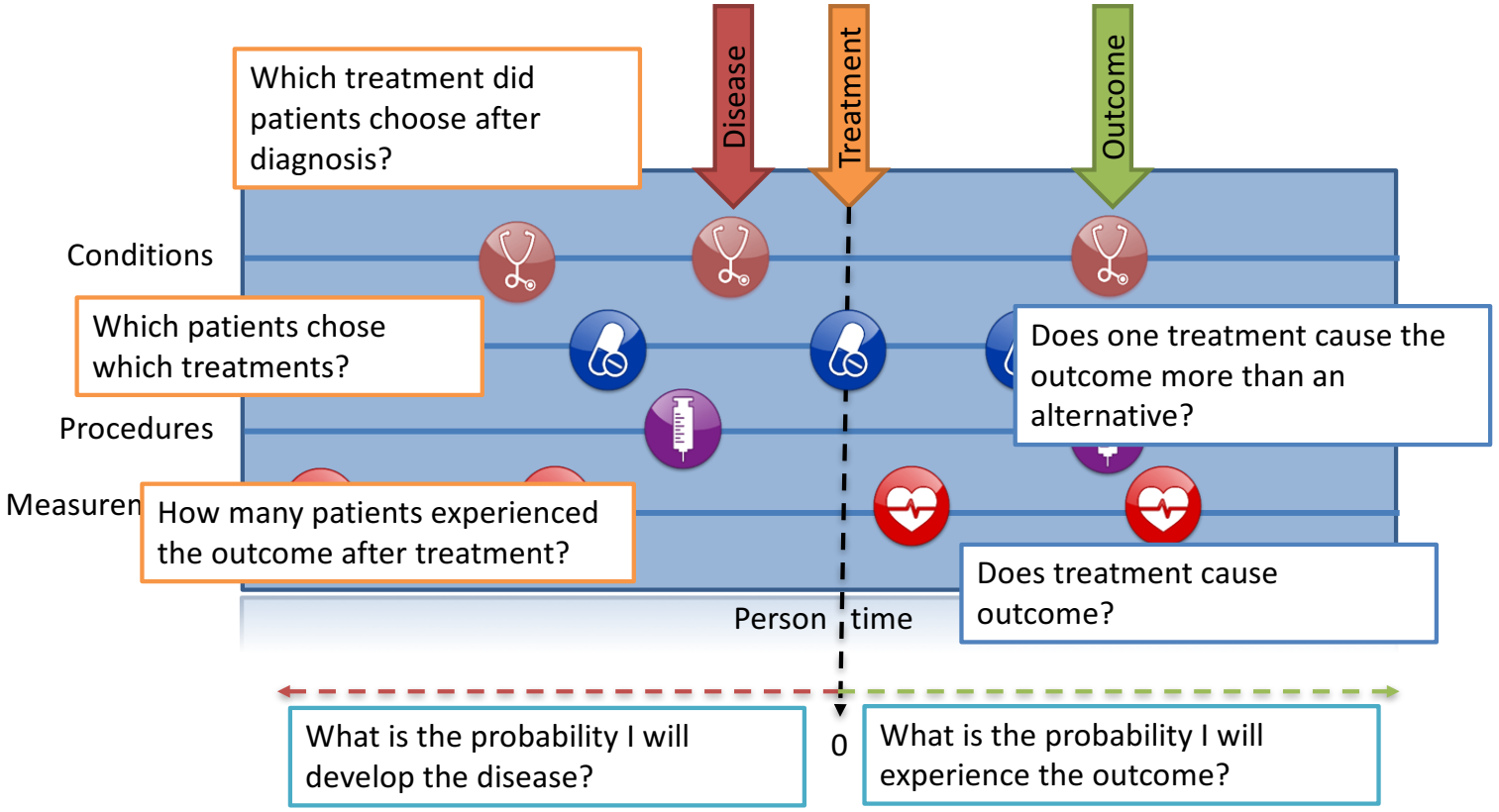
The health data originates from patient journeys



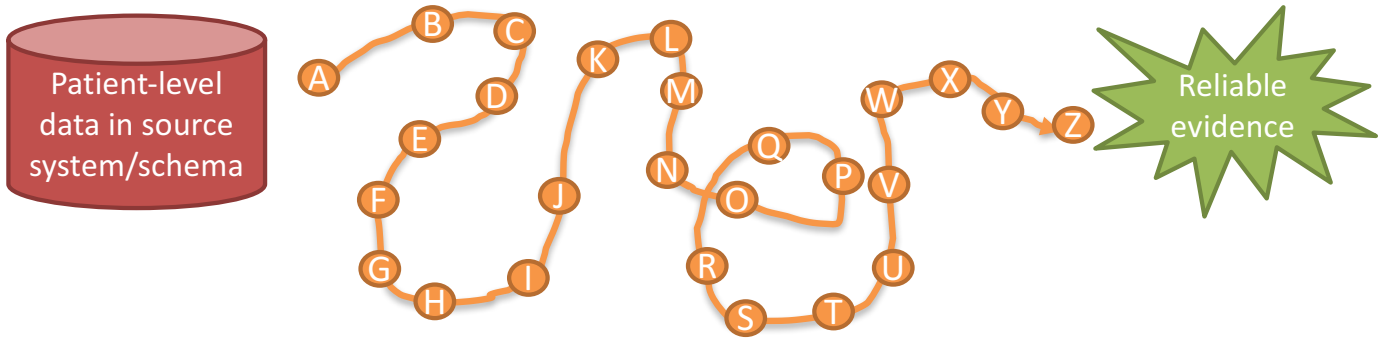
Each observational database is just an (incomplete) compilation of patient journeys



Questions asked across the patient journey

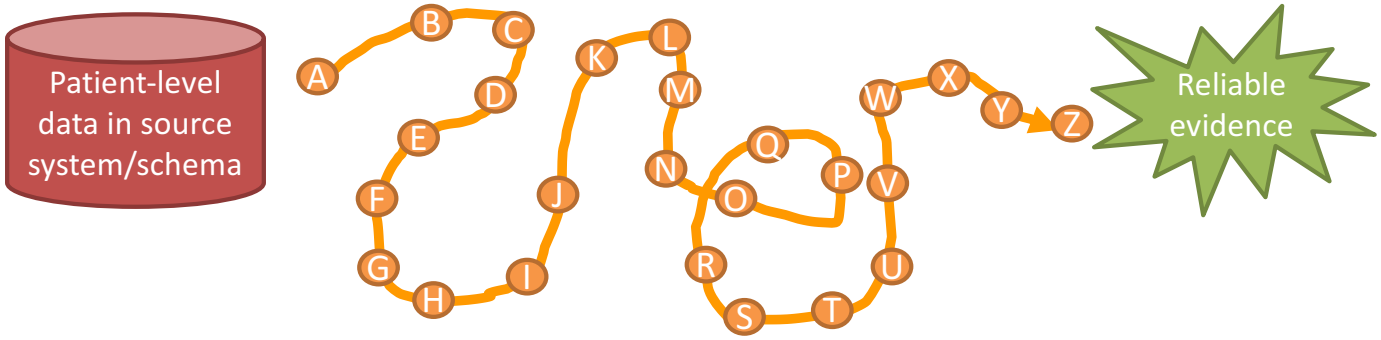


The journey to real-world evidence: a fully reproducible data flow



Minimum requirements to achieve reproducibility

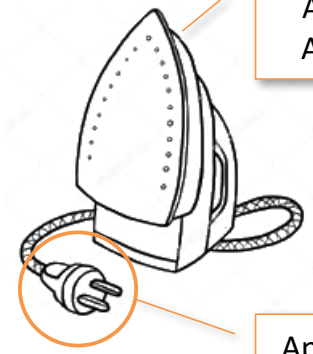
Desired attribute	Question	Researcher	Data	Analysis		Result
Reproducible	Identical	Different	Identical	Identical	=	Identical



- Complete documented specification that fully describes all data manipulations and statistical procedures
- Full analysis code that executes end-to-end (from source to results) without manual intervention

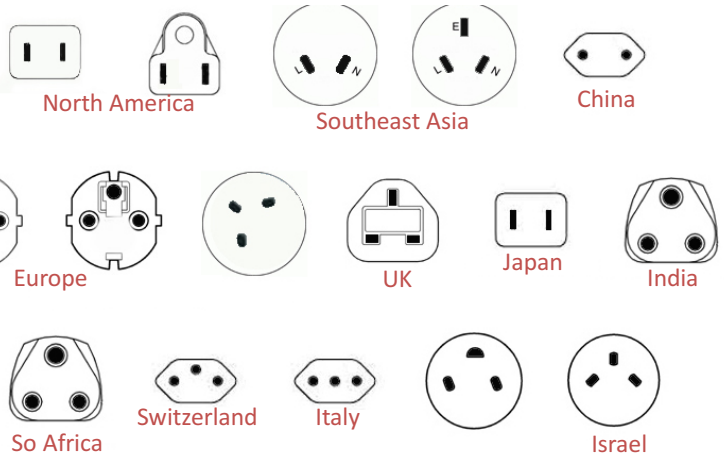
Current Approach: “One Study – One Script“

"What's the adherence to my drug of interest?"

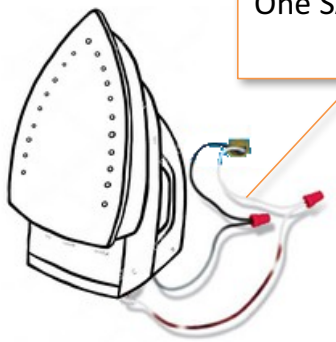


Analytical method: Adherence to Drug

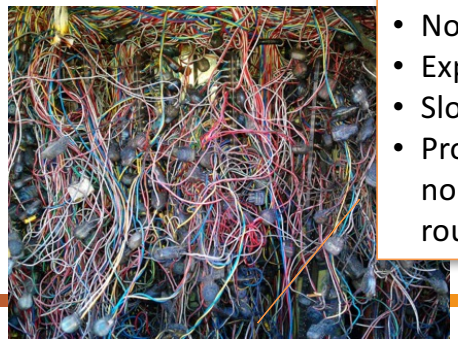
Application to data



Curren

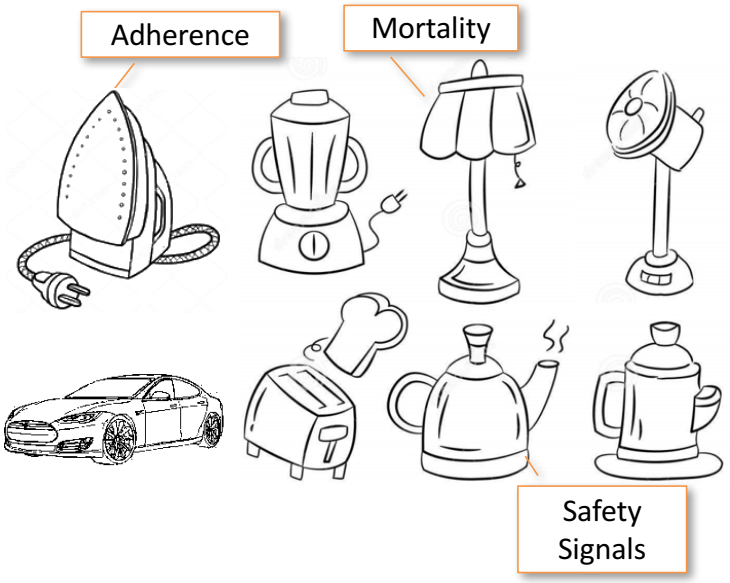


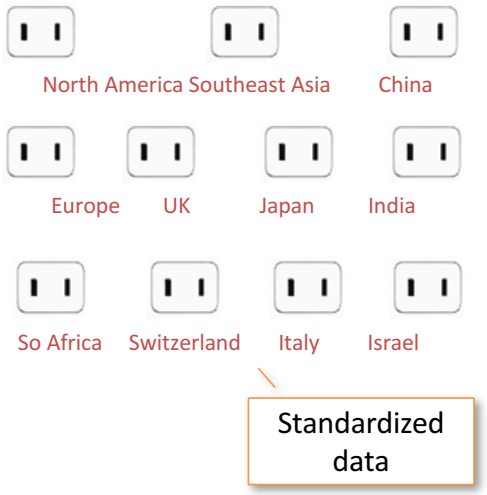
One SAS or R script for each study



- Not scalable
- Not transparent
- Expensive
- Slow
- Prohibitive to non-expert routine use

Solution: Data Standardization Enables Systematic Research



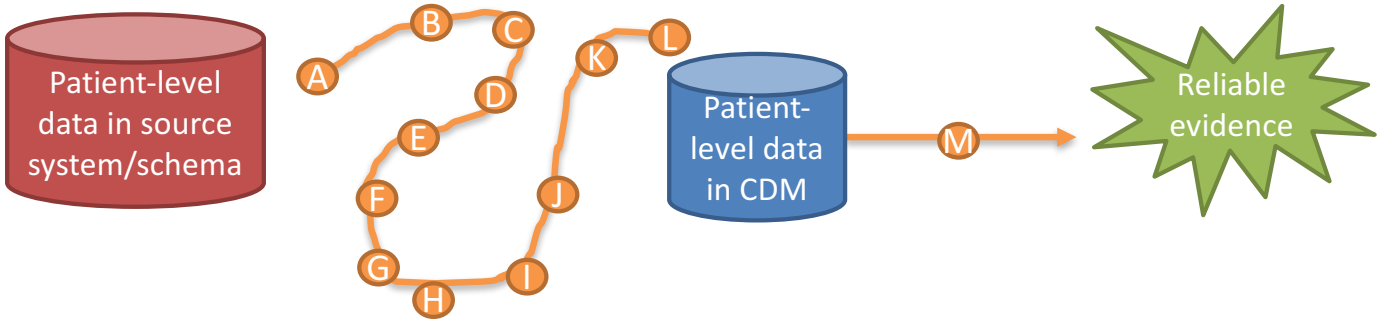


Analytical Tools

CDM

How a common data model + common analytics can support reproducibility

Desired attribute	Question	Researcher	Data	Analysis		Result
Reproducible	Identical	Different	Identical	Identical	=	Identical



- Use of common data model splits the journey into two segments: 1) data standardization, 2) analysis execution
- ETL specification and source code can be developed and evaluated separately from analysis design
- CDM creates opportunity for re-use of data step and analysis step

Observational Health Data Sciences and Informatics (OHDSI) mission

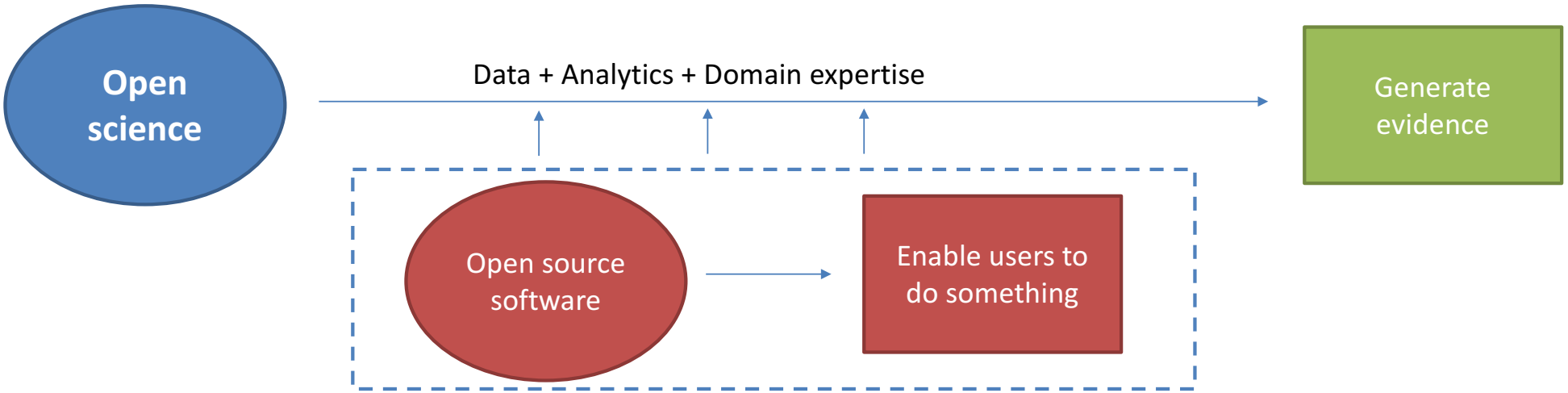
To improve health by empowering a community to collaboratively generate the evidence that promotes better health decisions and better care

