



How to extract transform and load observational data?

Martijn Schuemie

Janssen Research & Development

Department of Pharmacology & Pharmacy,
The University of Hong Kong



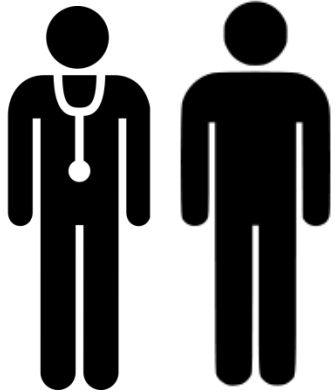
Outline

- Observational data & research networks
- OMOP Common Data Model (CDM)
- Transforming data to the CDM
- Available tools for the CDM

Observational data & research networks



Observational health data



Subjective:

- Complaint and symptoms
- Medical history



Objective:

- Observations
- Measurements – vital signs, laboratory tests, radiology/ pathology findings

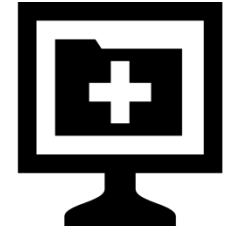
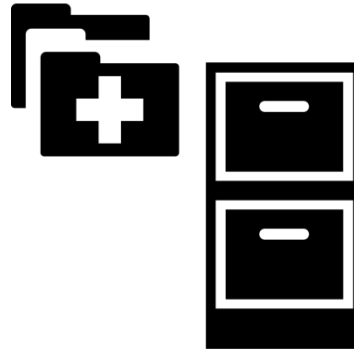
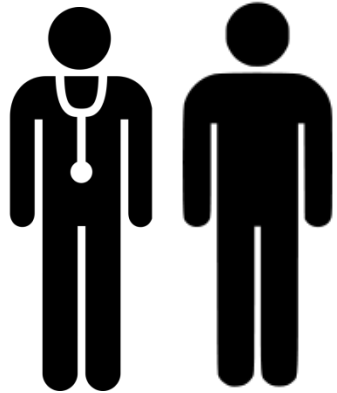


Assessment and Plan:

- Diagnosis
- Treatment



Observational health data



- All captured in the medical record
- Medical records are increasingly being captured in EHR systems
- Data within the EHR are becoming increasingly structured

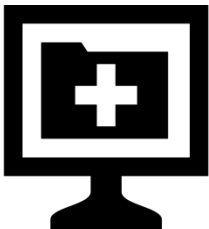
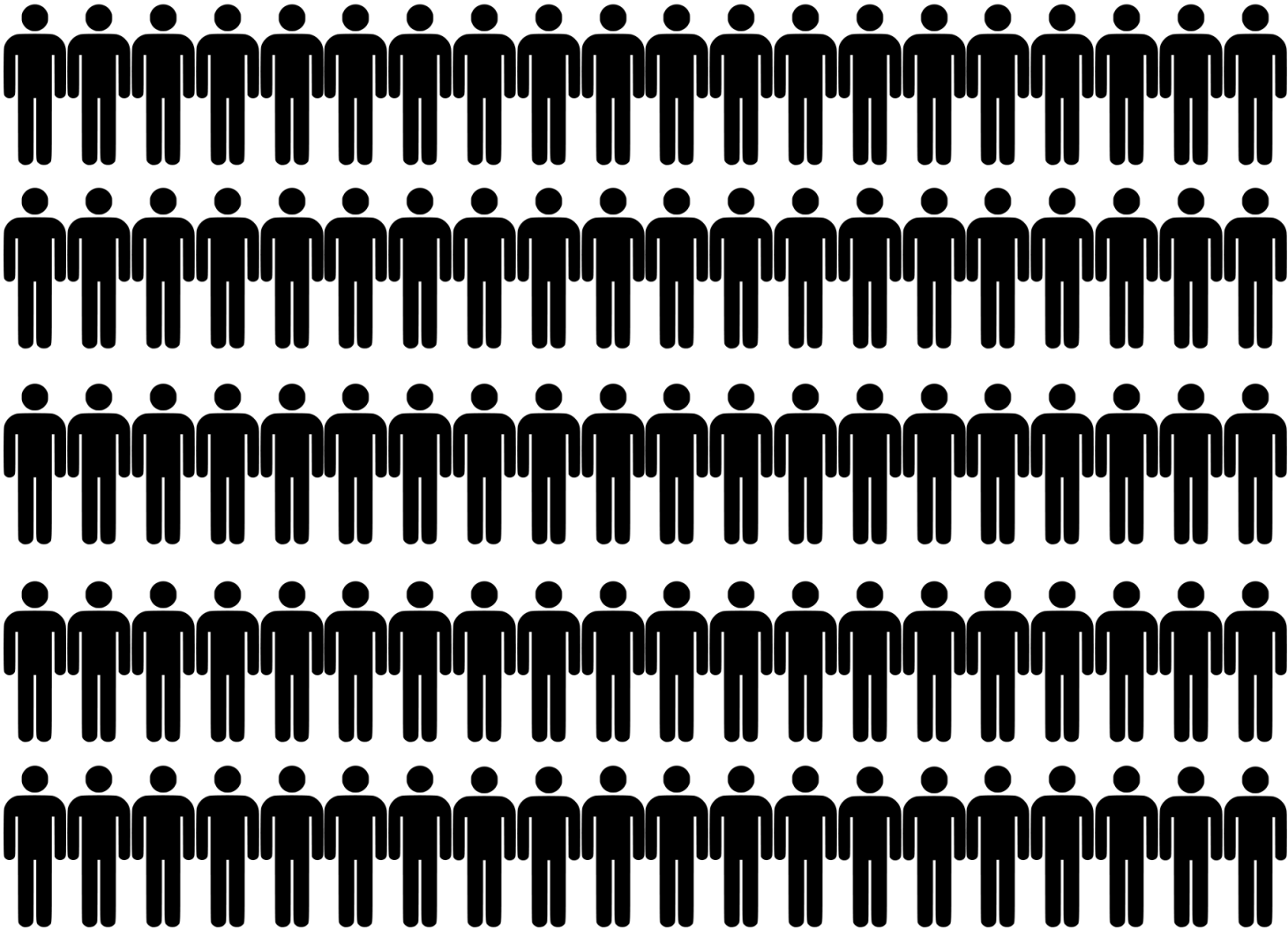


