What is Cohort Diagnostics?

Gowtham Rao

2021-07-20

Contents

1 Introduction

The CohortDiagnostics software enables iterative decision making by enabling the comparison of one or more cohort definition design choices for similar clinical ideas, to infer from the variations introduced by said choices on sensitivity, specificity, and consistency over a network of data sources. Cohort Diagnostics enables decisions such as the feasibility to develop cohort definitions for a clinical idea, improvement of definitions by comparing diagnostic performance to each other and a-priori expectations. Also, it may be used to generate characterization-based evidence.Cohort Diagnostics is an OHDSI HADES package and its use is considered a recommended best practice step prior to performing an OHDSI network study.

The CohortDiagnostics package allows one to generate a wide set of diagnostics to evaluate cohort definitions against a database in the Common Data Model (CDM). These diagnostics include incidence rates (optionally stratified by age, gender, and calendar year), cohort characteristics (comorbidities, drug use, etc.), and the codes found in the data triggering the various rules in the cohort definitions.

The CohortDiagnostics package in general works in two steps:

1. Generate the diagnostics against a database in the CDM.
2. Explore the generated diagnostics in a Shiny application included in the CohortDiagnostics package.

2 Diagnostics

The Diagnostics in Cohort Diagnostics may be grouped into

1. Data source level diagnostics
2. Concept set level diagnostics
3. Cohort level data diagnostics

2.1 Data source level diagnostics

There are currently two data source level diagnostics in Cohort Diagnostics. These are the ‘concepts in the data source’ and ‘Data source information’. The data source level diagnostics are not constrained to the

people who meet the cohort definition - and so may be used to infer about the data source heterogeneity. If data source heterogeneity is observed, then researchers should attempt to explain the observed heterogeneity by understanding the source data and make determination if that understanding might introduce limitations on the interpretation of the results of the proposed study.

Data Source information: answers questions such as - What is the version of the OMOP vocabulary used in the data sources? How many subjects are in the data source (irrespective of the cohort definition)? What is the distribution of the observation period in the data source? This information allows a researcher to determine the suitability of a data source for the research question.

Concepts in data source: shows the counts of concept id(s) that are part of the cohort definitions being evaluated. If one data source appears to have a different distribution of the concept counts compared to other data source(s), it is possibly a representation of differences in coding practices.

2.2 Concept set level diagnostics

Orphan concepts: are there any concepts that appear to have substring similarity with the concepts in the cohorts being diagnosed but are not present in the cohort definition, seem to capture the clinical idea behind the cohort definition and have sufficient counts in one or more data sources. If yes, was this code missed in the original cohort definition i.e., should the cohort definition include these concepts?

2.3 Data diagnostics

These are main diagnostics in cohort diagnostics and help us to infer about the cohorts by reviewing and comparing descriptive characteristics of the instantiated cohorts across and within data sources. Inferences may be drawn at the data source level, cohort definition level and concept-id level. - Cohort count and inclusion rules: Are the instantiated cohorts of sufficient size, what is the impact of the design choices on relative cohort counts, if inclusion rules are used in the cohort definition - are there rules that seem to dramatically change the counts, if yes, is that acceptable from generalizability standpoint i.e. are we identifying a specialized sub-population?

- Time distribution helps determine if the data source has sufficient time for people in the cohort either before, during or after cohort period.
- Index event breakdown helps to understand what concept in cohort concept set expression were present for the people in the data source on index date, i.e. concepts that triggered cohort entry. If we observe that a few concepts predominate in the cohort definition we may conclude that the other concepts in the definition are not meaningfully triggering cohort entry. In some instances there may be database heterogeneity in such predominance that may point to limitation of the generalizability of findings. We may also want to compare the counts of concepts in index event breakdown to the corresponding counts for the same concept in 'Concepts in data source' - to infer the impact of prevalence of concept in data source on determining entry into the cohort.
- Incidence rate describes the rate by which persons first enter the cohort. This diagnostics allows us to infer if there are unexpected variations in the rate over calendar periods, e.g. are there years that appear to be outlier with unusual dip/spike in incidence rate. If yes, is that explainable and would that introduce a bias/limitation to the study results?
- Characterization diagnostics include cohort characterization, compare cohort characterization, temporal characterization and compare temporal characterization. While cohort characterization and temporal characterization allows us to understand the characteristics of one cohort definition with ability to compare across data sources; compare cohort characterization and compare temporal characterization allows us to compare across cohort definitions within the same data source. The characterization output, derived from OHDSI/FeatureExtraction, presents both baseline and temporal characteristics (before and after index date) for cohorts. While characteristics of a cohort allow us to infer if a cohort definition is more likely capturing the subjects we want to study, differences in characteristics across the comparators highlights heterogeneity that may need to be explained as potential limitations or sources of biases.

3 Features

- Show cohort inclusion rule attrition.
- List all source codes used when running a cohort definition on a specific database.
- Find orphan codes, (source) codes that should be, but are not included in a particular concept set.
- Compute cohort incidence across calendar years, age, and gender.
- Break down index events into the specific concepts that triggered them.
- Compute overlap between two cohorts.
- Characterize cohorts, and compare these characterizations. Perform cohort comparison and temporal comparisons.
- Explore patient profiles of a random sample of subjects in a cohort.