OHDSI's global research community

- >200 collaborators from 25 different countries
- Experts in informatics, statistics, epidemiology, clinical sciences
- Active participation from academia, government, industry, providers
- Currently records on about 500 million unique patients in >100 databases

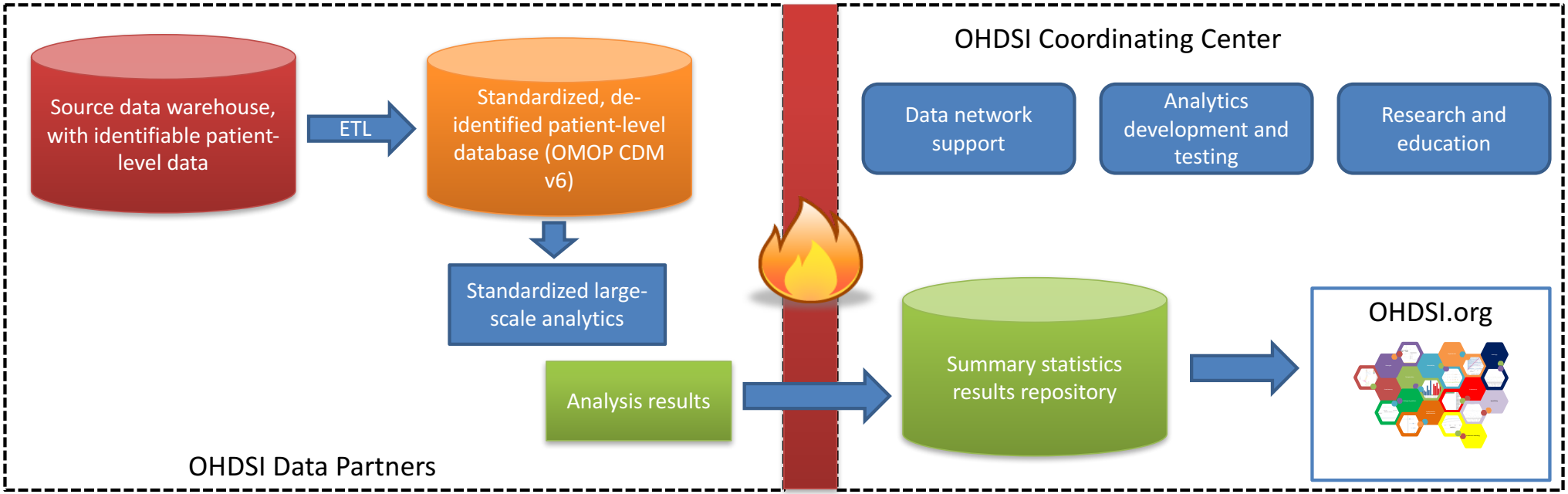
http://ohdsi.org/who-we-are/collaborators/

