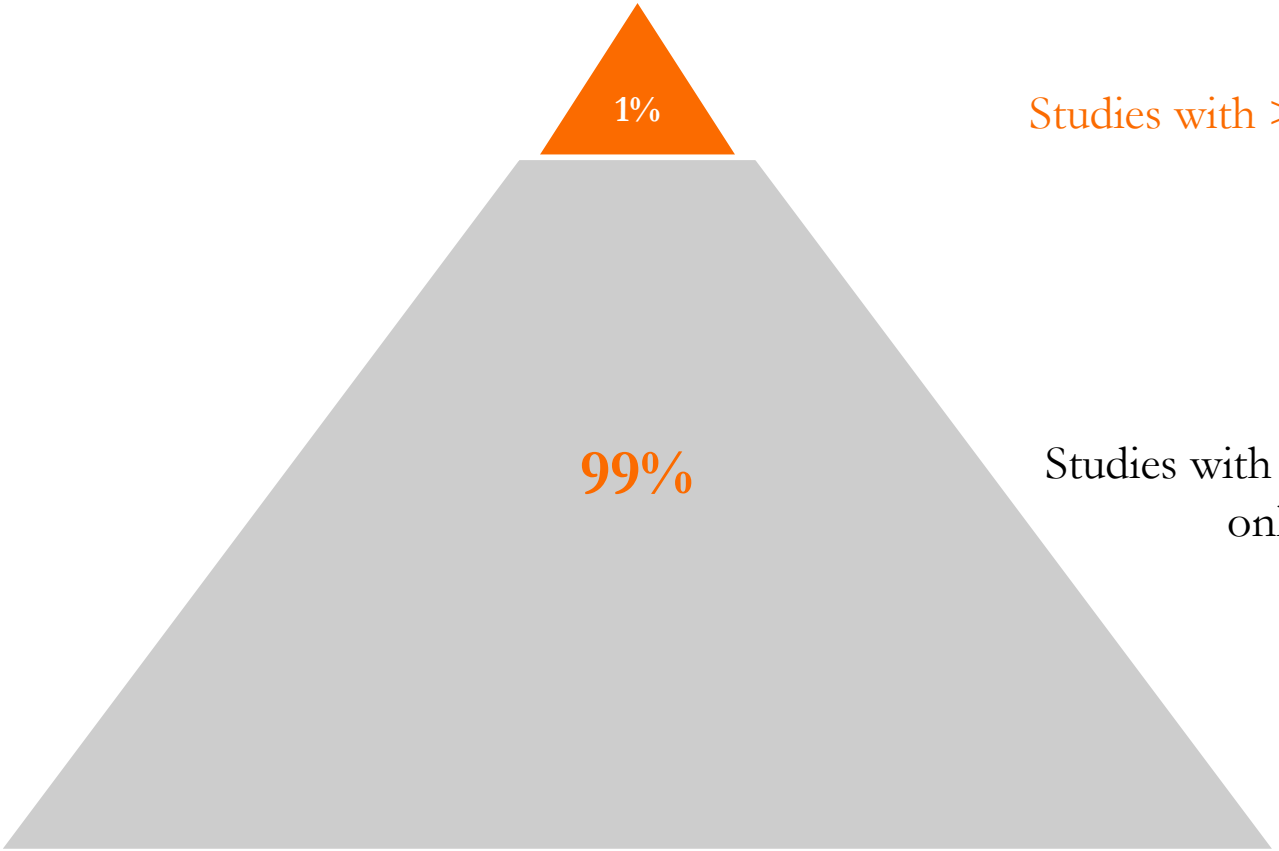
Concept Heterogeneity and Granularity in the OHDSI Network