The average primary care physician sees 20 different patients a day



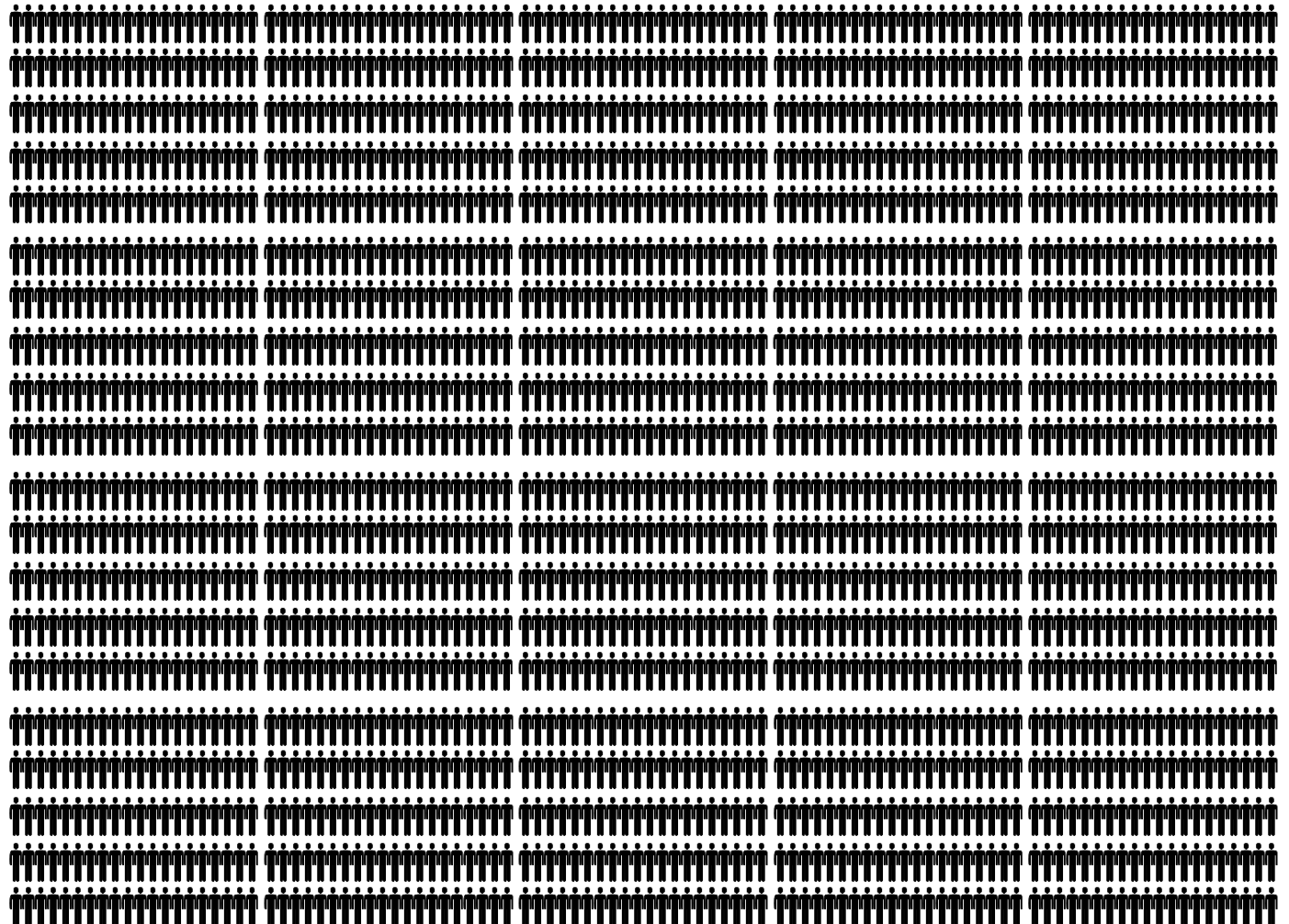


...that's 100 visits in a week...



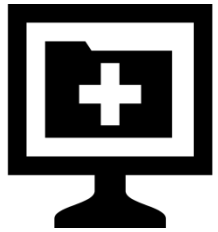


...which can reflect a panel of over 2,000 patients in a year





Databases for research



Evidence
generation



- Electronic Health Records
- Insurance claims
- Registries

- Diseases
- Health care provided
- Effects of treatments
- Differences between patients
- Personalized medicine



Key point

Existing observational health care data such as **Electronic Health Records** and **insurance claims databases** have great potential for research

现有的卫生保健观测数据，例如电子健康记录、保险索赔数据库等具有巨大的潜力为研究所用



Research networks

- Data may be at different sites
- Sites often cannot share data at the patient level
- Data can be in very different formats