Open Science

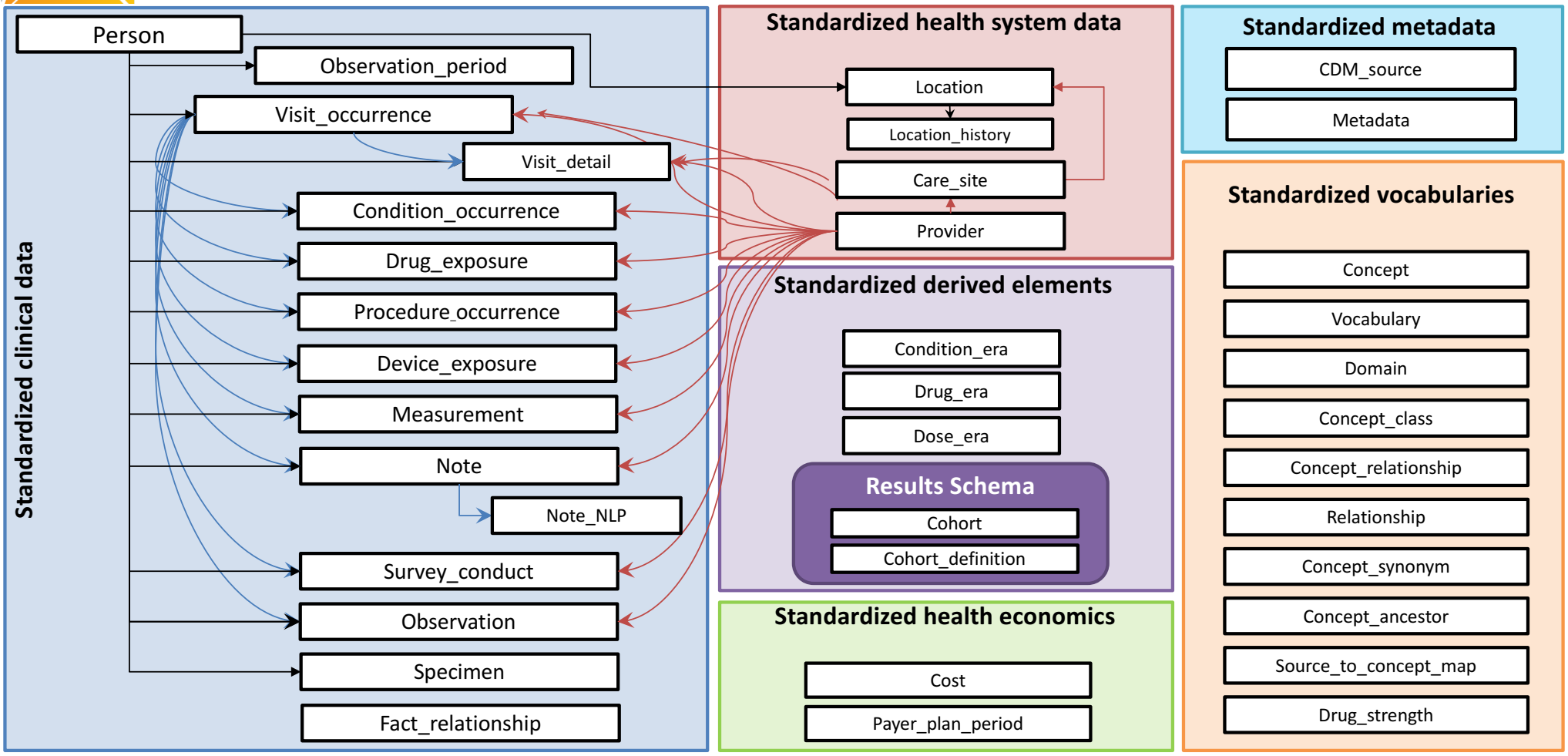


Standardized, transparent workflows

How OHDSI Works

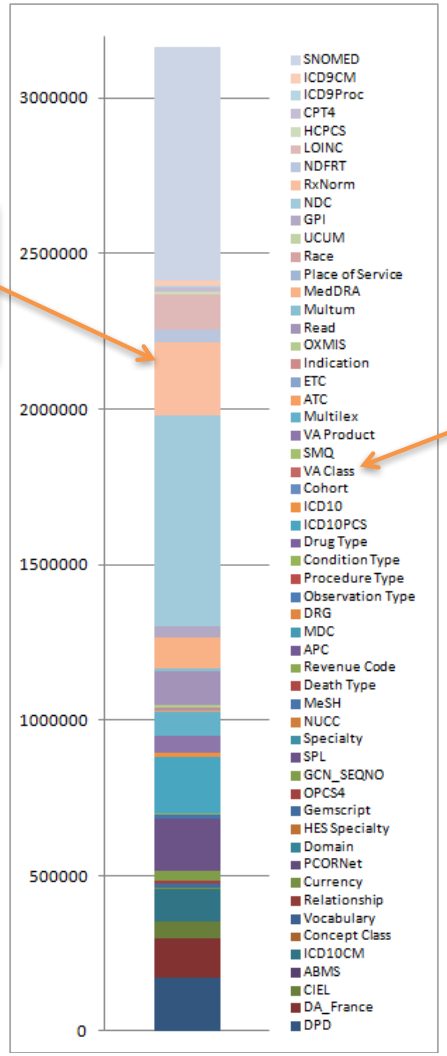


Deep information model
OMOP CDM Version 6



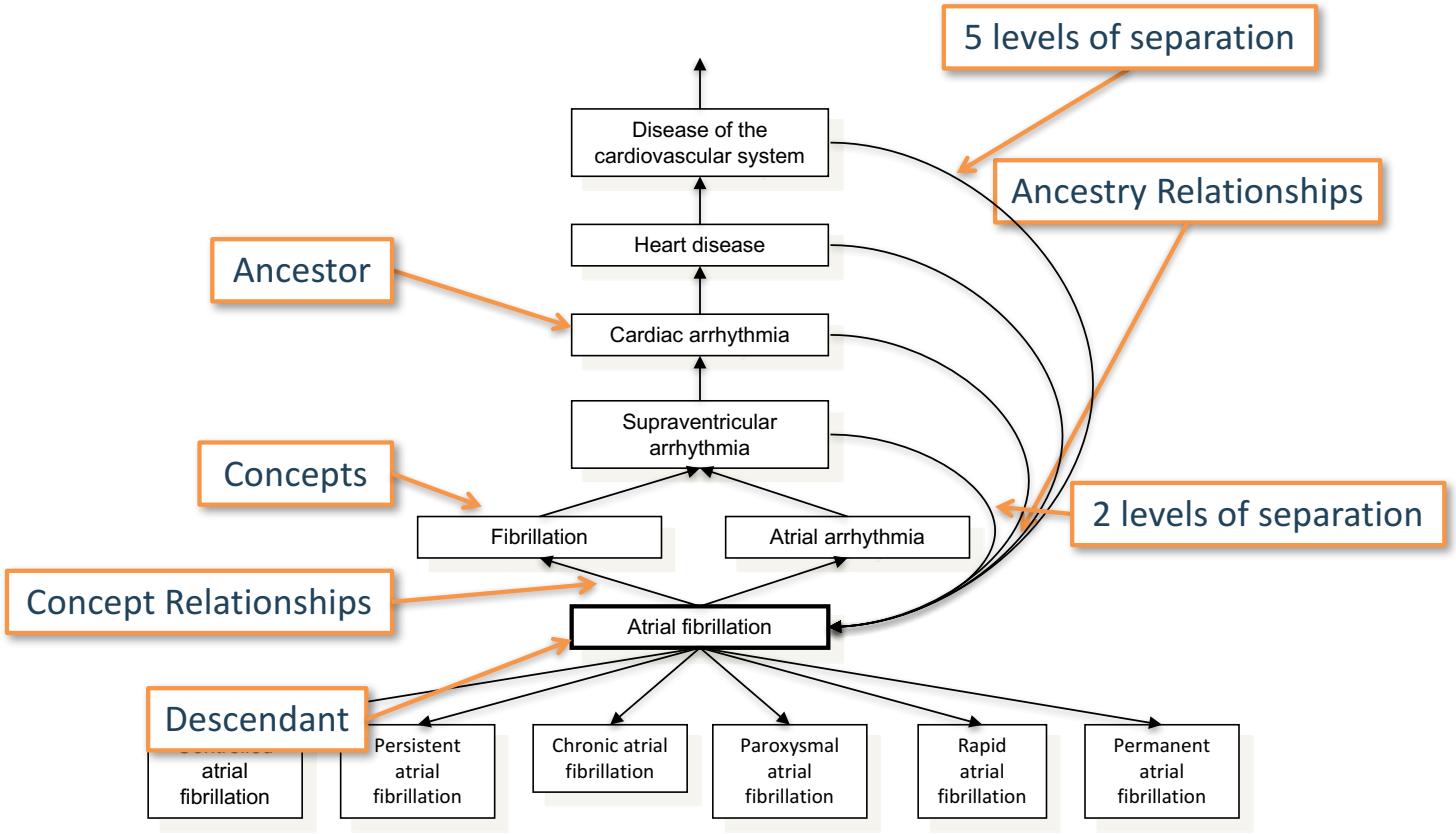
Single Concept Reference Table

All vocabularies stacked up in one table



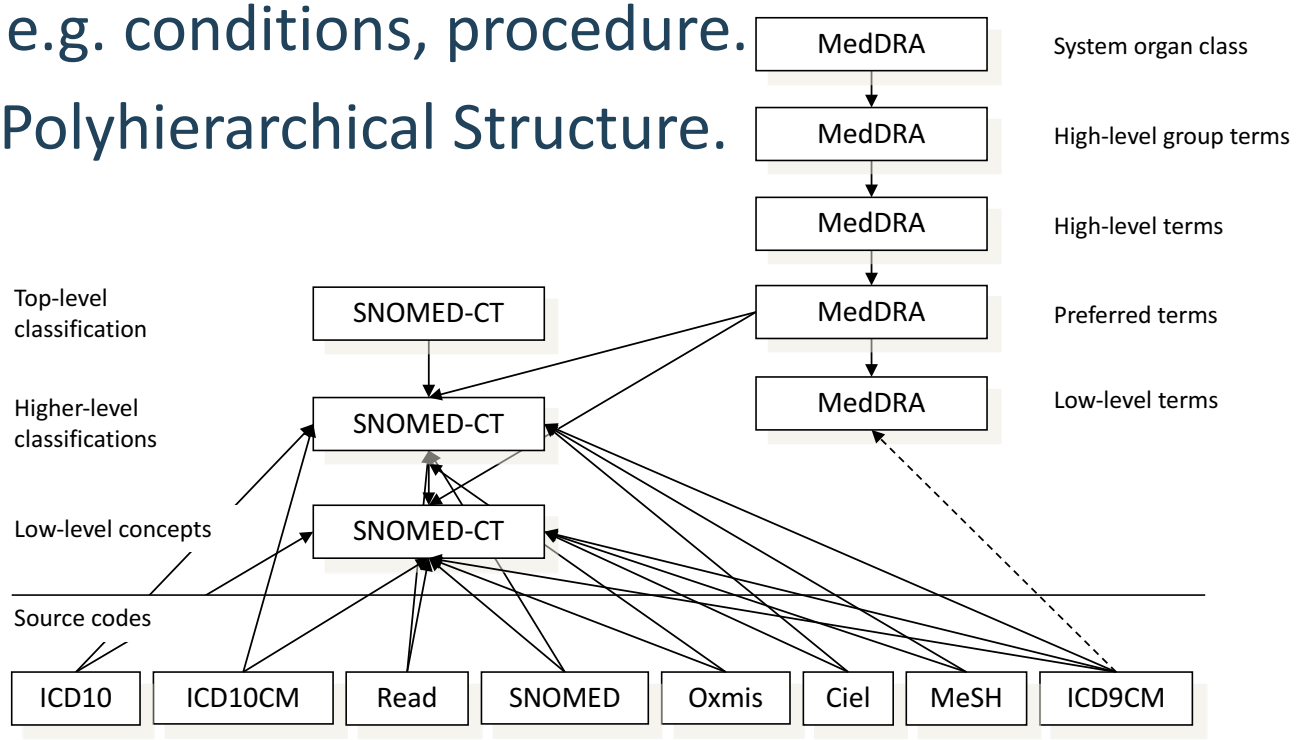
Vocabulary ID

Ancestry Relationships: Higher-Level Relationships



Use of SNOMED in the Standardized Vocabularies

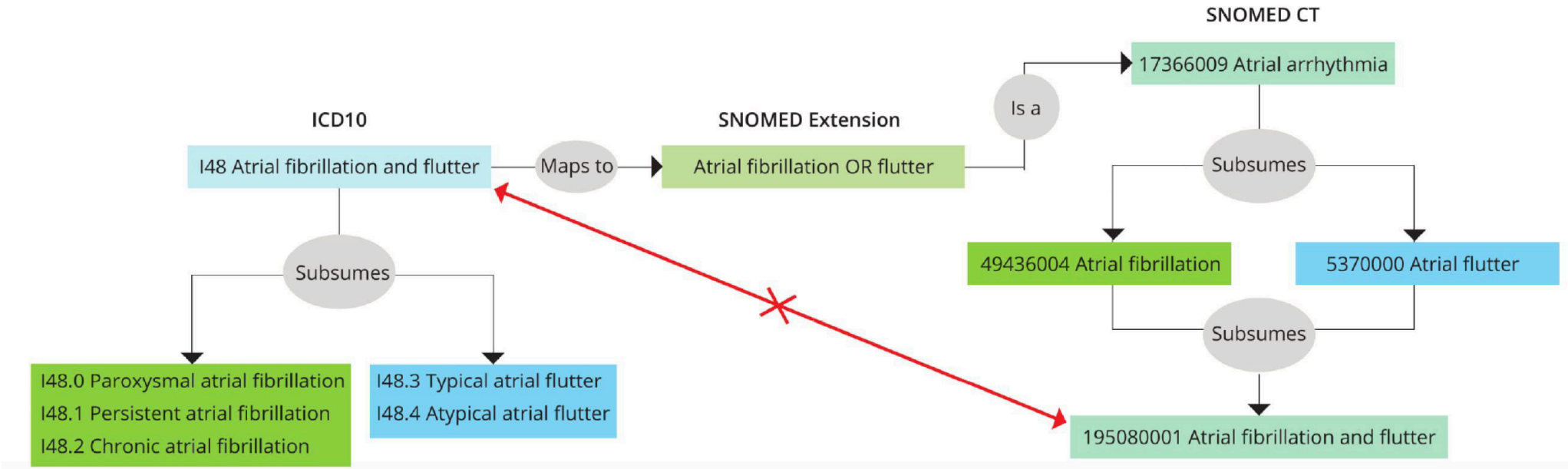
- SNOMED is the standard in several domains, e.g. conditions, procedure.
- Powerful Polyhierarchical Structure.



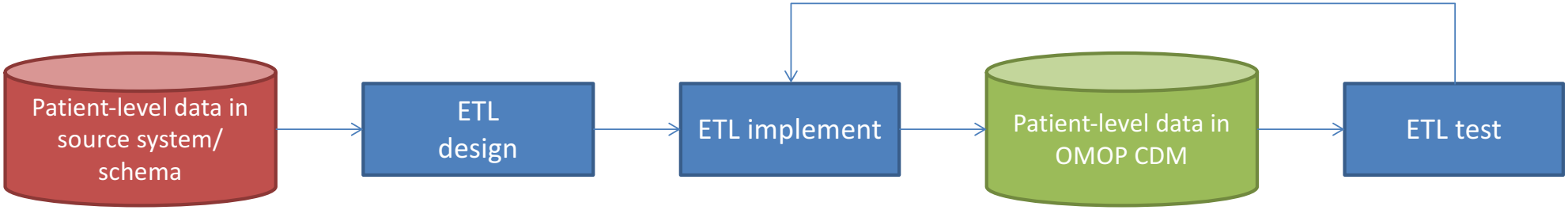
SNOMED Challenges

- We have to making mappings from many source coding systems to SNOMED in Europe.
- We want to use SNOMED across the world: how to deal with countries that do not (yet) have a license?
- We will require SNOMED extensions to accommodate differences in granularity or classification differences.

Example: SNOMED Challenges



Preparing your data for analysis



OHDSI tools built to help

WhiteRabbit: profile your source data

Rabbit-In-A-Hat: map your source structure to CDM tables and fields

ATHENA: standardized vocabularies for all CDM domains

Usagi: map your source codes to CDM vocabulary

CDM: DDL, index, constraints for Oracle, SQL Server, PostgresQL; Vocabulary tables with loading scripts

ACHILLES: profile your CDM data; review data quality assessment; explore population-level summaries

OHDSI Forums: Public discussions for OMOP CDM Implementers/developers

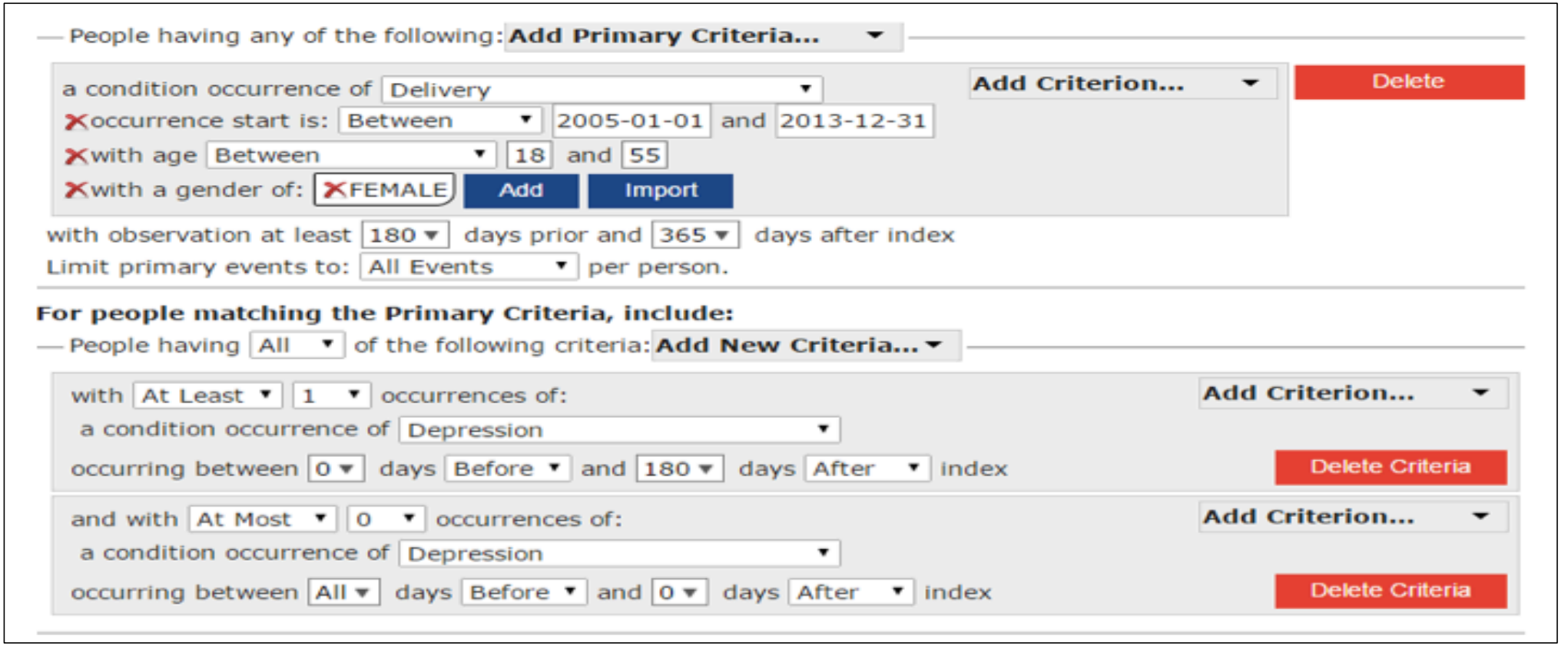
ACHILLES Heel Data Curation

Data Quality Messages

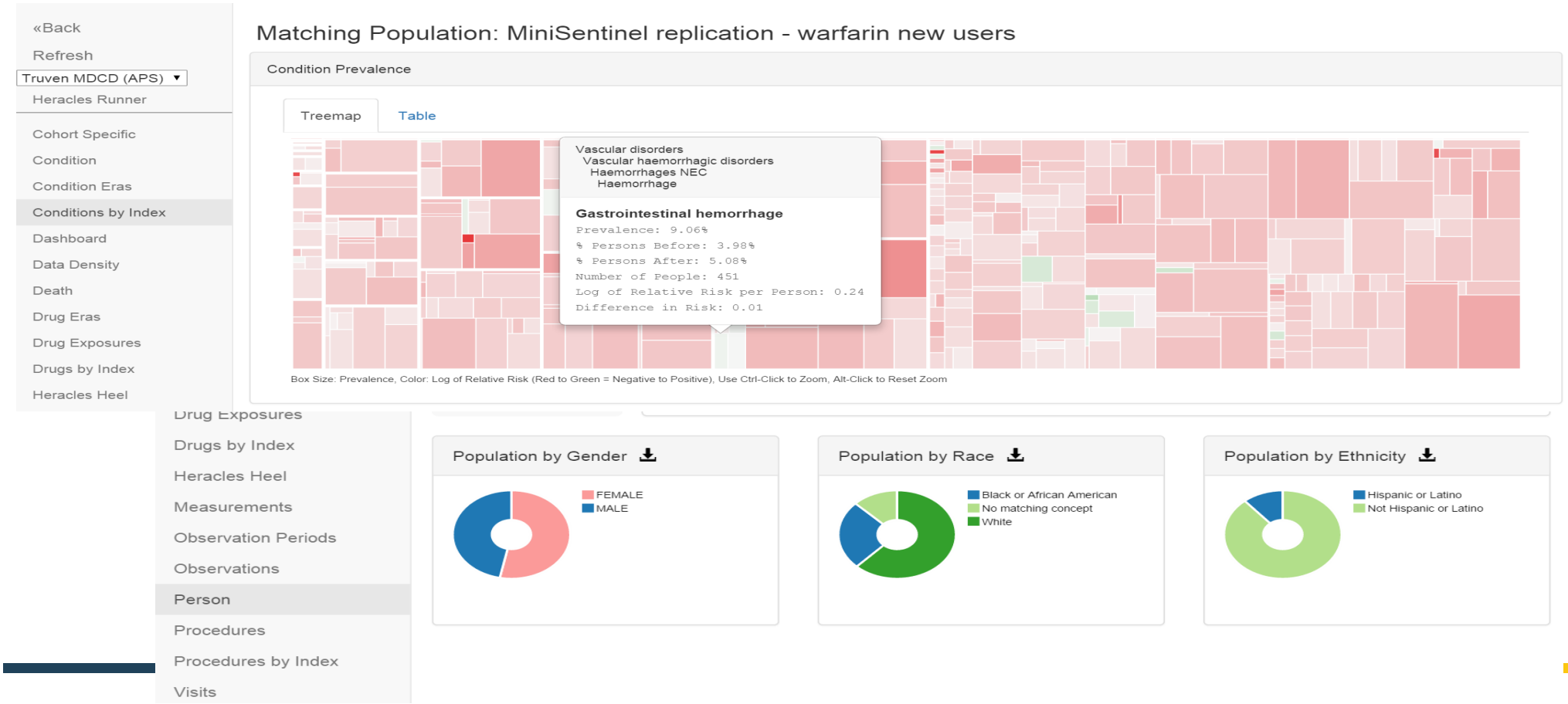
Search: Show / hide columns

Message Type	Message
ERROR	101-Number of persons by age, with age at first observation period; should not have age < 0, (n=848)
ERROR	103 - Distribution of age at first observation period (count = 1); min value should not be negative
ERROR	114-Number of persons with observation period before year-of-birth; count (n=851) should not be > 0
ERROR	206 - Distribution of age by visit_concept_id (count = 7); min value should not be negative
ERROR	301-Number of providers by specialty concept_id; 224 concepts in data are not in correct vocabulary (Specialty)
ERROR	400-Number of persons with at least one condition occurrence, by condition_concept_id; 115 concepts in data are not in correct vocabulary (SNOMED)
ERROR	406 - Distribution of age by condition_concept_id (count = 753); min value should not be negative

ATLAS to build, visualize, and analyze cohorts



Characterize the cohorts of interest



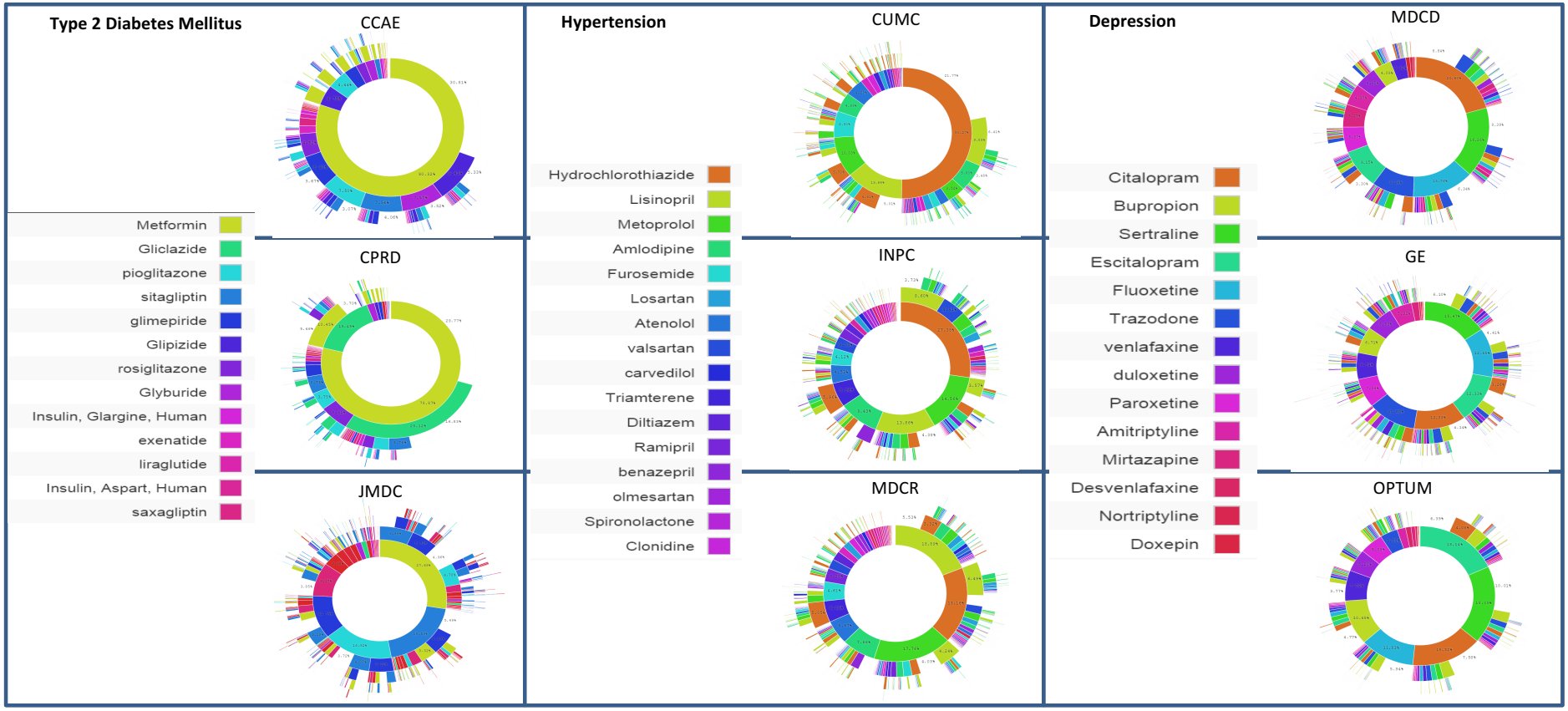
What is OHDSI's strategy to deliver reliable evidence?