Presenter:
Anna Ostropolets
PhD Student
DBMI, Columbia University

HOW THE CURRENT PUBLISHED RESEARCH IN CLINICAL INFORMATICS LOOKS LIKE

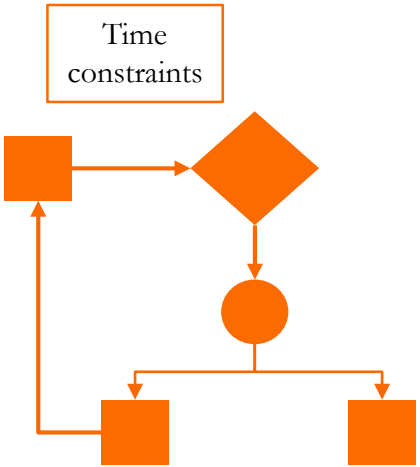


Studies with >1 database

Studies with 1 database only

SCENARIO 1: A PERFECT STUDY

You have an interdisciplinary team, you designed your study, created the cohorts, discussed them, validated all the codes…

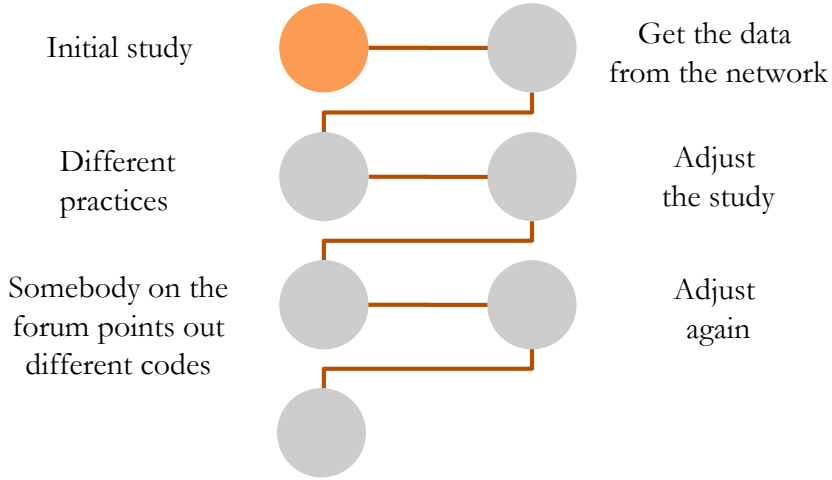


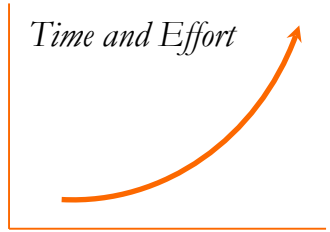
LOINC codes:
-
-

ICD10CM codes:
-
-
-
-
-

SCENARIO 1: A PERFECT STUDY

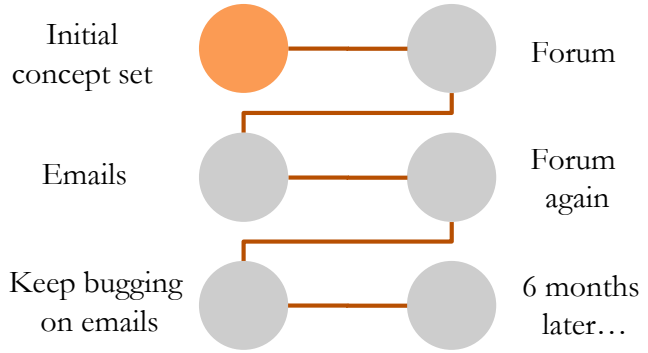
You wrote a protocol, start the study, but then your data partners told you that the events may be coded differently in their datasets

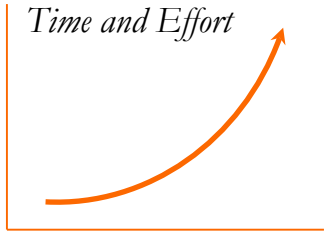




SCENARIO 2: A STUDY FOCUSED ON RARE EVENTS

You study a rare disorder or procedure and don't know which data partners have these events in their databases





OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

SCENARIO 3: STUDYING CONCEPT HETEROGENEITY

We have more than 150 databases from all over the world

IS GRANULARITY OF CODES DIFFERENT IN EHR AND CLAIMS DATA?

