Name:	Martijn Schuemie
Affiliation:	Janssen R&D
Email:	schuemie@ohdsi.org
Presentation type (s):	Poster

A benchmark for population-level estimation methods

Martijn J. Schuemie[1], M. Soledad Cepeda[1], Marc A. Suchard[2], Yuxi Tian[2], Alejandro Schuler[3], Patrick B. Ryan[1], George Hripcsak[4]

[1]Janssen R&D, Titusville, New Jersey, USA; [2]University of California, Los Angeles, California, USA; [3]Stanford, Stanford, California, USA; [3]New York-Presbyterian Hospital, New York, New York, USA

Abstract

Here we present a benchmark for evaluating methods used for effect estimation (effect of an exposure on the risk of an outcome) and comparative effect estimation (effect of an exposure relative to another exposure on the risk of an outcome). The benchmark consists of a gold standard of 200 real negative and 600 synthetic positive controls, and a set of metrics that can be computed after a method is applied to the controls. We believe this benchmark can help provide an understanding of how methods perform in general. This in turn can inform the choice of method for answering a particular research question.

Introduction

When designing an observational study, there are many study designs to choose from, and many additional choices to make, and it is often unclear how these choices will affect the accuracy of the results. (e.g. If I match on propensity scores, will that lead to more or less bias than when I stratify? What about power?) The literature contains many papers evaluating one design choice at a time, but often with unsatisfactory scientific rigor; typically, a method is evaluated on one or two exemplar study from which we cannot generalize, or by using simulations which have an unclear relationship with the real world.

Here we present a new benchmark for evaluating population-level estimation methods, one that can inform on how a particular study design and set of analysis choices perform in general. The benchmark consists of a gold standard of research hypothesis where the truth is known, and a set of metrics for characterizing a methods performance when applied to the gold standard. We distinguish between two types of tasks: (1) estimation of the average effect of an exposure on an outcome relative to no exposure (*effect estimation*), and (2) estimation of the average effect of an exposure on an outcome relative to another exposure (*comparative effect estimation*). The benchmark allows evaluation of a method on either or both tasks.

This work builds on previous efforts in EU-ADR[1], OMOP[2], and the WHO[3], adding the ability to evaluate methods on both tasks, and using synthetic positive controls as real positve controls have been observed to be problematic in the past.

Gold standard construction

The gold standard consists of 800 entries, with each item specifying a target exposure, comparator exposure, outcome, nesting cohort, and true effect size. An example entry: target = Diclofenac, comparator = Celecoxib, outcome = Lyme disease, nesting cohort = Arthralgia, true effect size = 1. Each entry can be used for both tasks, since the true effect size holds both when comparing the target exposure to no exposure as well as when comparing the target exposure to the comparator exposure. The nesting cohort identifies a more homogeneous subgroup of a population, and can be used to evaluate methods such as the nested case-control design.

A set of 200 entries are negative controls, where the relative risk is believed to be 1. These negative controls were selected by first picking four outcomes and four exposures of interest. Using these as starting point, we generated candidate lists of negative controls using LAERTES[4], which draws on literature, product labels, and spontaneous reports. These candidates were used to construct target-comparator-outcome triplets where neither the target nor the exposure causes the outcome, and the target and comparator were either previously compared in a randomized trial per ClinicalTrials.gov, or both had the same 4-digit ATC code (same indication) but not the same 5-digit ATC code (different class). These candidates were ranked on prevalence of the exposures and outcome and manually reviewed until 25 were approved per initial outcome or exposure. Nesting cohorts were selected by manually reviewing the most prevalent conditions and procedures on the first day of the target or comparator treatment.

The remaining 600 entries are positive controls, which were automatically derived from the 200 negative controls by adding synthetic additional outcomes during the target exposure until a desired incidence rate ratio was achieved between before and after injection of the synthetic outcomes. The target incidence rate ratios were 1.25, 2, and 4. To preserve (measured) confounding, predictive models were fitted for each outcome during target exposure and used to generate probabilities from which the synthetic outcomes were sampled.

Metrics

Once a practical method has been used to produce estimates for the gold standard the following metrics are computed:

- Area under the received operating curve, when comparing positive controls to negative controls
- Mean squared error
- Bias distribution
- Coverage of the confidence interval
- Type I and type II error

These metrics are computed both overall, as well as stratified by true effect size and by each of the 4 initial outcomes and 4 initial exposures.

Conclusions

The OHDSI Population-Level Estimation Benchmark allows characterization of performance of observational study designs and analytic choices, and can help provide an understanding of how methods perform in general. This in turn can inform the choice of method for answering a particular research question.

References

1. Schuemie MJ, Coloma PM, Straatman H, et al. Using electronic health care records for drug safety signal detection: a comparative evaluation of statistical methods. *Medical care*. 2012; 50: 890-7.

2. Ryan PB, Stang PE, Overhage JM, et al. A comparison of the empirical performance of methods for a risk identification system. *Drug safety*. 2013; 36 Suppl 1: S143-58.

3. Caster O, Juhlin K, Watson S and Noren GN. Improved statistical signal detection in pharmacovigilance by combining multiple strength-of-evidence aspects in vigiRank. *Drug safety*. 2014; 37: 617-28.

4. Voss EA, Boyce RD, Ryan PB, van der Lei J, Rijnbeek PR and Schuemie MJ. Accuracy of an automated knowledge base for identifying drug adverse reactions. *Journal of biomedical informatics*. 2017; 66: 72-81.