- **Methodological research**
  - Develop new approaches to observational data analysis
  - Evaluate the performance of new and existing methods
  - Establish empirically-based scientific best practices
- **Open-source analytics development**
  - Design tools for data transformation and standardization
  - Implement statistical methods for large-scale analytics
  - Build interactive visualization for evidence exploration
- **Clinical evidence generation**
  - Identify clinically-relevant questions that require real-world evidence
  - Execute research studies by applying scientific best practices through open-source tools across the OHDSI international data network
  - Promote open-science strategies for transparent study design and evidence dissemination

Evidence OHDSI seeks to generate from observational data

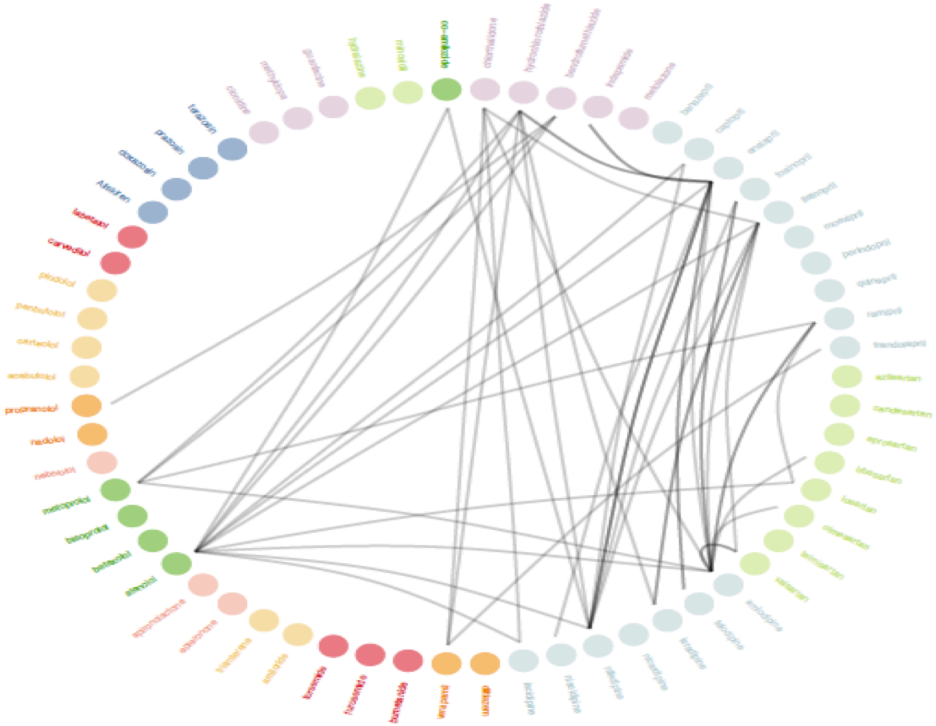
- **Clinical characterization**
  - Natural history: Who has diabetes, and who takes metformin?
  - Quality improvement: What proportion of patients with diabetes experience complications?
- **Population-level effect estimation**
  - Safety surveillance: Does metformin cause lactic acidosis?
  - Comparative effectiveness: Does metformin cause lactic acidosis more than glyburide?
- **Patient-level prediction**
  - Precision medicine: Given everything you know about me, now I started using metformin, what is the chance I will get lactic acidosis?
  - Disease interception: Given everything you know about me, what is the chance I will develop diabetes?

Clinical Characterization: Population-level heterogeneity across systems, and patient-level heterogeneity within systems

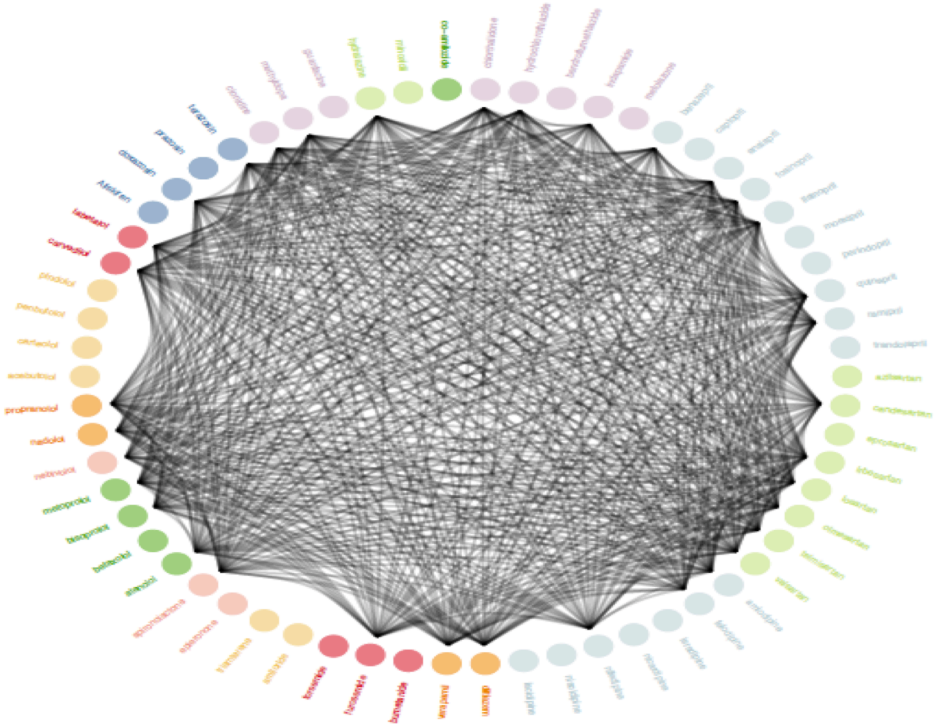


Population-Level Effect Estimation: LEGEND Study

Head-to-head HTN drug comparisons



- Trials: 40
- $N = 102 - [1148] - 33K$



- Comparisons: 10,278
- $N = 3502 - [212K] - 1.9M$

Journey of Patient-Level Prediction

An example of large-scale analysis enabled by data standardization

Clinicians are confronted with prediction questions on a daily basis. What options do they have?

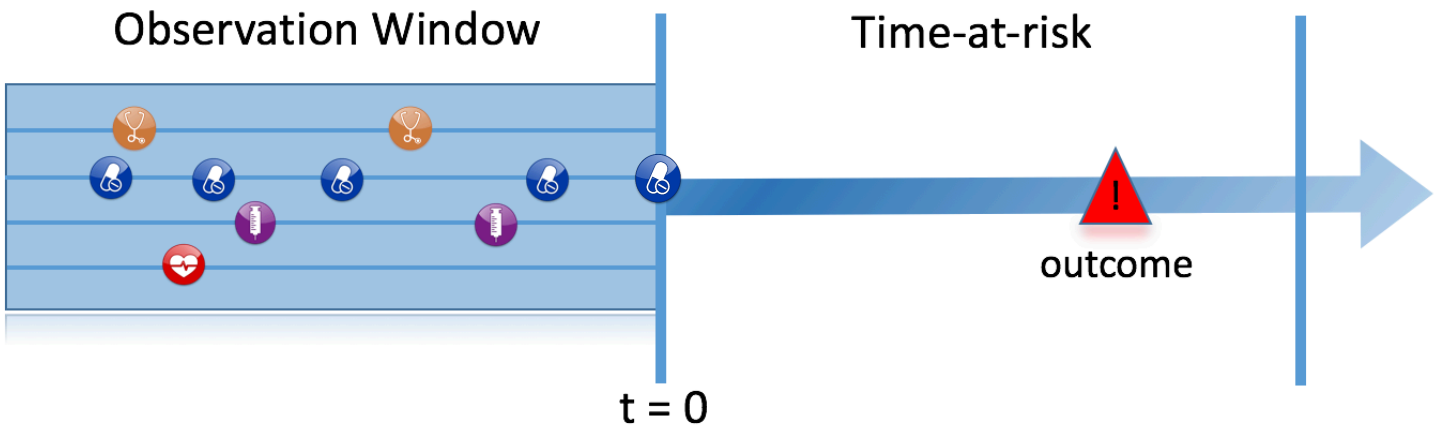
Deny ability to predict at the individual patient level

Quote an overall average to all patients

Utilize knowledge and personal experience

Provide a personalized prediction based on an advanced clinical prediction model

Problem definition



Among a target population (T), we aim to predict which patients at a defined moment in time (t=0) will experience some outcome (O) during a time-at-risk Prediction is done using only information about the patients in an observation window prior to that moment in time.

What are the key inputs to a patient-level prediction study?

Input parameter	Design choice
Target cohort (T)	
Outcome cohort (O)	
Time-at-risk	
Model specification -which model(s)? -which parameters? -which covariates?	

Types of prediction problems in healthcare

Type	Structure	Example
Disease onset and progression	Amongst patients who are newly diagnosed with **<insert your favorite disease>**, which patients will go on to have **<another disease or related complication>** within **<time horizon from diagnosis>**?	Among newly diagnosed AFib patients, which will go onto to have ischemic stroke in next 3 years?
Treatment choice	Amongst patients with **<indicated disease> who are treated with either <treatment 1> or <treatment 2>**, which patients were treated with **<treatment 1>** (**on day 0**)?	Among AFib patients who took either warfarin or rivaroxaban, which patients got warfarin? (as defined for propensity score model)
Treatment response	Amongst patients who are new users of **<insert your favorite chronically-used drug>**, which patients will **<insert desired effect>** in **<time window>**?	Which patients with T2DM who start on metformin stay on metformin after 3 years?
Treatment safety	Amongst patients who are new users of **<insert your favorite drug>,** which patients will experience **<insert your favorite known adverse event from the drug profile>** within **<time horizon following exposure start>**?	Among new users of warfarin, which patients will have GI bleed in 1 year?
Treatment adherence	Amongst patients who are new users of **<insert your favorite chronically-used drug>**, which patients will achieve **<adherence metric threshold>** at **<time horizon>**?	Which patients with T2DM who start on metformin achieve >=80% proportion of days covered at 1 year?

Growing interest in prediction modelling

Current status of predictive modelling

Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review

RECEIVED 27 October 2015
REVISED 25 January 2016
ACCEPTED 20 February 2016

Benjamin A Goldstein[1,2], Ann Marie Navar[2,3], Michael J Pencina[1,2], John PA Ioannidis[4,5]

ABSTRACT

Objective Electronic health records (EHRs) are an increasingly common data source for clinical risk prediction, presenting both unique analytic opportunities and challenges. We sought to evaluate the current state of EHR based risk prediction modeling through a systematic review of clinical prediction studies using EHR data.

Methods We searched PubMed for articles that reported on the use of an EHR to develop a risk prediction model from 2009 to 2014. Articles were extracted by two reviewers, and we abstracted information on study design, use of EHR data, model building, and performance from each publication and supplementary documentation.

Results We identified 107 articles from 15 different countries. Studies were generally very large (median sample size = 26 100) and utilized a diverse array of predictors. Most used validation techniques ($n = 94$ of 107) and reported model coefficients for reproducibility ($n = 83$). However, studies did not fully leverage the breadth of EHR data, as they uncommonly used longitudinal information ($n = 37$) and employed relatively few predictor variables (median = 27 variables). Less than half of the studies were multicenter ($n = 50$) and only 26 performed validation across sites. Many studies did not fully address biases of EHR data such as missing data or loss to follow-up. Average c-statistics for different outcomes were: mortality (0.84), clinical prediction (0.83), hospitalization (0.71), and service utilization (0.71).

Conclusions EHR data present both opportunities and challenges for clinical risk prediction. There is room for improvement in designing such studies.

Goldstein BA, J Am Med Inform Assoc. 2016.

Current status of predictive modelling

- Inadequate internal validation
- Small sets of features
- Incomplete dissemination of model and results
- No transportability assessment
- Impact on clinical decision making unknown

→ Relatively few prediction models are used in clinical practice

OHDSI Mission for Patient-Level Prediction

OHDSI aims to develop a systematic process to learn and evaluate large-scale patient-level prediction models using observational health data in a data network

PLP Framework Paper

JAMIA
A SCHOLARLY JOURNAL OF INFORMATICS IN HEALTH AND BIOMEDICINE

Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data

Jenna M Reps, Martijn J Schuemie, Marc A Suchard, Patrick B Ryan, Peter R Rijnbeek

Journal of the American Medical Informatics Association, ocy032,
https://doi.org/10.1093/jamia/ocy032
Published: 27 April 2018 **Article history** ▾

Abstract

Objective

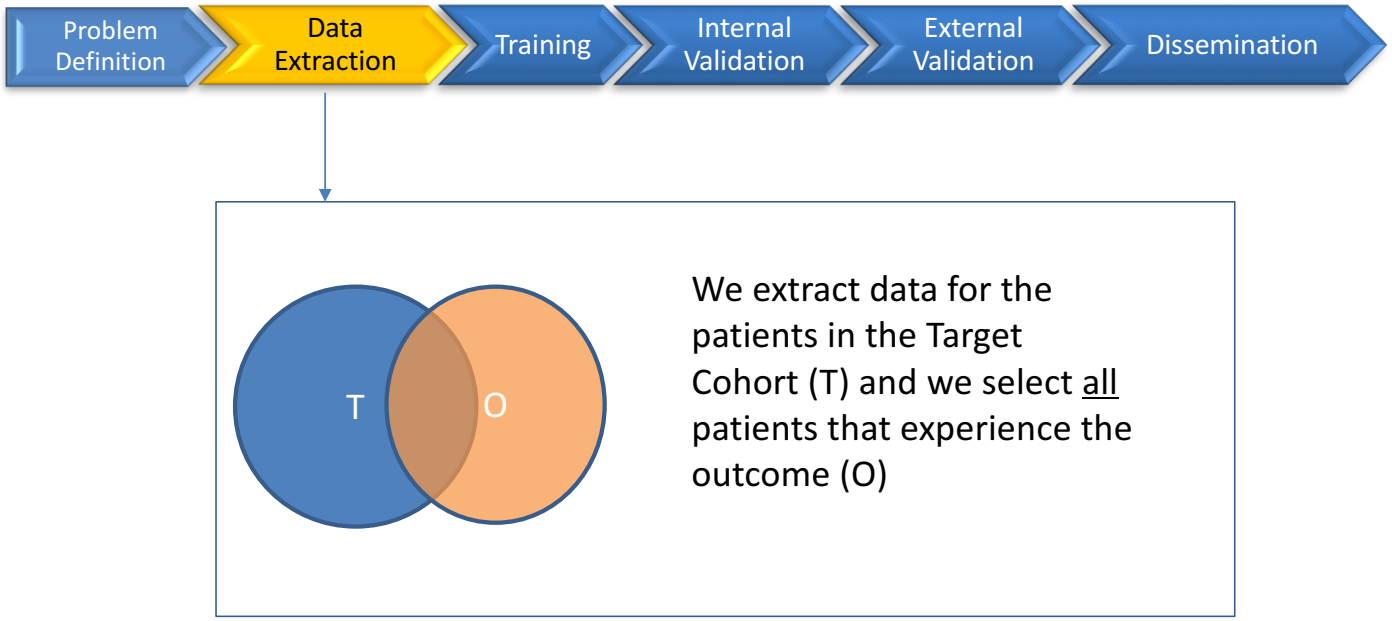
To develop a conceptual prediction model framework containing standardized steps and describe the corresponding open-source software developed to consistently implement the framework across computational environments and observational healthcare databases to enable model sharing and reproducibility.

Prediction Model Development

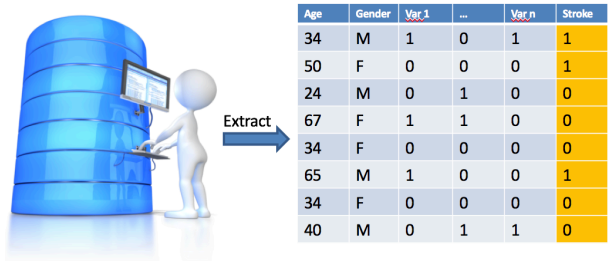
Problem pre-specification. A study protocol should unambiguously pre-specify the planned analyses.

Transparency. Others should be able to reproduce a study in every detail using the provided information. All analysis code should be made available as open source on the OHDSI Github.