Patient level, identifiable
information



Practice



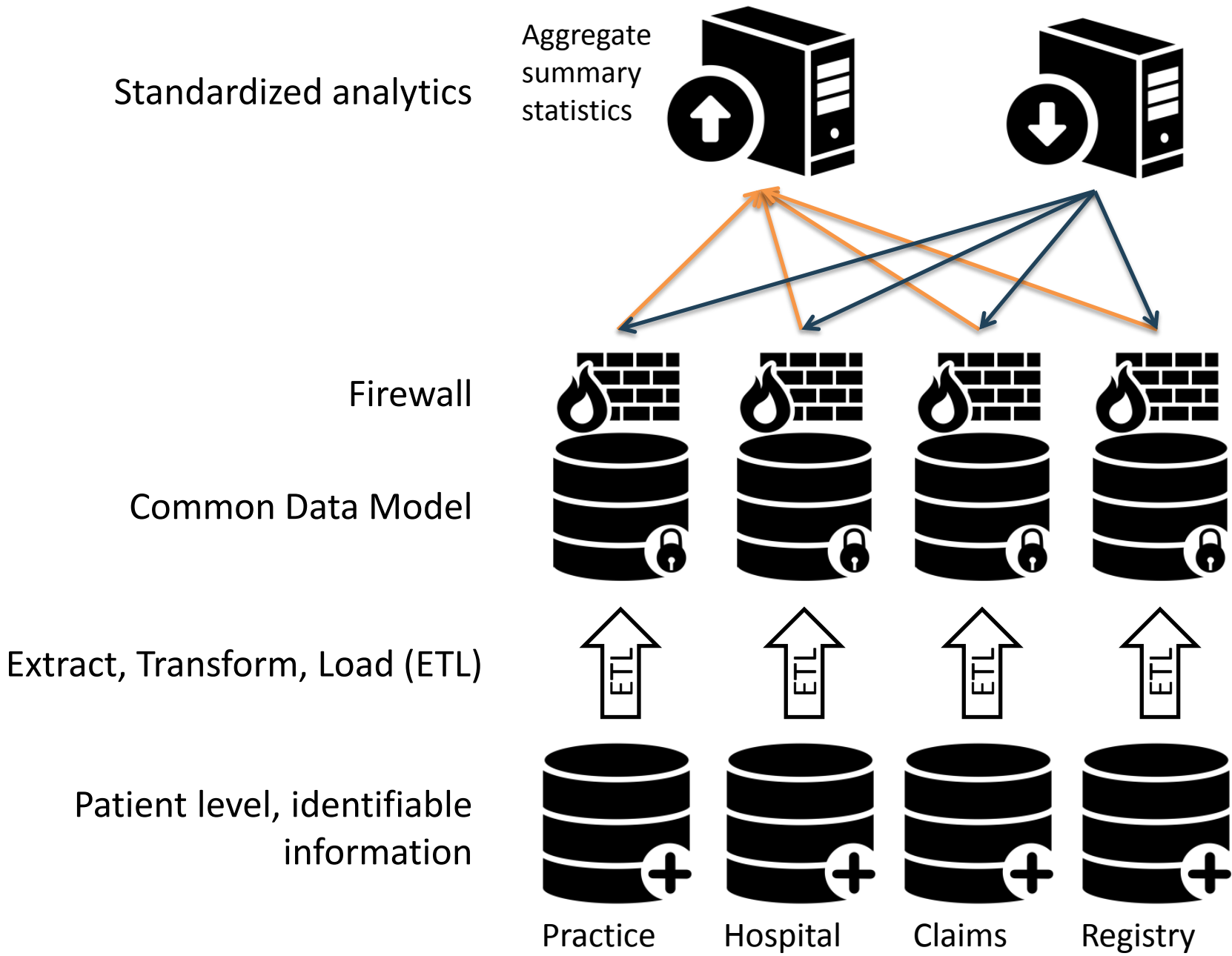
Hospital



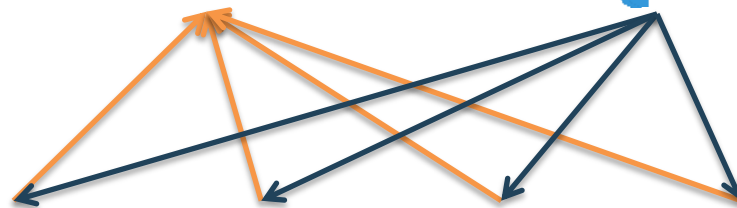
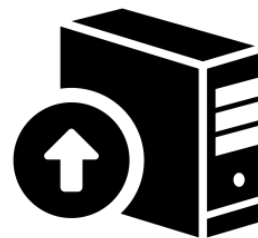
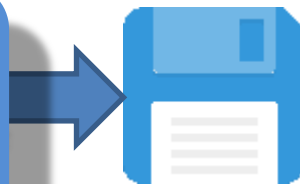
Claims



Registry



Count people on drug A
and B, and the number of
outcomes X



Firewall

Common Data Model



Extract, Transform, Load (ETL)



Patient level, identifiable
information



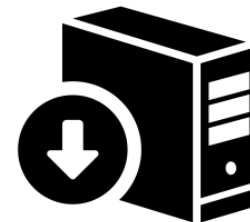
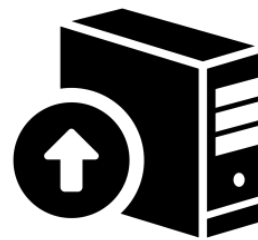
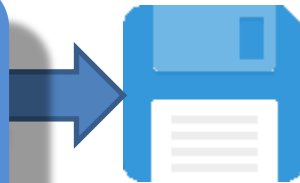
Practice

Hospital

Claims

Registry

Count people on drug A
and B, and the number of
outcomes X



Firewall

Practice:

100 patients on A, 1 has X
200 patients on B, 4 have X

Hospital:

1000 patients on A, 10 have X
2000 patients on B, 40 have X

Etc.



Aggregated data

No privacy concerns

Registry



Distributed research networks

- Europe



- United States



- Asia



- Global





Key point

Observational data can often not be shared.
Using a **common data model** allows analysis programs to ‘visit’ the data instead.

观测数据通常不是共享的。而使用**公共数据模型**可以实现分析程序对不同来源数据的统一“访问”

OMOP Common Data Model (CDM)



Common Data Model

- A common **structure**

Person

- person_id
- year_of_birth
- month_of_birth
- day_of_birth
- gender_concept_id

- A common **vocabulary**

How do we store gender?

