Prevalence Methods

Review of Operational Dfinition for Prevalence

Table of contents

1 Introduction

In this document we summarize the operational definitions of prevalence for longitudinal observational databases. Most assumptions of the operational definition for prevalence come from Rassen et al.

2 Rassen Method

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 extra effort to correctly enumerate.

2.0.1 Key Terms

- Observation Period: the span of time during which a person is considered actively engaged with the healthcare system and their medical activity is observable in the database. In claims, this is proxied by enrollment logic. In EHR data, it is typically defined as either: all observable time after birth or first observed event to last available date (common ETL assumption).

- 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 time: A defined span of time prior to the period of interest during which the database is queried for existing evidence of 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. If the chronic disease occurs during the lookback time, then it is considered to have prevalent 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.

2.1 Establishing the Eligible Population

The first step is to determine the "eligible" patients in the prevalence pool. These are the set of persons who will be evaluated for presence of the disease during the period of interest. Eligibility is determined using a couple of parameters:

1) Minimum Observation Period Length: This is the required time that persons in the database must have been observed to be eligible. This can be any number in days; typical options would be 1 day or 365 days.

2) First or any observation period: Determine whether to use first observation period or any observation period to evaluate the prevalence of a disease during the period of interest. In claims data, it is possible for patients to leave the database and return.

2.2 Defining the numerator

Once the eligible population is defined, the next step is to determine how to count the presence of the condition of interest. Note the existence of the condition is based on establishing a period of interest and a lookback period. There are two types of lookback periods: any time before or a fixed lookback based on a number of days relative to the anchor of the period of interest.

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.

For example, let's assume the period of interest is year 2022. For numerator option 1, we would count an event in the numerator if the condition was observed on either Jan 1, 2022 or sometime before based on the lookback option. If the lookback option was any time, a diagnosis in 2016 of a chronic event would be considered prevalent in the period of interest. If the lookback option was a fixed time of 365 days, then the diagnosis from 2016 would not be counted because it is beyond the fixed lookback time.

2.3 Defining the denominator

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

Common arrangements of the prevalence calculation:

- Option 1 numerator and "day 1" denominator, commonly known as point prevalence
- Option 1 numerator and "complete-period" denominator commonly known as complete point prevalence
- Option 2 numerator and "complete-period" denominator commonly known as complete period prevalence
- Option 1 numerator and "any-time" denominator; typically used in IHD and naive tools

3 Process for Prevalence

Here we define the process of defining a prevalence estimate.

1) Define the condition of interest: Construct an all-by-all condition cohort in ATLAS
2) Define the period of interest (one year or multi-year)

3) Define the "eligible" observed population

- determine the minimum observation period length
- determine observation period choice (first or all)
- partition the population by demographic requirements (age, sex, race)

4) Define the denominator:

- select denominator option ("day 1", "complete-period", "any-time", "sufficient-time")

5) Define the numerator

- select numerator option (events in lookback time or in both lookback time and period of interest)
- define lookback time (all or fixed)