Prediction Model Development



Prediction Model Development



Age	Gender	Var 1	...	Var n	Stroke
34	M	1	0	1	1
50	F	0	0	0	1
24	M	0	1	0	0
67	F	1	1	0	0
34	F	0	0	0	0
65	M	1	0	0	1
34	F	0	0	0	0
40	M	0	1	1	0

Data is extracted from the OMOP CDM using the Feature Extraction R-Package.

This allows for specification of the candidate predictors and time windows.

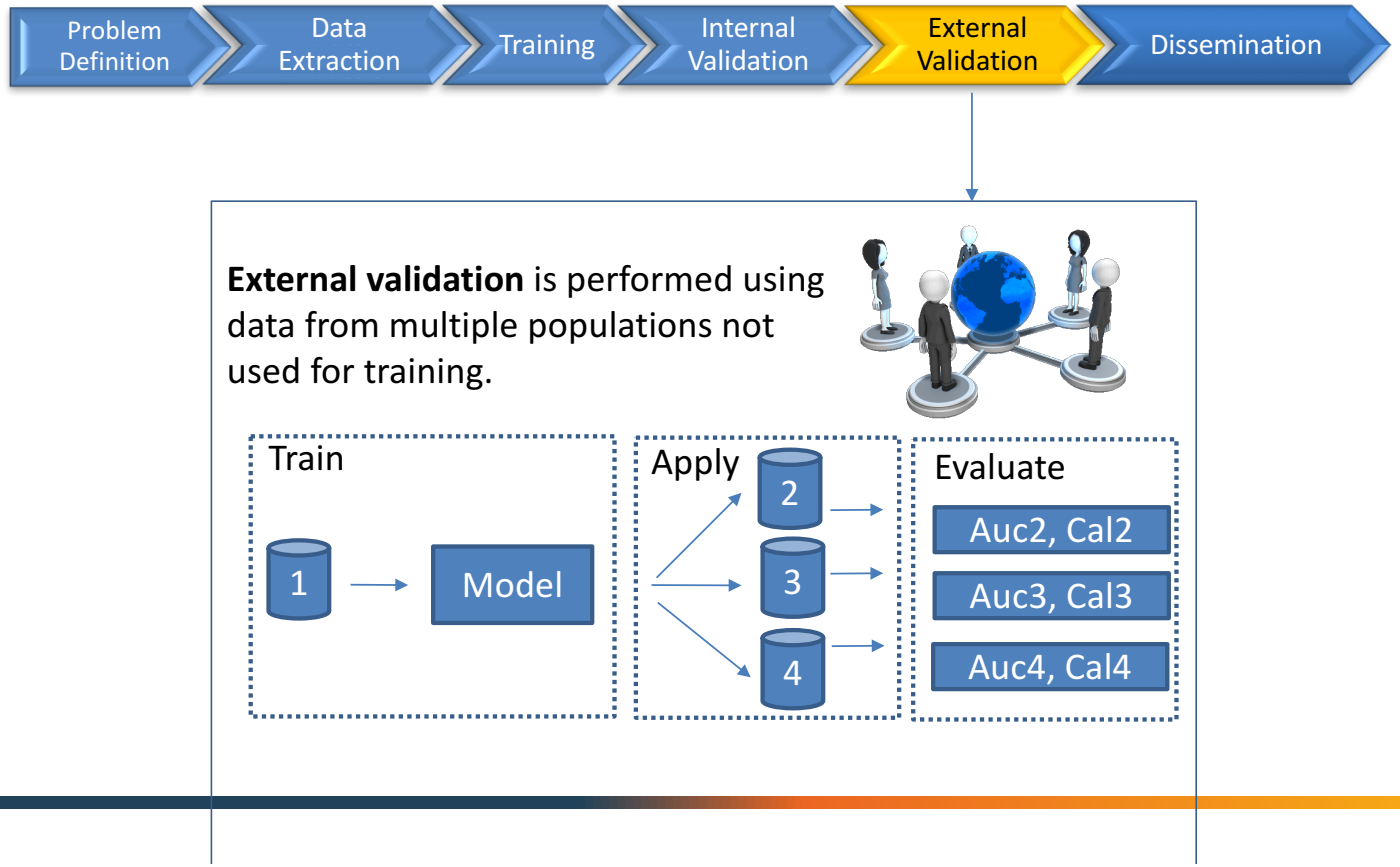
Prediction Model Development

Model training and **Internal validation** is done using a train test split:

1. Person split: examples are assigned randomly to the train or test set, or

2. Time split: a split is made at a moment in time (temporal validation)

External Validation



Dissemination

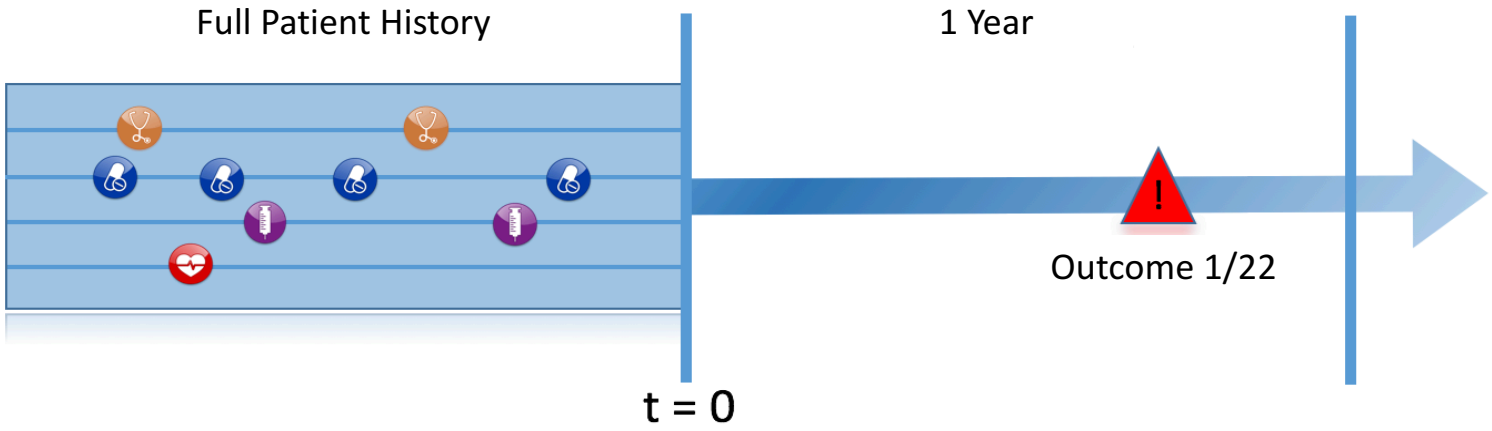
Dissemination of study results should follow the minimum requirements as stated in the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement [1].

- Internal and external validation
- Sharing of full model details
- Sharing of all analyses code to allow full reproducibility

Website to share protocol, code, models and results for all databases

[1] Moons, KG et al. Ann Intern Med. 2015;162(1):W1-73

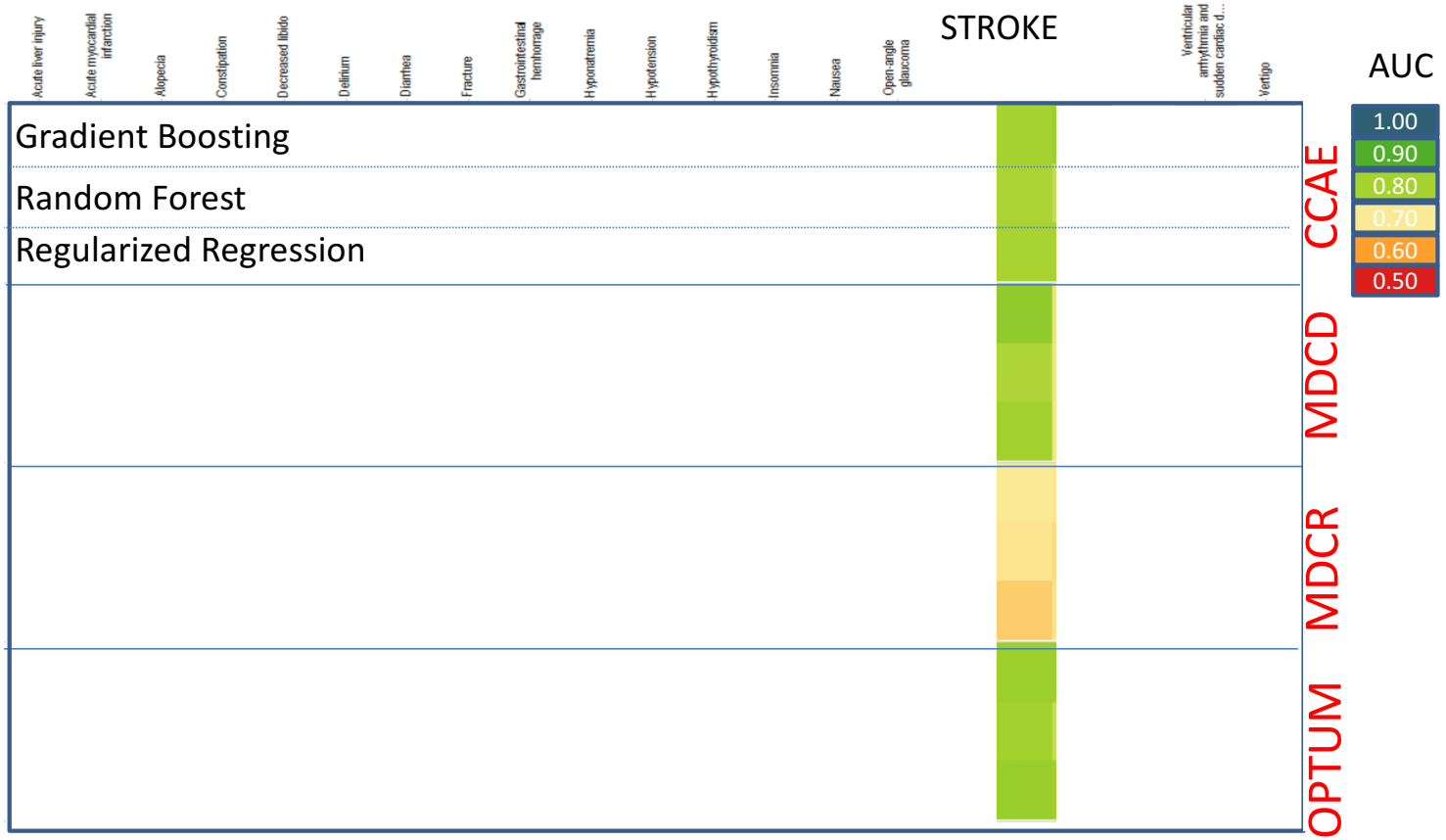
Example Study



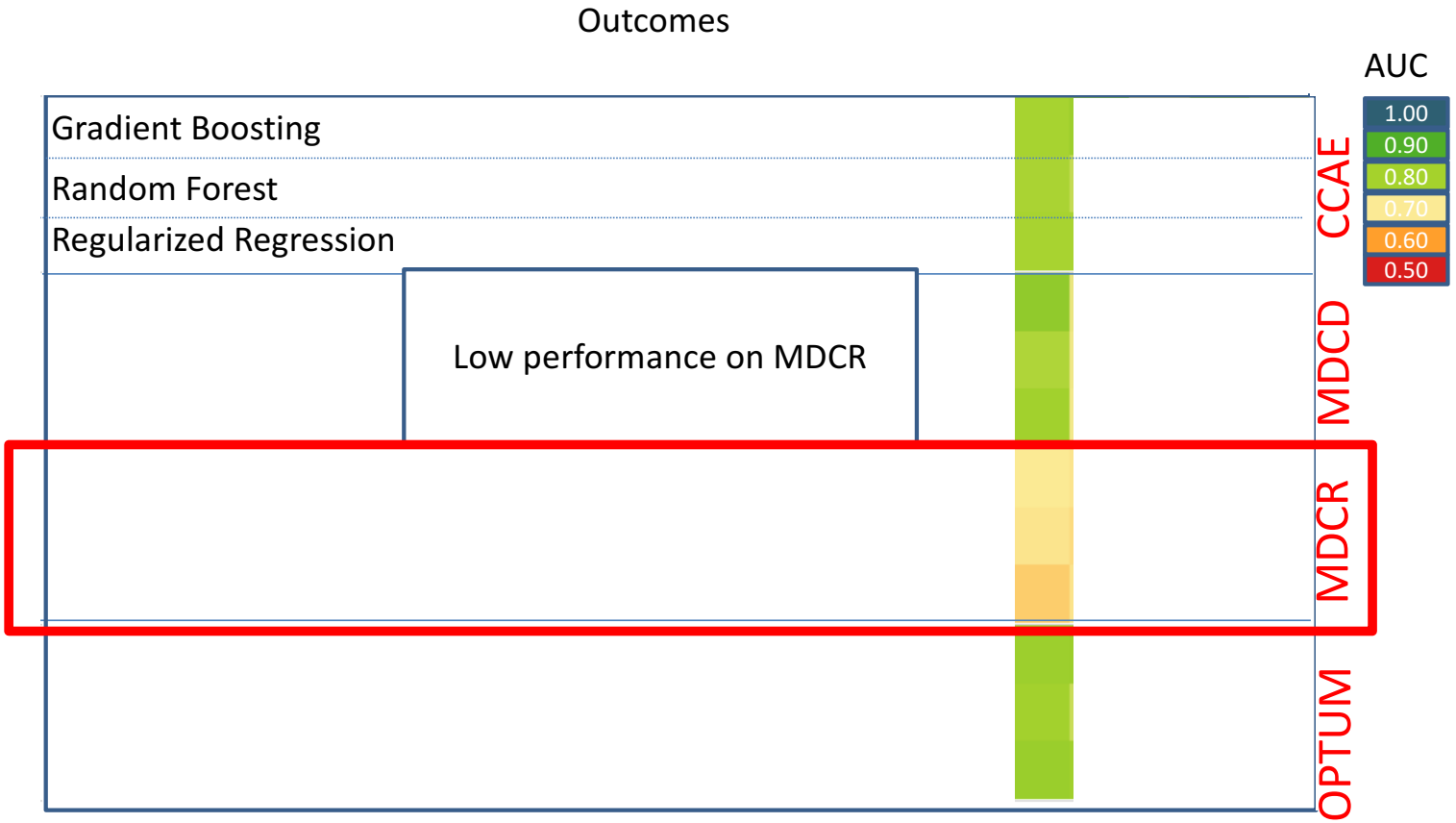
First Pharmaceutically Treated Depression

Among patients **in 4 different databases**, we aim to develop prediction models to predict which patients at a defined moment in time (**First Pharmaceutically Treated Depression Event**) will experience one out of **22 different outcomes** during a time-at-risk (**1 year**). Prediction is done using **all demographics, conditions, and drug use** data prior to that moment in time.