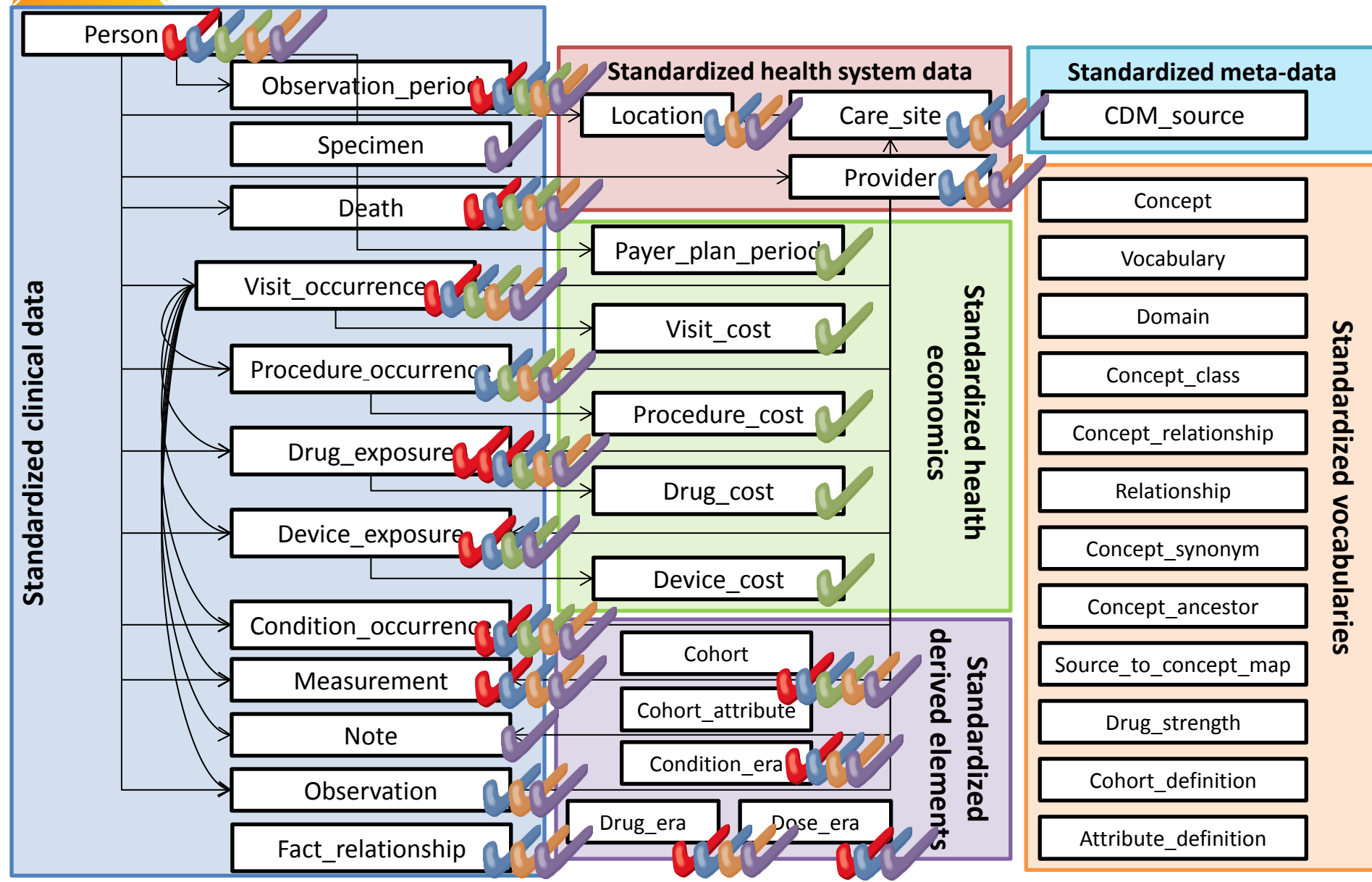
- M, F, U
- 0, 1, 2
- 8507 (male), 8532 (female), 8851 (unknown gender), 8570 (ambiguous gender)



OMOP Common Data Model

- Designed for **various types of data** (EHR, insurance claims) in **various countries**
- Developed by the OMOP / OHDSI community
- Currently on 5th version
- ‘Easy to get data in, easy to get data out’

One model, multiple use cases





CDM principles

- Person centric
- Data is split into domains (e.g. conditions, drugs, procedures)
- Preserve source values and map to standard values
- Store the source of the data
- Separate verbatim data from inferred data



OMOP Vocabulary

All codes are mapped to **standard coding systems**

- Drugs: RxNorm
- Conditions: SNOMED
- Measurements: LOINC
- Procedures: ICD9Proc, HCPCS, CPT-4

