

CohortPrevalence: OHDSI R Package for Estimating Prevalence

Martin Lavalley¹, Asieh Golozar², and Klaus Bondt²

¹Boehringer Ingelheim, ²Nemesis Health

Background

Prevalence is a key epidemiological measure used to understand the burden of disease in a population. In the pharmaceutical industry, prevalence is a key measure needed to support trial feasibility, regulatory and payer planning and market forecasting. Prevalence is defined as the number of affected person present in the population at a specific time divided by the number of person in the population at that time [1]. While a simple proportion to calculate in practice, the interpretation of how to count the numerator and denominator is variable from study to study [2]. Ascertaining prevalence from claims and electronic health records presents challenges: due to incomplete coding and biased representation of population. An optimal estimate of prevalence relies on three aspects: one, identification of cases (numerator); two, identification of the prevalence pool (denominator); and three, generalizability to the population (standardization). Given these issues with prevalence estimates, we propose a standardized framework for prevalence estimation using the OHDSI R package CohortPrevalence. The CohortPrevalence package provides researchers with pre-specified parameters for calculating based on the framework proposed by Rassen et al [3].

Methods

Prevalence is a simple proportion that calculates the number of persons with a disease at a specified point in time. It is naturally a cross-sectional method, originating from survey methodology, thus its use in longitudinal observational databases requires clear communication of enumeration. Below we present the methodology for calculating chronic conditions in longitudinal observational databases. This yields the crude prevalence estimate. To generalize to the target population we next perform direct method standardization by age-sex population weights.

Key Terms

- **Observation Period:** the span of time when a person's healthcare activity is observable in the database.
- **Period of Interest:** the specific time for which prevalence is anchored for its enumeration. This is often one year but can span multiple years.

- **Lookback Window:** A span of time prior to the period of interest used to assess existing evidence of chronic disease. In longitudinal observational databases, we are unable to “dip” into the data at a single point in time to determine whether a chronic condition is present. Instead, we define a period where we surveil for existing disease.
- **Prevalence Pool:** The denominator population is the source of which prevalent cases (numerator) are drawn. Conversely, if a patient cannot contribute to the numerator, it is excluded from the denominator.
- **Lead-in period:** the minimum duration of continuous observation required before a patient can be considered at risk and included in the denominator. LiP ensures that every patient had sufficient time for an existing condition to be documented before the prevalence estimation point.

Defining the numerator

Patients with the condition of interest are identified through a numerator cohort. Each record in the numerator cohort represents a patient’s period of cohort membership and contains a cohort start date, corresponding to the earliest recorded event satisfying the cohort definition, and a cohort end date, corresponding to the date when the condition is no longer considered present. For most chronic conditions, this is set to the end of the patient’s observation period in the database. However, in scenarios where chronic conditions can be resolved such as heart failure being resolved through heart transplantation, the cohort end date corresponds to the date when the condition is resolved. For conditions from which patients can recover, the cohort end date corresponds to the date the patient recovers or dies. The validity of all prevalence estimates is directly dependent on the assumption made towards the numerator cohort definition; including index date, end date and exclusion criteria.

In addition to the numerator cohort (case definition) we also need to define our means of counting cases relative to the period of interest and the lookback window. Per Rassen, there are two types of numerators:

- Option 1: The number of patients who have been observed to have the condition of interest on the first day of the period of interest or within the lookback time
- Option 2: The number of patients who have been observed to have the condition of interest at any time in the period of interest or within the lookback time.

While the above presents the formal approach to prevalence calculation, we also present a “rough” calculation which is common alternative performed by analysts in the pharmaceutical industry. We alternatively define the “rough” calculation as counting cases based on whether the cohort end date falls into an enumeration window relative to the period of interest and lookback period. Part of our aim is to present several parameterization of estimation in order to exhaustively define the range of viable

estimates. Given our assumption that a “true” prevalence is not attainable our best direction is to present all tweaks to the calculation.