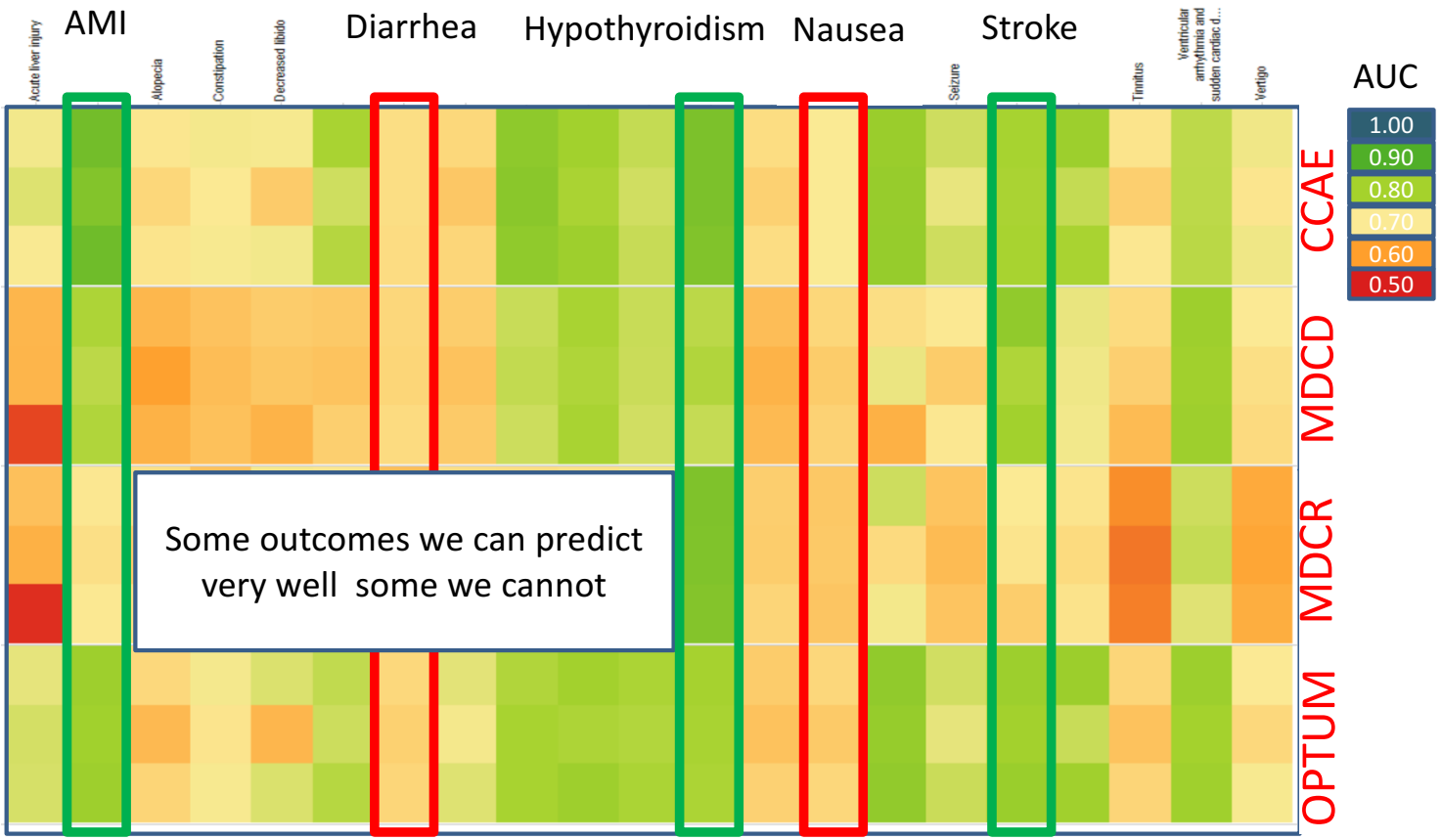
Model Discrimination Stroke



Model Discrimination

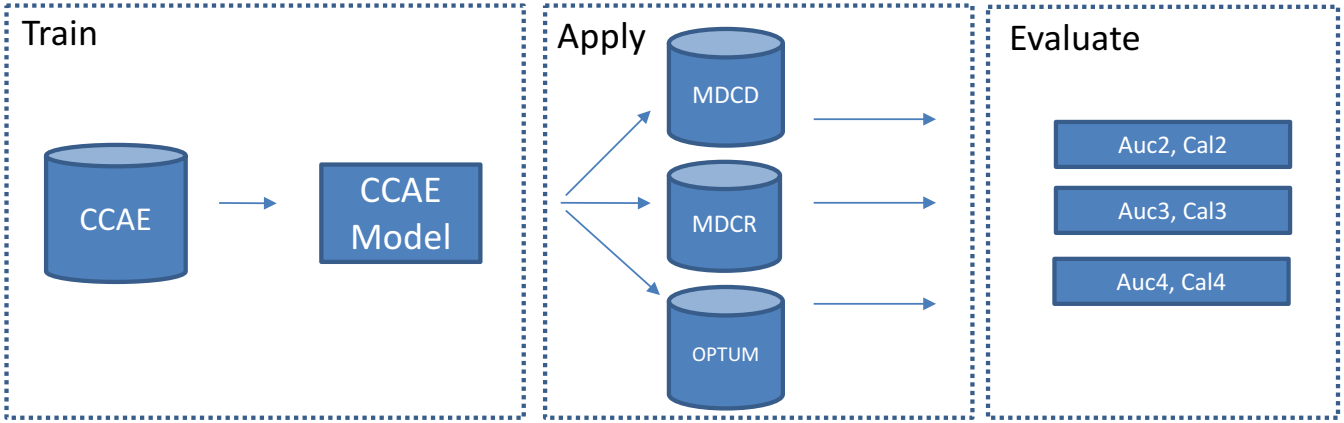


Model Discrimination



Transportability Assessment

How well do the models perform on other databases?

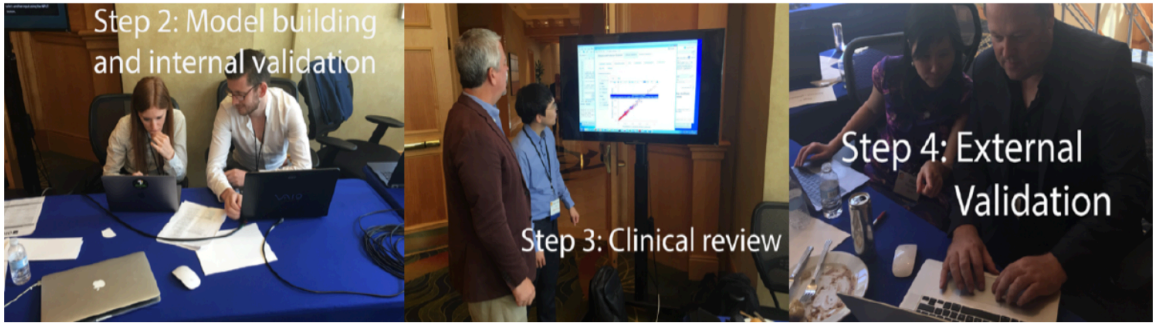


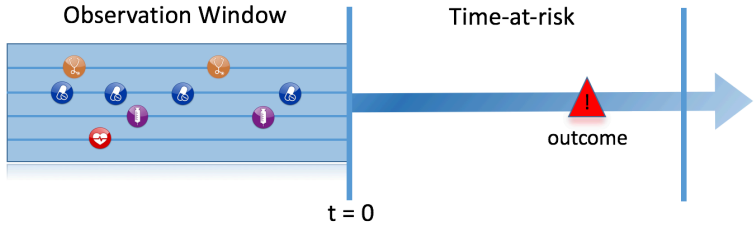
Patient-Level Prediction Workgroup

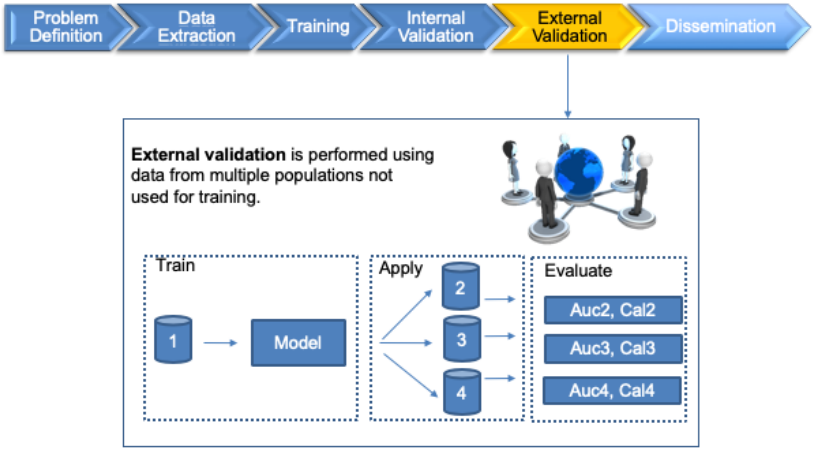
OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Patient-Level Prediction Team Work







www.github.com/OHDSI/PatientLevelPrediction

Jenna M Reps, Martijn J Schuemie, Marc A Suchard, Patrick B Ryan, Peter R Rijnbeek; Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data, Journal of the American Medical Informatics Association, Volume 25, Issue 8, 1 August 2018, Pages 969–975, https://doi.org/10.1093/jamia/ocy032

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

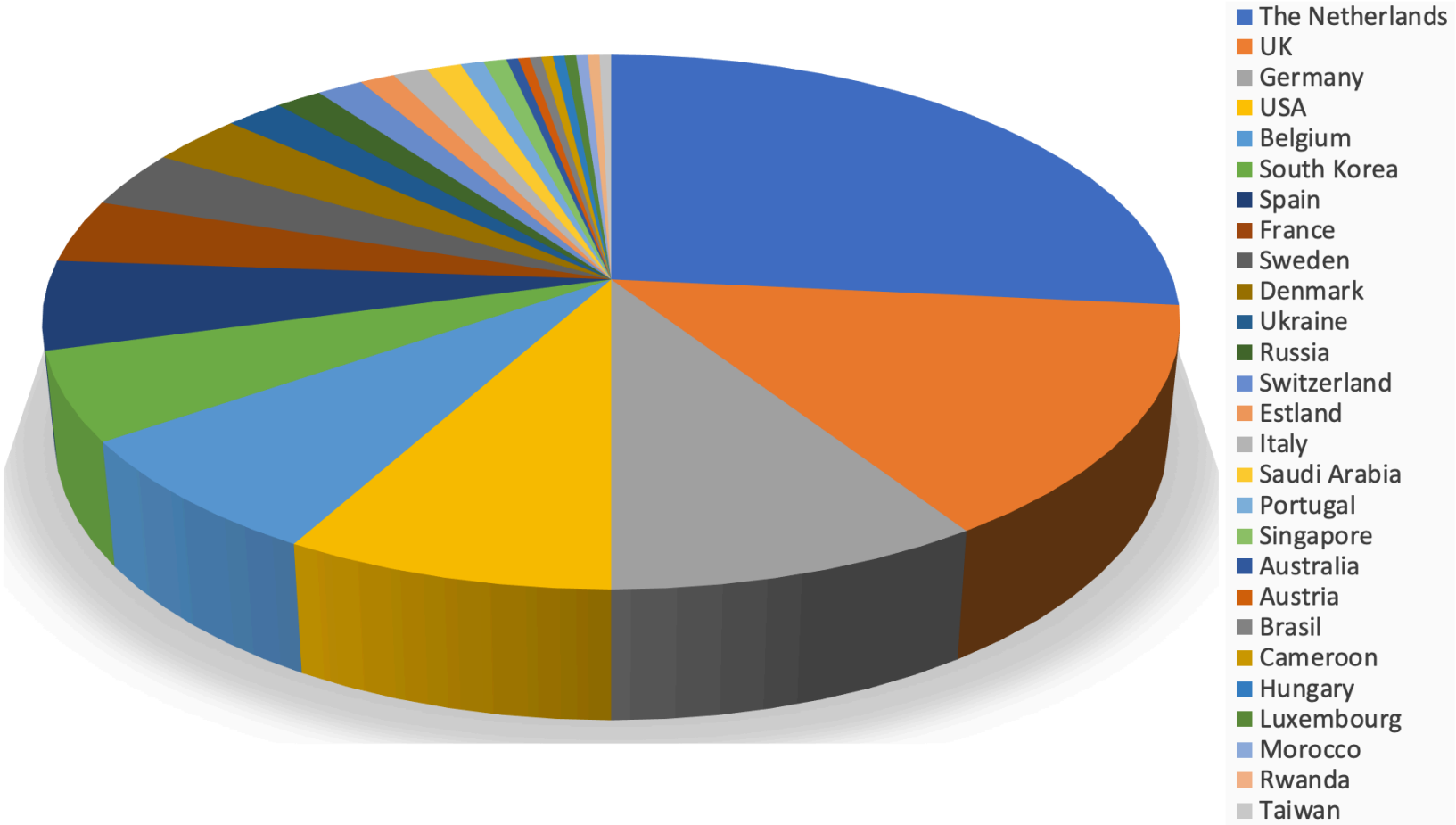
First Annual OHDSI Symposium, March 23th 2018

Meeting Goals Second OHDSI Symposium

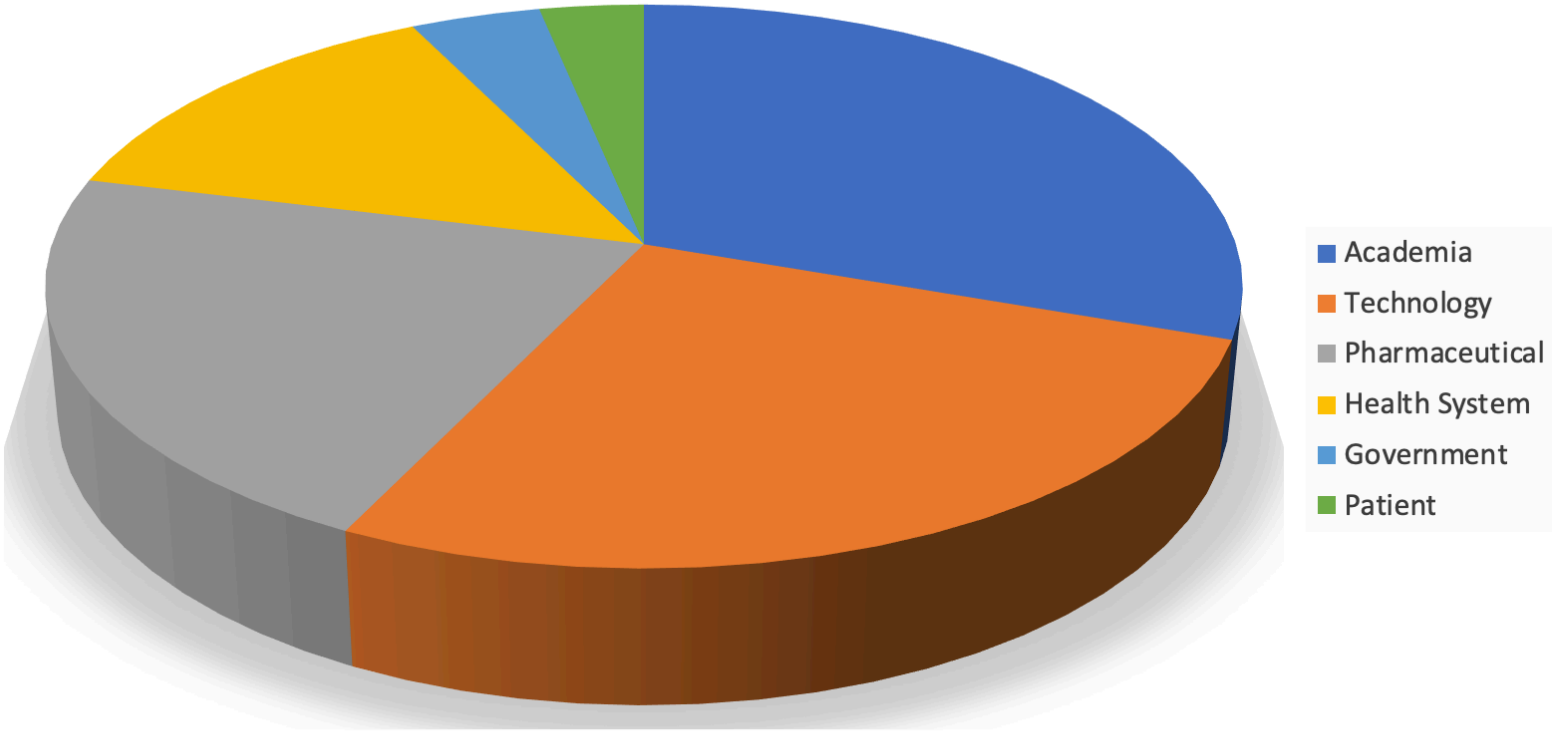
- Provide a platform to stimulate community building
- Demonstrate the OHDSI approach to Reliable and Reproducible Evidence Generation
- Educate and train the community