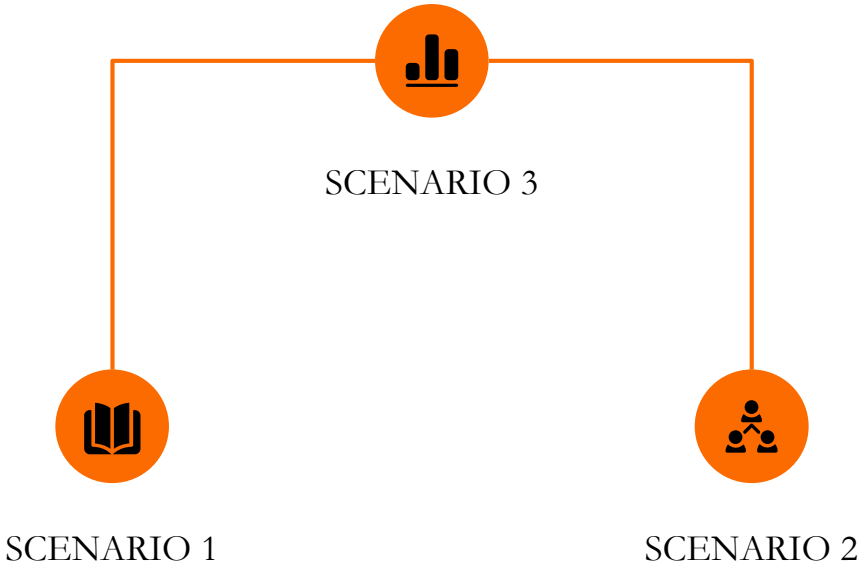
DOES GRANULARITY OR HETEROGENEITY DEPEND ON A COUNTRY?

DO DATABASES SHARE COMMON CONCEPTS?

ARE THERE DIFFERENT PATTERNS OF CODES UTILIZATION IN US AND EUROPE?

DOES GRANULARITY DEPEND ON A DISORDER/PROCEDURE/DRUG?

3 SCENARIOS: WHAT CAN WE DO TO MAKE THE NETWORK STUDIES EASIER AND TO LEAR MORE FROM OUR DATA?



CONCEPT PREVALENCE STUDY

CONCEPT PREVALENCE STUDY

METHODS: What do we collect?

Counts of records in the OMOP event tables and their domain

A snapshot of CONCEPT_RELATIONSHIP without custom (local) mappings

No patient-level information, no patient counts, all counts <100 are rounded up to 100

WE'VE ALREADY COLLECTED 19 DATABASES

1. Stanford Medicine Research Data Repository (StaRR)
2. Tufts Medical Center Repository (CLARET)
3. Columbia University Medical Center Database
4, 5, 6. IQVIA Hospital , Ambulatory EMR and Open Claims Databases
7. NHIS-Korean National Sample Cohort Database
8. Ajou University Database
9. The Healthcare Cost and Utilization Project (HCUP) Database
10, 11, 12. IBM CCAE, IBM MDCD and IBM MDCR
13. Japan Medical Data Center (JMDC) Database
14. MIMIC3 (Korea) Database
15, 16, 17. OPTUM EXTENDED DOD, EXTENDED SES and PANTHER
18. PREMIER Healthcare Database
19. Australian ePBRN Database