Defining the denominator

The denominator (i.e. population at risk) is the part of the population who can be identified with the condition of interest during the period of interest. There are four options of denominators for prevalence:

- Option 1: “day 1” population - the number of persons in the population who were observed on the first day of the period of interest
- Option 2: “complete-period” population - the number of persons in the population who contribute all observable person-days in the period of interest. This is the strictest denominator
- Option 3: “any-time” population - the number of persons who who contributes at least 1 day in the period of interest. This is the most naive denominator and what IHD uses.
- Option 4: “sufficient-time” population - the number of persons who contributes *sufficient* time in the period of interest based on at least n observable person-days in the period of interest

This can be further partitioned by demographic constraints such as age and sex. Note that numerator 1 can only be used with denominator 1, while numerator can only be used with denominators 2-4.

Generalization to Population

Direct method standardization depends on the population reference utilized in the analysis. For US estimates, it is most common to standardize using the 2020 US census.

Results

Since we are presenting an R package we do not have specific results. Instead the basic package API is presented.

```
# set execution settings object from picard
executionSettings <- picard::createExecutionSettingsFromConfig(configBlock =
"omop_data")

#define output folder
outputFolder <- here::here("prevalence_results")

# Define analysis dimensions
exp <- CohortPrevalenceExperiment$new(
  name = "CKD Multi-Cohort Study",
  description = "3 cohort definitions x 3 prevalence types x 2 demographics"
)
```

```

# Add cohort reference. This assumes cohorts 1,2, and 3 were enumerated in a
cohort Table using CohortGenerator or analog
exp$addCohorts(tibble::tibble(
  cohortId = c(1, 2, 3),
  cohortName = c("CKD A", "CKD B", "CKD C")
))

# add the type of prevalence for calculation
exp$addPrevalenceTypes(list(
  createPrevalenceType("point_prevalence", lookBackDays = 0L, mode =
"formal"),
  createPrevalenceType("period_prevalence_pd2", lookBackDays = 365L, mode =
"formal"),
  createPrevalenceType("period_prevalence_pd3", lookBackDays = 0L, mode =
"rough")
))
# add demographic constraints of the population
exp$addDemographicConstraints(list(
  createDemographicConstraints(ageMin = 18, ageMax = 150),
  createDemographicConstraints(ageMin = 0, ageMax = 150)
))

# define the period of interest
exp$addPeriodsOfInterest(list(
  createYearlyRange(2015:2024)
))

# Execute all analyses
results <- generatePrevalence(
  prevalenceAnalysisList = exp$define(),
  executionSettings = settings
)

# standardize
results$standardizePrevalence(
  referencePopulation = usa_census_2020
)

# export results
results$export(outputFolder = outputFolder)

```

Conclusion

We propose this new open-source R package CohortPrevalence as a standardize framework for estimating prevalence in databases mapped to the OMOP CDM. The package follows the methods outlined above to estimate prevalence of chronic disease in longitudinal observational databases and provides alternatives common in the

pharmaceutical industry. Given that a “true” prevalence is impossible to ascertain, the best course is to perform multiple estimate based on justifiable assumptions relative to the disease in order to approximate the true burden of disease.

References

1. Celentano D, Szklo M. *Gordis Epidemiology*. 6th ed. Elsevier, Inc.; 2019.
2. Spronk, I., Korevaar, J.C., Poos, R., Davids, R., Hilderink, H., Schellevis, F.G., Verheij, R.A. and Nielen, M.M.J. (2019) 'Calculating incidence rates and prevalence proportions: not as simple as it seems', *BMC Public Health*, 19(1). doi: 10.1186/s12889-019-6820-3
3. Rassen JA, Bartels DB, Schneeweiss S, Patrick AR, Murk W. Measuring prevalence and incidence of chronic conditions in claims and electronic health record databases. *Clin Epidemiol*. 2019;11:1-15. doi:10.2147/CLEP.S164218