First Annual OHDSI Symposium, March 23th 2018

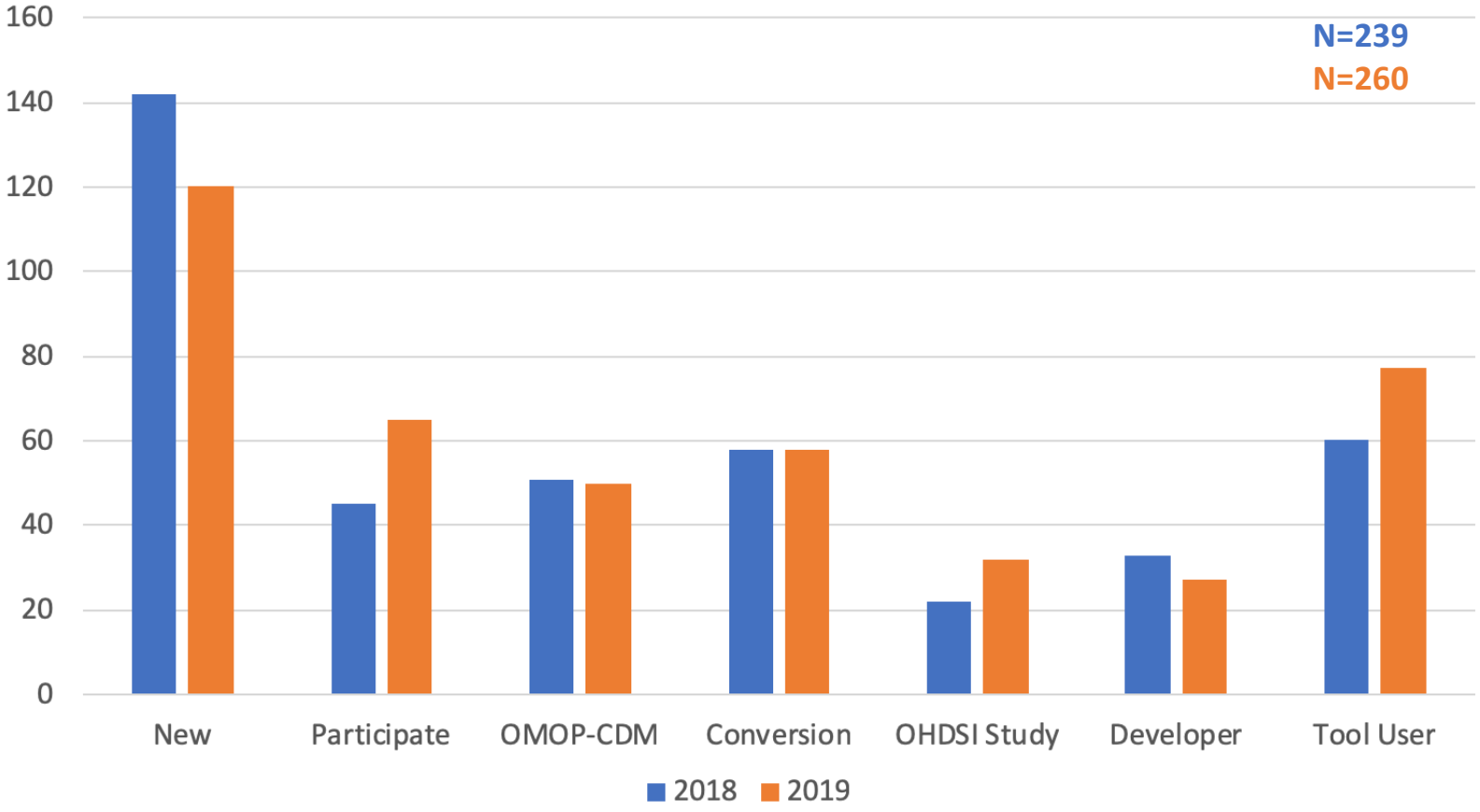
Breakdown of Participants: Countries



Breakdown of Participants: Stakeholders



Relationship with OHDSI



The Journey From Data to Evidence

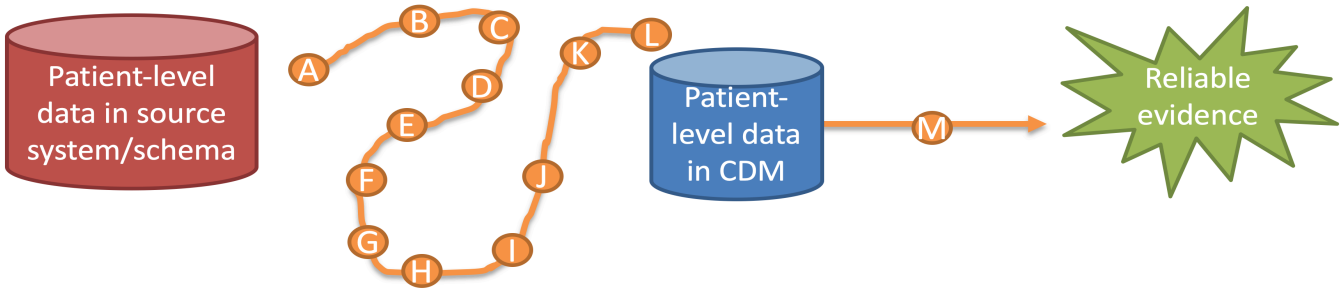
Lightning Talks

35 Posters and 8 Software Demos

8 Tutorials with 150 participants

- OMOP-CDM and Vocabularies
- Extract Transform Load
- Cohort Building
- Patient-Level Prediction
- Population-Level Effect Estimation

We collaboratively need to make the next steps!



- Increase the number of data sources in the open network
- Extend the vocabularies for the European market
- Optimize standardized procedures for quality control
- Further improve and extend the analytical toolset
- Build a sustainable eco-system for mapping support, study execution etc.
- Train all our stakeholders

EHDEN

EUROPEAN HEALTH DATA & EVIDENCE NETWORK

The European Health Data and Evidence Network (EHDEN) Project

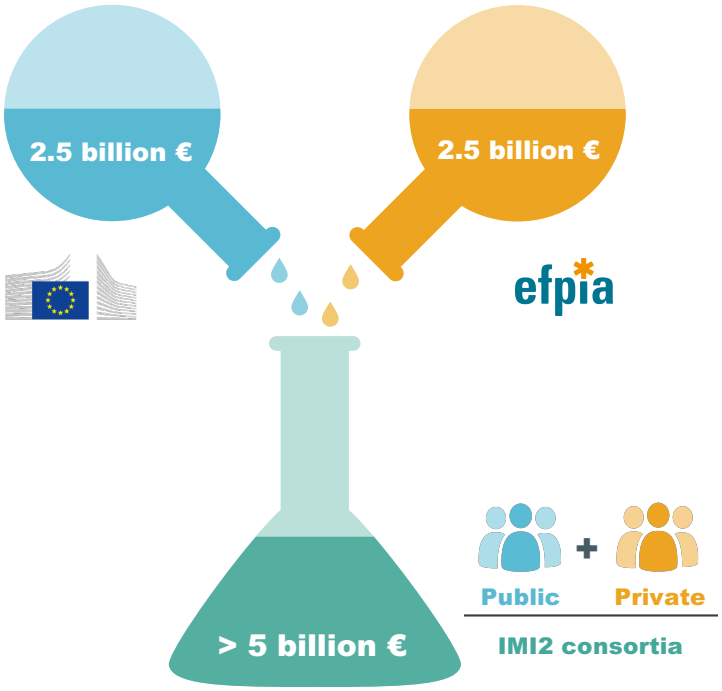
Peter Rijnbeek, PhD

Associate Professor Health Data Science
Department of Medical Informatics
Erasmus MC, The Netherlands

imi efpia

INNOVATIVE MEDICINES INITIATIVE

www.imi.europa.eu



IMI 1 (2008-2014)

2 bn € budget

59 projects

IMI 2 (2014-2024)

3.3 bn € budget

More ambitious, more open & greater scope

EHDEN Project Facts

Start date: 1st Nov 2018

End date: 30th Apr 2024

Total duration: 66 months

EHDEN
EUROPEAN HEALTH DATA & EVIDENCE NETWORK

Total budget: 28,917,357€

IMI2 Funding: 14,105,750€ (7M Harmonization Fund)

EFPIA contribution: 14,811,607€ (10M Harmonization Fund)

EHDEN HAD ITS KICK-OFF MEETING IN JANUARY

EHDEN: Vision and Mission

Vision

The European Health Data & Evidence Network (EHDEN) aspires to be the trusted observational research ecosystem to enable better health decisions, outcomes and care

Mission

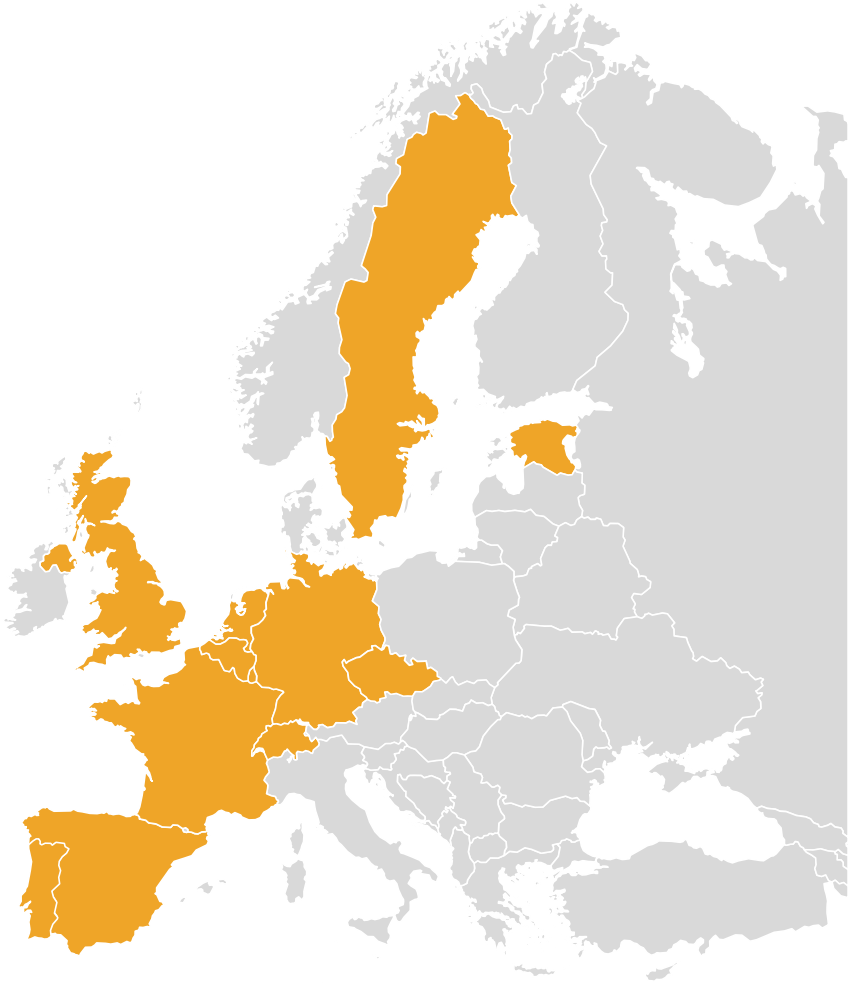
Our mission is to provide a new paradigm for the discovery and analysis of health data in Europe, by building a large-scale, federated network of data sources standardized to a common data model

EHDEN CONSORTIUM

BAYER · AstraZeneca · Celgene · Lilly · Janssen · abbvie · Pfizer · SERVIER · NOVARTIS · SANOFI · UCB

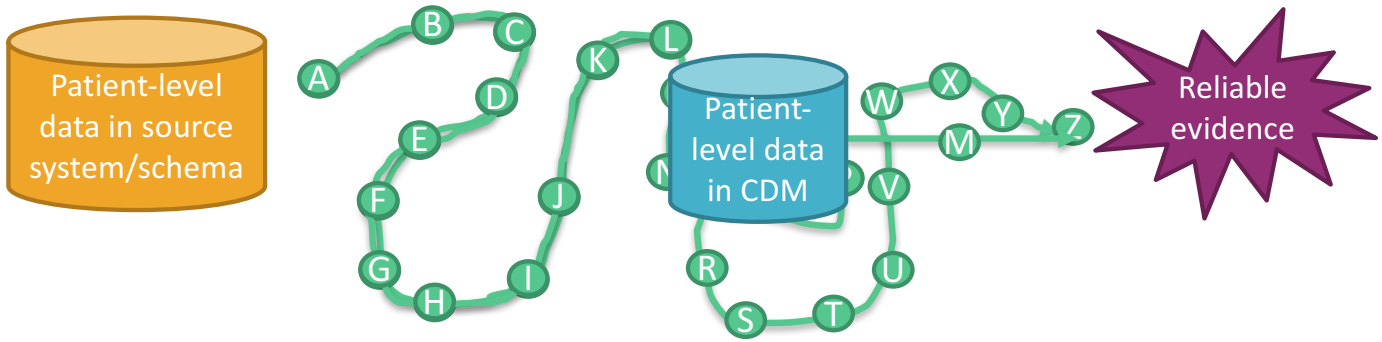
22 PARTNERS

Erasmus MC · SYNAPSE Research Management Partners · Tartu Ülikool 1632 Universitas Tartuensis · EPF European Patients Forum · ODYSSEUS Data Services Inc · NICE National Institute for Health and Care Excellence · ICHOM · University of Oxford · universidade de aveiro · Uppsala Monitoring Centre · thehyve



imi · efpia

GENERATING RELIABLE EVIDENCE USING THE OMOP CDM



EHDEN will build on expertise and tools from prior IMI projects, such as EMIF, and will collaborate intensively with the global OHDSI community.

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

BUILDING THE ECO-SYSTEM: PUTTING THE PIECES TOGETHER

BUILDING THE ECO-SYSTEM: HARMONISATION

The aim is to map 100 million health records across the EU via a common data model (OMOP), supporting research within a federated network, the BD4BO IMI2 programme, and outcomes-based healthcare

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

- Standards development within OHDSI
- Certified & qualified Small to Medium-sized Enterprises (SMEs) in the EU
- Harmonisation fund to support mapping to CDM

THE EHDEN OPEN CALL PROCESS

EHDEN's Data Source Prioritisation Committee governs the scope of the open calls for data sources.

Community of data sources

Open calls

Series of in-project calls across the EU

Tailored for project objectives and sustainability

Grant awarding

Grant applications are evaluated via a pre-defined set of criteria

Data source prioritisation committee, consisting of both internal and external experts, ranks the grant applications

Supporting SMEs

Open calls

Series of in-project calls across the EU

Focusing on SMEs able to support mapping and sustainability

Training & Certification

Training and certification will ensure high-quality mapping procedures

SME certification committee prioritizes SMEs for training and certification

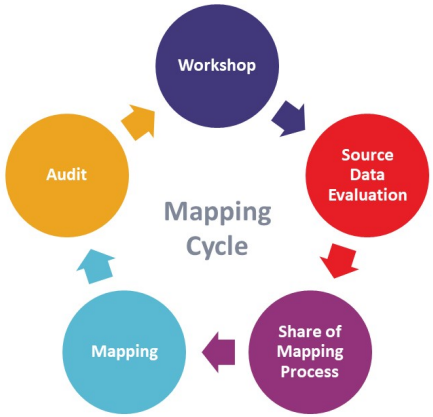
Widespread geographic coverage of Europe by certified SMEs is desired.