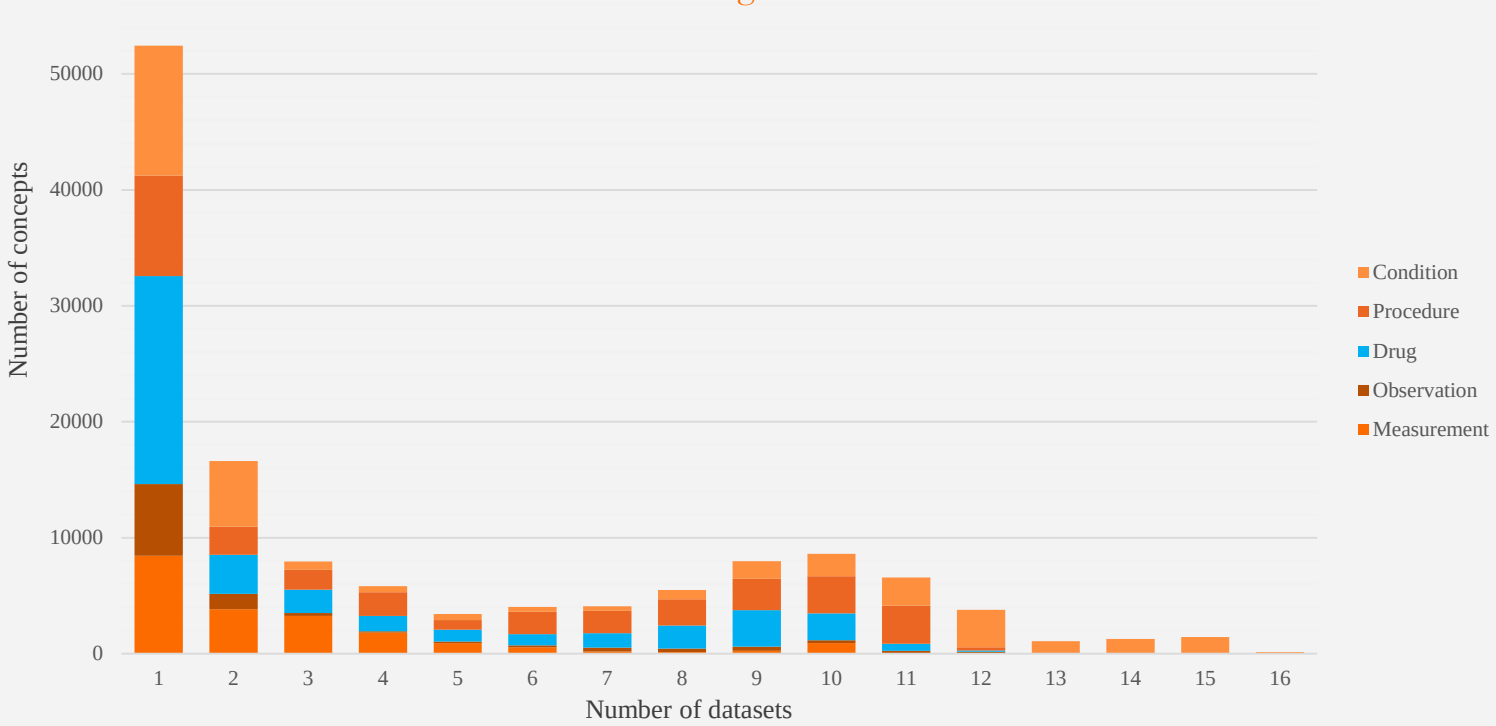
~271 billion
Records

12.5%
of all databases within the OHDSI network

257,385
Distinct concepts

MOST OF THE CONCEPTS CAN BE FOUND ONLY IN 1 DATASET

Condition is the least heterogeneous domain with the highest number of overlapping concepts across datasets, followed by Procedure and Drug domains. Measurement and Observation – highly heterogenous.



Child attention deficit disorder can only be found in 18% of datasets and has few patients. ADHD can be found in most of the databases and has many patients

HOW DO I GET INVOLVED?