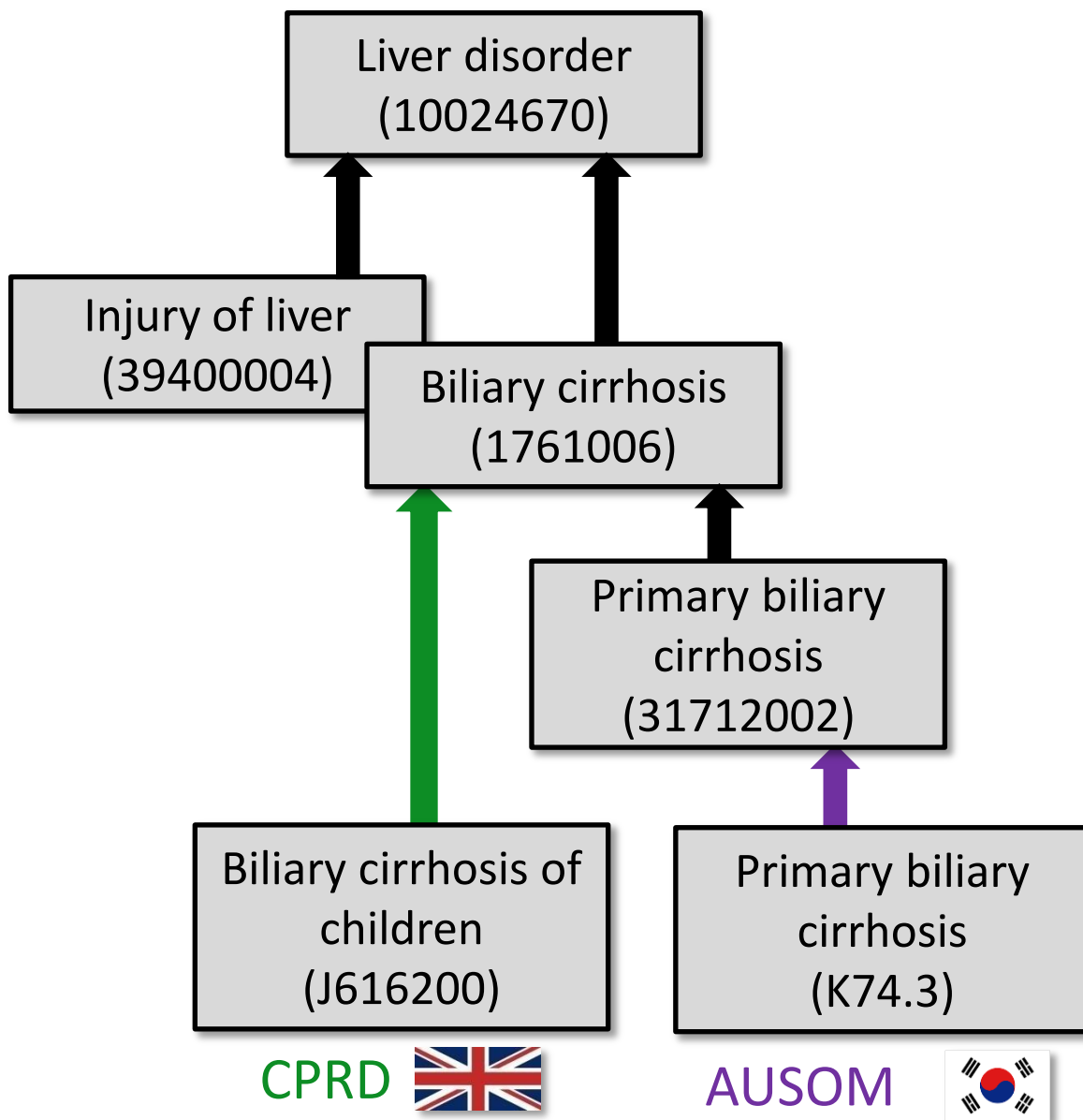


MEDDRA

SNOMED

Source codes

Vocabulary





Key point

The OMOP Common Data Model provides a **common structure** and a **common vocabulary**. Different levels of precision in different coding systems are resolved using a concept hierarchy.

OMOP公共数据模型对数据采用通用结构和通用词汇；利用统一的概念等级解决不同编码系统的各个级别的精确度问题。

Transforming data to the CDM



Process to create an ETL

1. Data experts and CDM experts together design the ETL
2. People with medical knowledge create the code mappings
3. A technical person implements the ETL
4. All are involved in quality control



Tools to support the process

1. Data experts and CDM experts together design the ETL



WHITE RABBIT



RABBIT IN A HAT

2. People with medical knowledge create the code mappings



USAGI

3. A technical person implements the ETL


?

4. All are involved in quality control

ACHILLES





Step 1. Designing the ETL

- Run  **WHITE RABBIT**
 - Scans the source database
 - Creates a report describing
 - Tables
 - Fields in the tables
 - Values in the fields
 - Frequency of values