EHDEN Harmonisation fund (up to 100 000 € / grant)

Data sources can choose the SME they want to work with from the pool of EHDEN certified SMEs

SMEs performing mappings will be paid by the data sources, via a grant from the harmonisation fund

Payments are milestone based



Mapped data sources are encouraged to be active members of the EHDEN community, participating in research studies.

SME Certification Procedure: Pilot Call

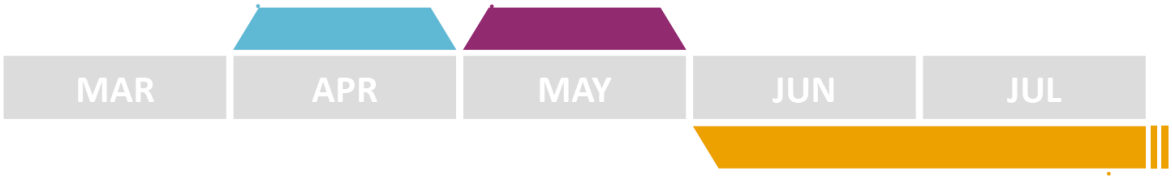
SME call proces overview

Open call

You will be able to submit applications for the open call for SMEs from the 1st of April untill the 30th of April via the EHDEN website.

Evaluation

Following an eligibilty check, applications will be evaluated by the SME certification committee.



Training & Certification

Certification and training of selected SMEs in all necessary competencies.

Eligibility criteria

The call is open to Small and Medium-sized Enterprises under the official EU definition:

- Less then 250 staff headcount.
- <= 50m € turnover or <= 43m € balance sheet total.
- Legally established as a business and based in an EU member state or H2020 associated country.

For more details, see **bit.ly/EU_SME**

SME Certification Procedure

Training

EHDEN's aim is that all selected SMEs as companies will obtain all the necessary competencies to work with a data source within the EHDEN federated network following the established quality standards.

Training topics include, but are not limited to:

- The EHDEN project and its objectives
- OMOP-CDM and the Standardized Vocabularies
- ETL steps and their implementation with OHDSI tools and approaches developed in EHDEN.
- Fundamentals of proper documentation of the ETL process to assure transparency and reproducibility.
- Expertise in the installation and use of the OHDSI Tools for federated data analyses

The course material will be made available through our online learning platform, the EHDEN academy (**https://academy.ehden.eu**).

Certification

The final step in the certification program is a face-to-face training and assessment.

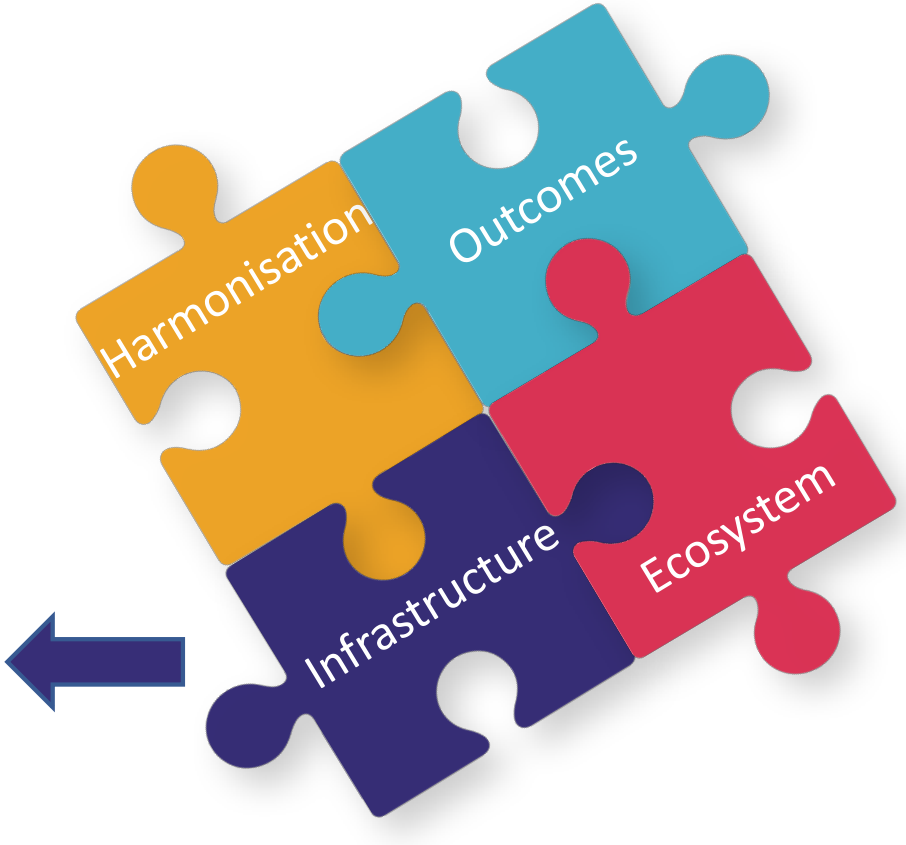
Once the SME receives the certificate it obtains the following privileges:

- Member of the EHDEN SME Community
- Listed in the Certified SME Catalogue
- Use of the EHDEN Certification Badge

Following certification, SMEs can support and map data sources which have received an EHDEN grant.

BUILDING THE ECO-SYSTEM: INFRASTRUCTURE

- Build the infrastructure to enable federated studies on an unprecedented scale in Europe
- Leverage and further develop OHDSI and other tools for high quality analyses
- Use Case driven development
- Ensure interoperability on a global level

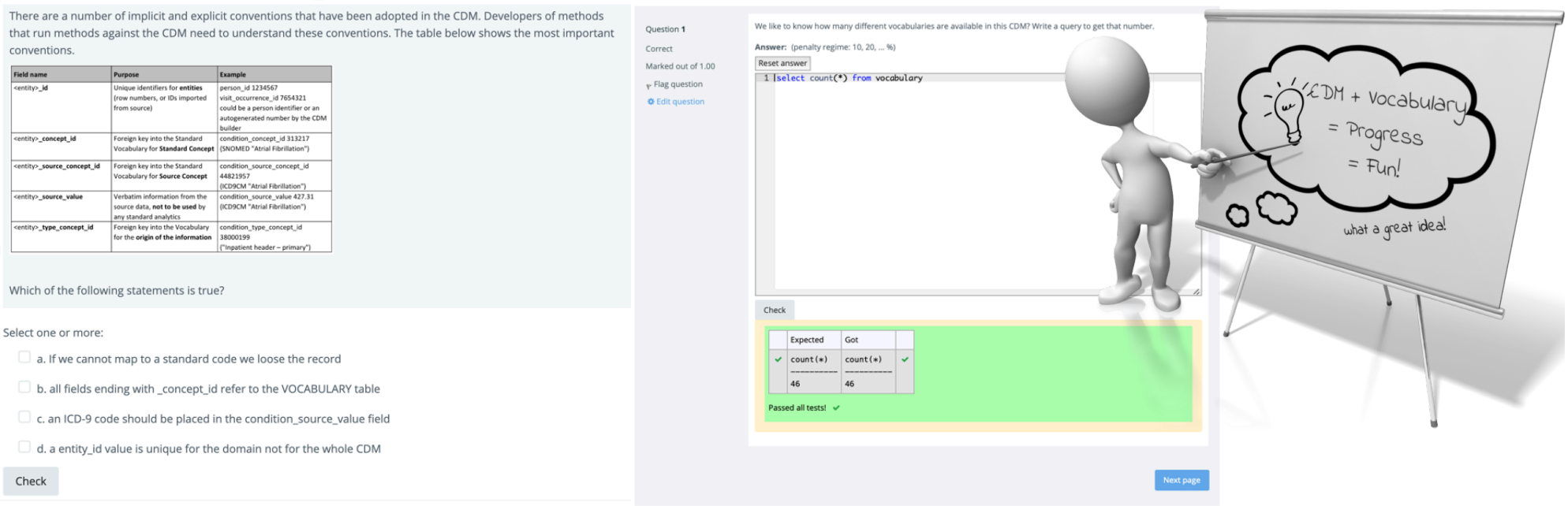


THE EHDEN ACADEMY

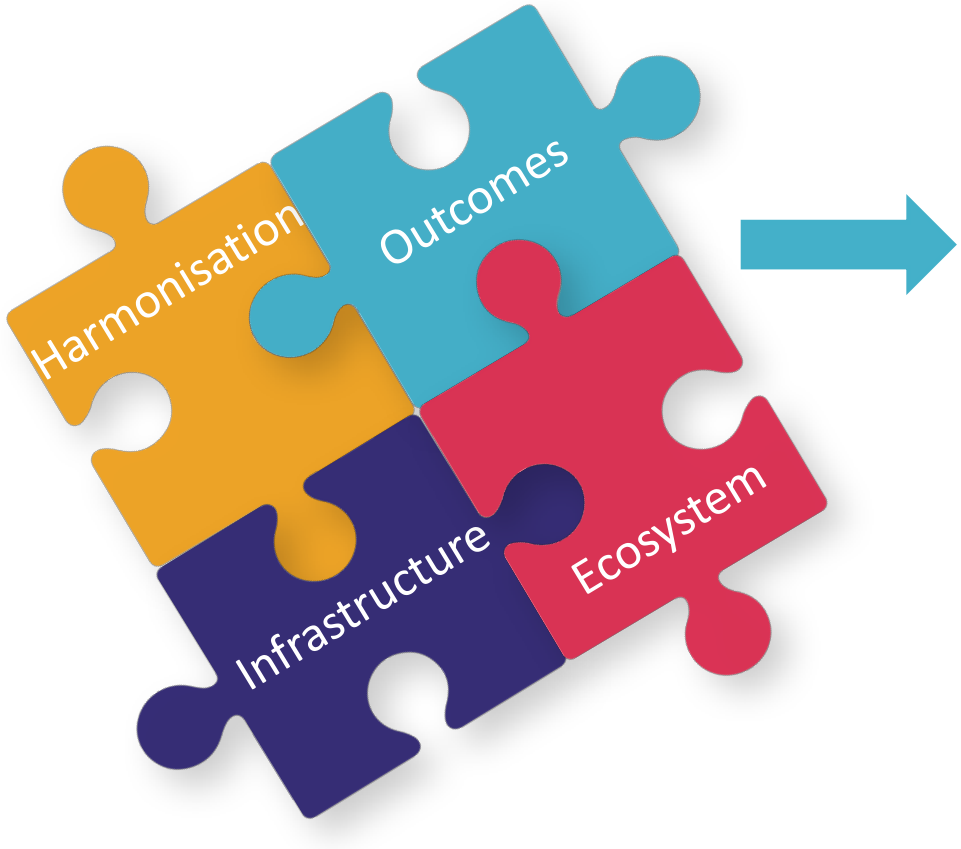
EHDEN is developing the EHDEN Academy (academy.ehden.eu):
an online learning platform that will host courses from OHDSI and EHDEN.

EHDEN Academy

It will contain: Video Lectures, Quizzes, SQL Questions, OHDSI-IN-A-BOX VM integration, and more..



BUILDING THE ECO-SYSTEM: OUTCOMES

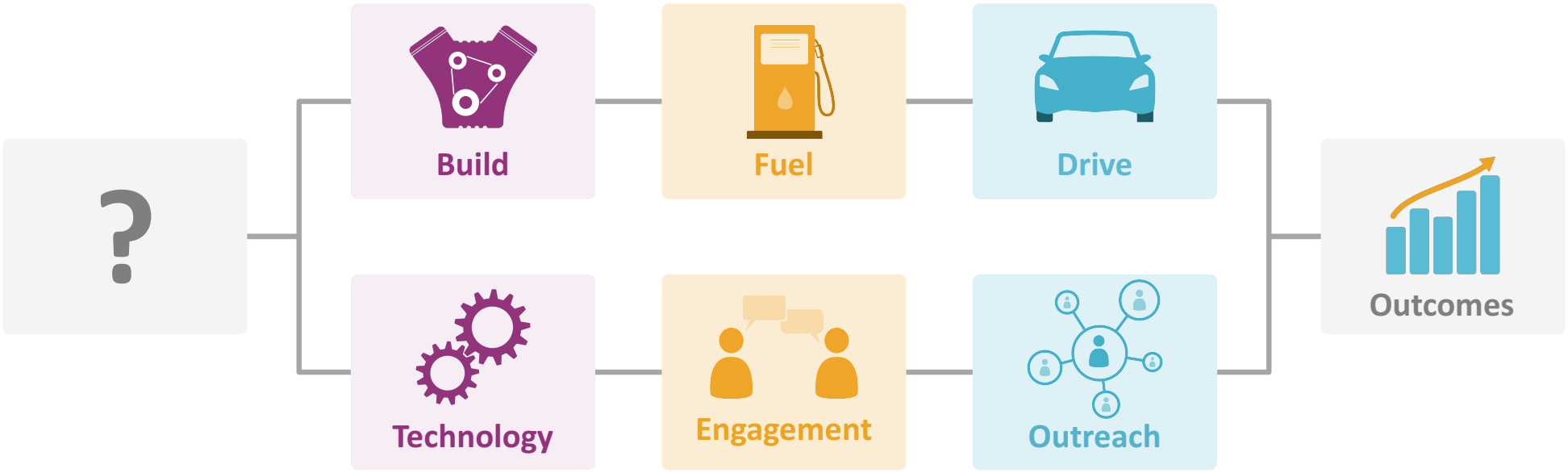


- Incorporation of outcome standards (ICHOM)
- Supporting outcomes-based research and medicine
- Demonstrate the additive value of EHDEN through use cases

BUILDING THE ECO-SYSTEM: ECOSYSTEM

- A trusted federated data network expanding beyond EHDEN project lifetime
- European-wide network of certified SMEs
- Enable new and augmented health services during and after EHDEN
- Stimulate active collaboration within the community
- Sustainability is a responsibility of all stakeholders

KEY COMPONENTS FOR EHDEN



Work Package Structure

WP1: Evidence Workflow Development

Incorporating the use cases for supporting development and validation of the EHDEN socio-technical approach, inclusive of BD4BO projects

WP2: Outcome Driven Healthcare

Related to all activities specific to e.g. BD4BO projects outcome focus, and ICHOM standards incorporation

WP3: Personalized Medicine

Focusing on the support of outcomes/value based healthcare, inclusive of clinical prediction models, with the incorporation of 'novel' patient data

WP4: Technical Implementation

Key priority is socio-technical development of the EHDEN federated framework and relevant services

WP5: Data Workflow Implementation & Service Deployment

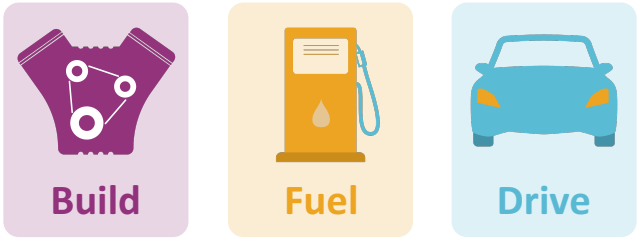
Development, oversight and evaluation of the ecosystem development from SME qualification/ certification to data source engagement, OMOP CDM mapping and evaluation

WP6: Outreach and Sustainability

Ensuring the development of value propositions for key stakeholders, and developing the sustainable operational model for EHDEN during and post IMI phase

WP7: Project Management and Dissemination

Concentrating on intra-project project management, internal communications and external dissemination, and responding to IMI deliverables



Final Remarks

EHDEN and OHDSI will collaboratively work on the implementation of an ecosystem for federated analysis in Europe at an unprecedented scale.

We hope that we can further intensify our collaboration with SNOMED to speak the same language. We are convinced this will empower our research community to generate the reliable evidence our patients need!

NEED MORE INFORMATION?

EHDEN
EUROPEAN HEALTH DATA & EVIDENCE NETWORK

enquiries@ehden.eu

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

www.ohdsi.org

www.ehden.eu

@IMI_EHDEN

IMI_EHDEN

github.com/EHDEN

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.