I. Go to the GitHub

https://github.com/ohdsi-studies/ConceptPrevalence

II. Read Readme

III. Run the package

IV. Upload your results to our AWS bucket or send it to my email (encrypted)

V. Share your ideas and feedback

III. Run the package

You specify the connection details and the package does everything for you

I

Install 2 packages

```
install.packages("devtools")
devtools::install_github("https://github.com/ohdsi-studies/ConceptPrevalence")
```

Will also install SQL Render and Database Connector

II

Library the package

```
library('ConceptPrevalence')
```

III

Specify your connection details

dbms <- 'your_dbms' *("mysql"/"oracle"/"postgresql"/"redshift"/"sql server"/"pdw"/"netezza"/"bigquery")*
user <- 'user' *(your username)*
password <- 'password' *(your password)*
server <- Sys.getenv('server')
port <- Sys.getenv('port')
cdmName <- 'your_cdm_name' (e.g. Optum, CUMC etc.)
cdmDatabaseSchema <- "your_cdm_schema" *(the schema where event tables are stored)*
vocabDatabaseSchema <- "your_vocab_schema" *(the schema where vocabulary tables are stored)*
resultDatabaseSchema <-"your_results_schema" *(the schema with writing permissions)*

III. Run the package

You specify the connection details and the package does everything for you

IV

Establish the connection with your database

```
connectionDetails <- DatabaseConnector::createConnectionDetails(
dbms = dbms, server = server, user = user, password = password, port = port )
```

V

Run the package

```
ConceptPrevalence::calculate (
connectionDetails, cdmName, cdmDatabaseSchema, vocabDatabaseSchema, resultDatabaseSchema )
```

Output

5 csv files:

count_standard.csv
count_source.csv
mappings.csv
vocab_version.csv
cdm_info.csv

Upload your results to our AWS bucket (or email me)

You just send 5 tables via R, AWS bucket or email

DOCUMENTATION
Protocol and GitHub

I. GitHub

https://github.com/ohdsi-studies/ConceptPrevalence

GitHub contains
- R package itself, including SQL that extracts counts from the tables (inst/sql/sql_server)
- Protocol (extras)

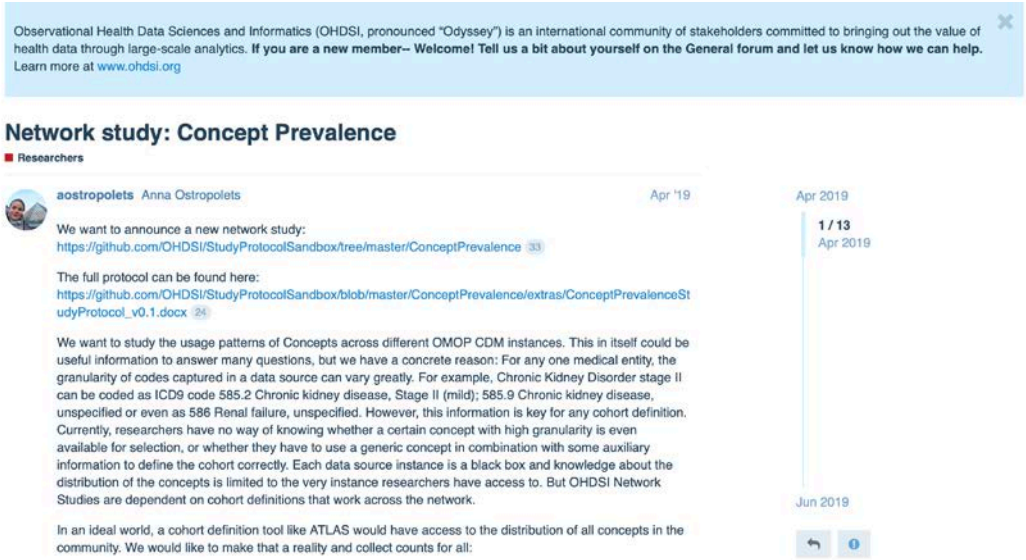
II. Protocol

https://github.com/ohdsi-studies/ConceptPrevalence/extras/

Protocol describes:
- Why this study matters
- What we are doing, including data analysis and data protection
- What we will do with the data

WHAT IF I HAVE QUESTIONS?

I Forum



https://forums.ohdsi.org/t/network-study-concept-prevalence/6562

II Just shoot me an email

ao2671@cumc.columbia.edu
aostropolets@gmail.com

THANKS!

Do you have any questions?