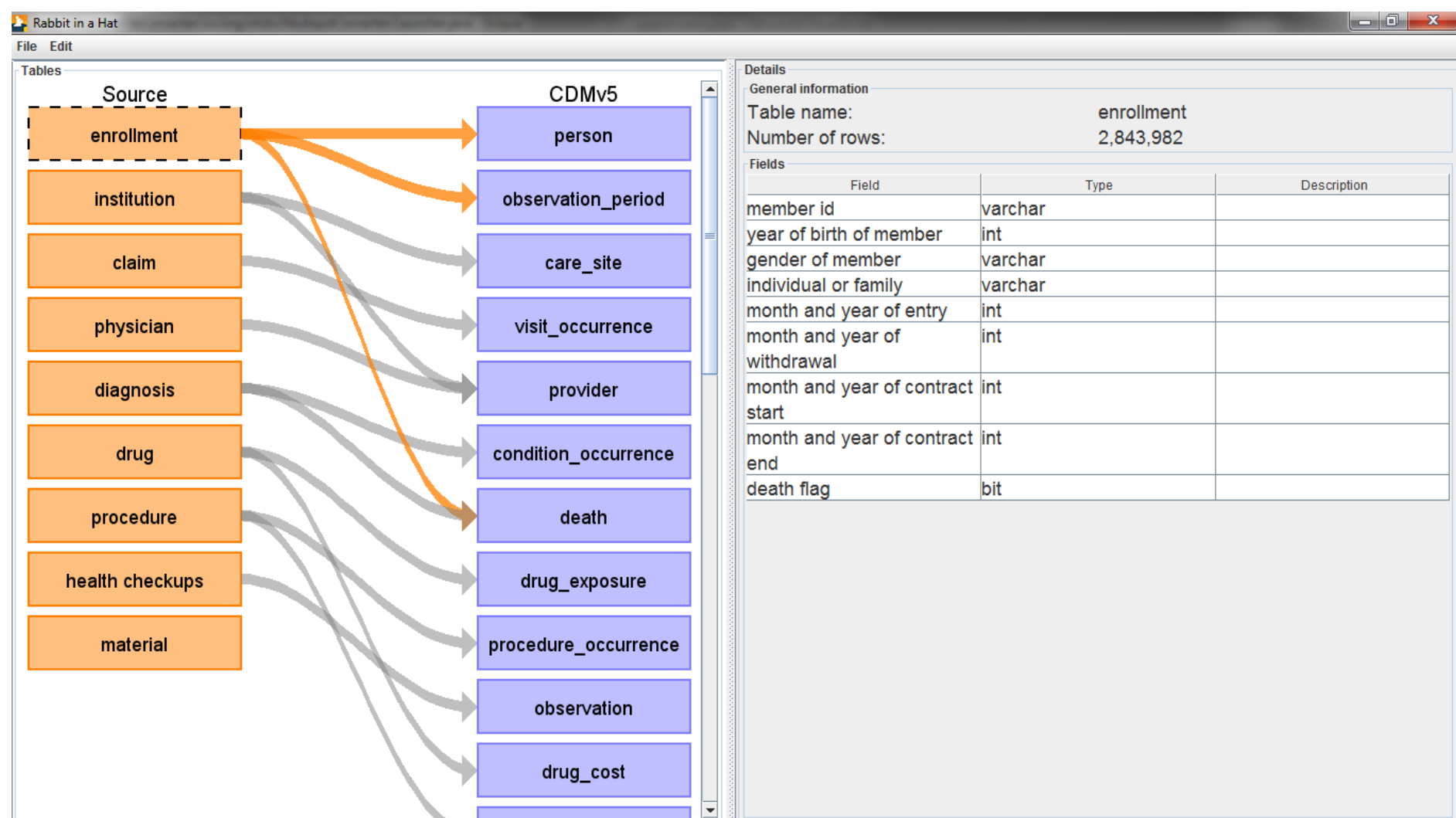


Step 1. Designing the ETL

- Run  **WHITE RABBIT**
- In an interactive session with both data experts and CDM experts, run  **RABBIT IN A HAT**





RABBIT IN A HAT





Step 1. Designing the ETL

- Run  **WHITE RABBIT**
- In an interactive session with both data experts and CDM experts, run  **RABBIT IN A HAT**
- Output document becomes basis of ETL specification




Step 2. Create code mappings

- Need to map codes to one of the OMOP CDM standards:
 - Drugs: RxNorm
 - Conditions: SNOMED
 - Lab values: LOINC
 - Specialties: CMS
 - ...



Step 2. Create code mappings

- Prepare data
 - Codes
 - Descriptions (translated to English if needed)
 - Frequencies
- Use existing mapping information if available
 - If ATC codes are available, use to select subsets
 - Use (partial) mappings in UMLS
- Run  **USAGI**



Usagi

Usagi

File Edit View Help

Status	Source c...	Source te...	Frequency	Dutch term	Match sc...	Term	Concept ID	Concept ...	Concept I...	Concept ...	Vocabulary	Concept ...	Valid start...	Valid end ...	Invalid re...
Unchecked	A99.00	General d...	5774012	Andere g...	0.82	Generaliz...	4244571	Generaliz...	0	Qualifier v...	SNOMED...	60132005	1970-01-...	2099-12-...	
Unchecked	K86.00	Hyperten...	3987206	Essenti...	0.82	uncompli...	4137841	Uncompli...	0	Qualifier v...	SNOMED...	263914008	1970-01-...	2099-12-...	
Unchecked	R44.00	Preventiv...	3702922	Immunis...	0.52	Prevental	4113919	Dichlorop...	0	Substance	SNOMED...	255999002	1970-01-...	2099-12-...	
Approved	T90.02	Diabetis ...	2275799	Diabetes ...	0.49	Type 2 Di...	201826	Type 2 di...	2	Clinical fi...	SNOMED...	44054006	1970-01-...	2099-12-...	
Approved	R05.00	Cough	1268829	Hoesten	0.58	Coughs	254761	Cough	2	Clinical fi...	SNOMED...	49727002	1970-01-...	2099-12-...	
Unchecked	R74.00	Upper re...	1061504	Acute infe...	1.00	acute upp...	40394956	Acute upp...	1	Clinical fi...	SNOMED...	195705000	1970-01-...	2013-01-...	D
Unchecked	A29.00	General s...	1035167	Andere al...	0.71	Wrist sym...	40479644	Wrist sym...	1	Clinical fi...	SNOMED...	441591007	2009-07-...	2099-12-...	

Source code

Source code	Source term	Frequency	Dutch term
R74.00	Upper respiratory infection acute	1061504	Acute infectie bovenste luchtwegen

Target concepts

Term	Concept ID	Concept name	Concept level	Concept class	Vocabulary	Concept code	Valid start date	Valid end date	Invalid reason
acute upper resp...	40394956	Acute upper respirat...	1	Clinical finding	SNOMED-CT	195705000	1970-01-01	2013-01-30	D

Remove concept

Search

Query

☒ Use source term as query

☐ Query:

Filters

☐ Filter by automatically select concepts

☐ Filter by concept class:

☐ Filter invalid concepts

☐ Filter by vocabulary:

Results

Score	Term	Concept ID	Concept name	Concept level	Concept class	Vocabulary	Concept code	Valid start date	Valid end date	Invalid reason
1.00	acute upper re...	40394956	Acute upper re...	1	Clinical finding	SNOMED-CT	195705000	1970-01-01	2013-01-30	D
1.00	Upper respirat...	40394568	Pyrexial cold	1	Clinical finding	SNOMED-CT	195648002	1970-01-01	2013-01-30	D
1.00	Upper respirat...	260427	Common cold	1	Clinical finding	SNOMED-CT	82272006	1970-01-01	2099-12-31	
1.00	Upper respirat...	40345737	Sniffles	1	Clinical finding	SNOMED-CT	266377009	1970-01-01	2013-01-30	D
1.00	acute upper re...	40316063	Acute upper re...	1	Clinical finding	SNOMED-CT	155516001	1970-01-01	2013-01-30	D
1.00	Upper respirat...	40316041	Common cold	1	Clinical finding	SNOMED-CT	155497009	1970-01-01	2013-01-30	D

Replace concept

Add concept

Approved / total: 2 / 1714 3.9% of total frequency

Approve



Step 3. Implement the ETL

No standard. Depends on the expertise of the person doing the work:

- SQL
- SAS
- C++ (CDMBuilder)
- Java (JCDMBuilder)
- Kettle



Step 4. Quality control

Not yet fully defined, but we currently use:

- Manually compare source and CDM information on a sample of persons
- Use Achilles software
- Replicate a study already performed on the source data



Key point

Transforming data to a common data model requires an **interdisciplinary team**, including clinicians, informatics specialists, and people that know the data.

