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

1% Studies with >1 database

99% 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…
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.

- Initial study
- Different practices
- Somebody on the forum points out different codes

Get the data from the network
Adjust the study
Adjust again

6 months later…

Time and Effort
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.

**Time and Effort**

- **Initial concept set**
- **Emails**
- **Keep bugging on emails**
- **Forum**
- **Forum again**
- 6 months later…
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 LEARN 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 heterogeneous.
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

Establish the connection with your database
connectionDetails <- DatabaseConnector::createConnectionDetails(
  dbms = dbms, server = server, user = user, password = password, port = port )

Run the package
ConceptPrevalence::calculate ( 
  connectionDetails, cdmName, cdmDatabaseSchema, vocabDatabaseSchema, resultDatabaseSchema )

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
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

II Just shoot me an email

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

Network study: Concept Prevalence

We want to announce a new network study:

The full protocol can be found here:

We want to study the usage patterns of Concepts across different OMOP CDM instances. This in itself could be useful information to answer many questions, but we have a concrete reason: For any one medical entity, the granularity of codes captured in a data source can vary greatly. For example, Chronic Kidney Disorder stage II can be coded as ICD10 code N58.2 Chronic kidney disease, Stage II (mild); N58.8 Chronic kidney disease, unspecified or even as N86 Renal failure, unspecified. However, this information is key for any cohort definition. Currently, researchers have no way of knowing whether a certain concept with high granularity is even available for selection, or whether they have to use a generic concept in combination with some auxiliary information to define the cohort correctly. Each data source instance is a black box and knowledge about the distribution of the concepts is limited to the very instance researchers have access to. But OHDSI Network Studies are dependent on cohort definitions that work across the network.

In an ideal world, a cohort definition tool like ATLAS would have access to the distribution of all concepts in the community. We would like to make that a reality and collect counts for all.

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

Do you have any questions?