实现数据到公共数据模型的转换需要一个跨学科领域团队的合作，包括临床医生、信息学专家以及了解数据内容的人员

Available tools for the CDM



Introducing OHDSI

- International collaborative





Introducing OHDSI

- International collaborative
- Coordinating center in Columbia University
- Academia and industry
- Data network
- Research (clinical + methodology)
- Open source software

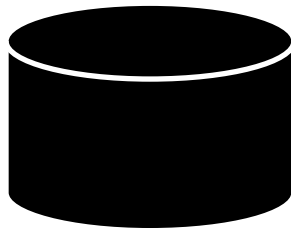


ACHILLES

Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems



Data in CDM



Aggregate statistics:

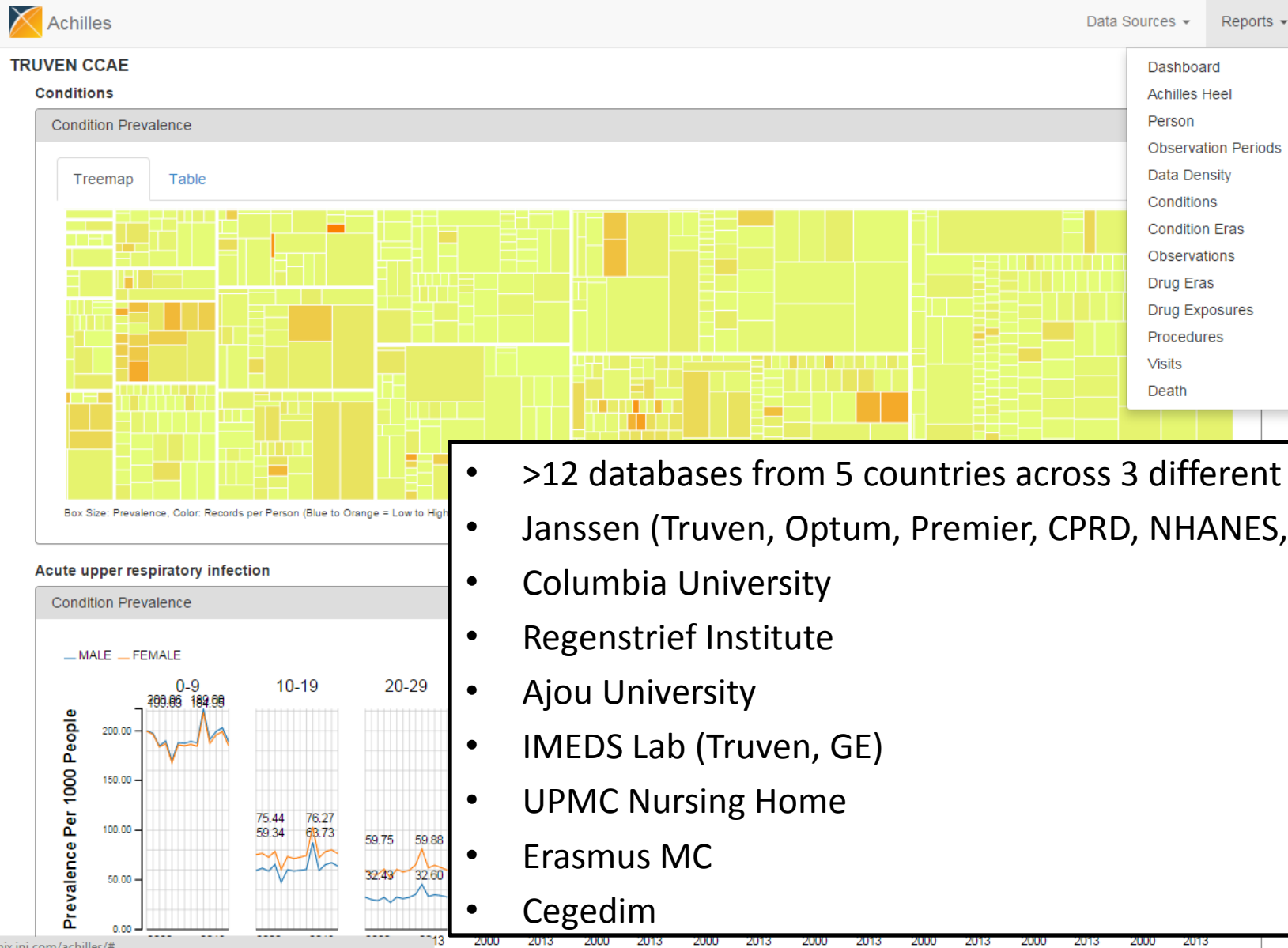
- Persons
- Conditions
- Drugs
- Lab results
- ...



Explore statistics
in a web browser



ACHILLES



- >12 databases from 5 countries across 3 different platforms:
- Janssen (Truven, Optum, Premier, CPRD, NHANES, HCUP)
- Columbia University
- Regenstrief Institute
- Ajou University
- IMEDS Lab (Truven, GE)
- UPMC Nursing Home
- Erasmus MC
- Cegedim



Achilles Heel

Achilles

wprdusmjtglay.wks.jnj.com/achilles/#/HCUP/achillesheel

Apps Reference JnJ OHDSI Chinese EMIF Distractions Hong Kong Welcome · R packag... Labcodeviewer web...

Achilles Data Sources Reports

HCUP

Achilles Heel Report

Data Quality Messages

Search: Show / hide columns

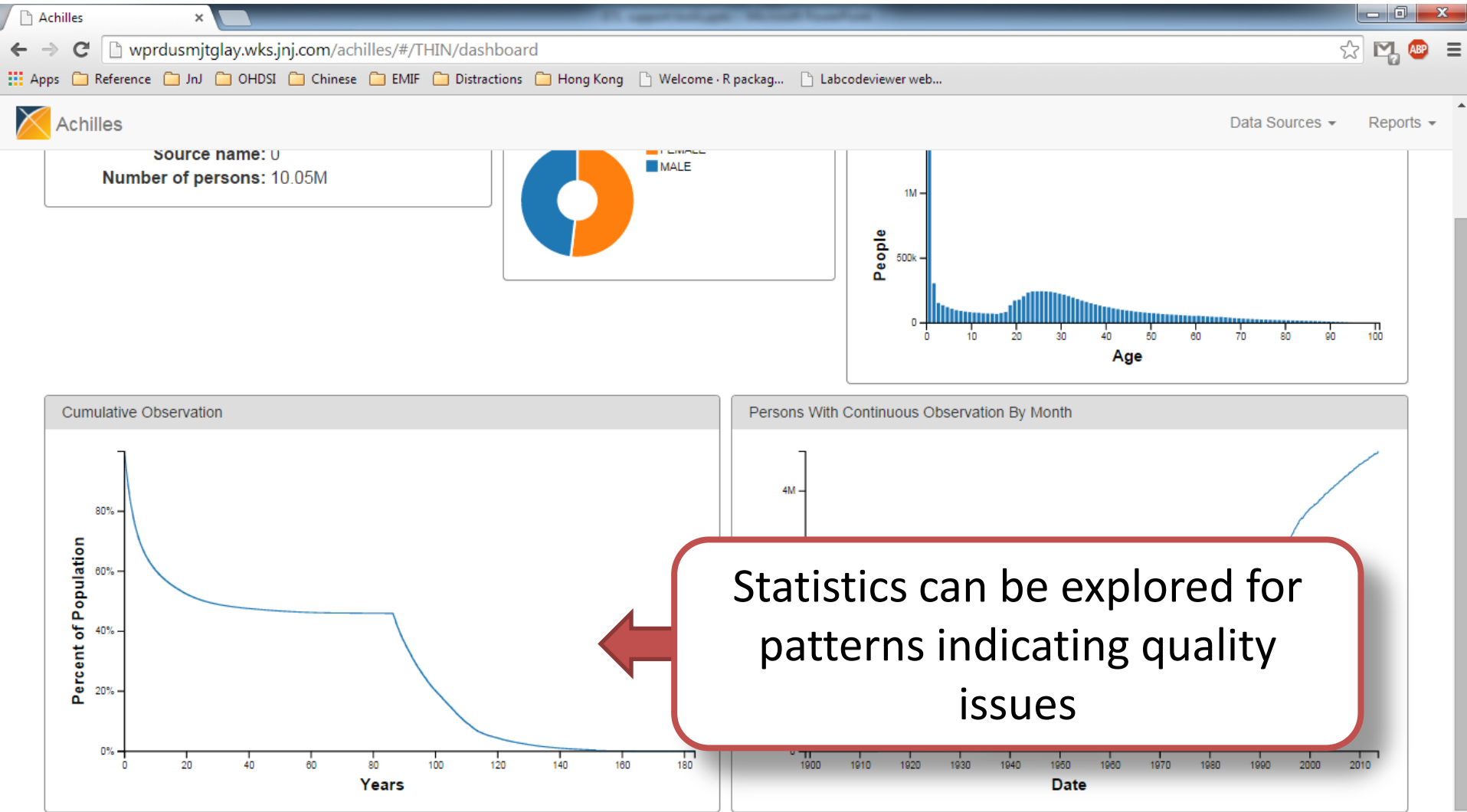
Message Type	Message
ERROR	3-Number of persons by year of birth; should not have year of birth < 1900, (n=6847)
ERROR	101-Number of persons by age, with age at first observation period; should not have age > 100, (n=38865)
ERROR	400-Number of persons with at least one condition occurrence, by condition_concept_id; 121 concepts in data are not in correct vocabulary (SNOMED)
ERROR	1000-Number of persons with at least one condition era, by condition_concept_id; 121 concepts in data are not in correct vocabulary (SNOMED)
WARNING	5-Number of persons by ethnicity; data with unmapped concepts
WARNING	400-Number of persons with at least one condition occurrence, by condition_concept_id; data with unmapped concepts
	art month, by condition_concept_id; 1766 concepts have a 100% change in monthly count of events
	occurrence, by procedure_concept_id; data with unmapped concepts
	art month, by procedure_concept_id; 777 concepts have a 100% change in monthly count of events
	n, by condition_concept_id; 1623 concepts have a 100% change in monthly count of events
	id; data with unmapped concepts

Previous 1 Next

Lists data quality issues



Achilles





Open source tools

Orange: Under development

Black: Complete

- Support for ETL
 - **White Rabbit + Rabbit in a Hat** to help design an ETL
 - **Usagi** to help create code mappings
- Data exploration / study feasibility
 - **HERMES** to explore the vocabulary
 - **ACHILLES** to explore a database
 - **CIRCE** to create cohort definitions
 - **HERACLES** to explore cohort characteristics
- Methods for performing studies
 - **Cohort method** for new user cohort studies using large-scale propensity scores
 - **IC Temporal pattern discovery**
 - **Self-Controlled Case Series**
 - **Korean signal detection algorithms**
- Knowledge base
 - **MedlineXmlToDatabase** for loading Medline into a local database
 - **LAERTES** for combining evidence from literature, labels, ontologies, etc.



Key point

OHDSI is a research community that develops open source tools that run on the Common Data Model.

OHDSI 是一个研究团体，主要致力于开发在公共数据模型上运行分析的开源工具



Concluding thoughts

- Wealth of observational data to generate wealth of evidence for health care
- ETL into Common Data Model can be large initial investment
- Common Data Model allows sharing of results, methods, software tools

Thank you !

谢谢！