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[1,4], George Hripcsak[4,5]

[1]Janssen R&D, Titusville, NJ, USA; [2]University of California, Los Angeles, CA, USA; [3]Stanford, Stanford, CA, USA; [4]Columbia University, New York, NY, USA; [5]NewYork-Presbyterian Hospital, New York, NY, USA

Background

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?) 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 hypotheses 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. **effect estimation**: estimation of the average effect of an exposure on an outcome relative to no exposure.
2. **comparative effect estimation**: estimation of the average effect of an exposure on an outcome relative to another exposure.

The benchmark allows evaluation of a method on both tasks.

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

Limitations

Given the nature of the negative controls it is unlikely that any of the exposures will be contra-indicated for the related outcome of interest, precluding the ability to evaluate a method's sensitivity to contra-indication.

The process for adding synthetic outcomes can only preserve measured confounding, so performance on positive controls with respect to unmeasured confounding may be slightly optimistic.

Availability

The benchmark is available in the MethodEvaluation R package: https://github.com/OHDSI/MethodEvaluation

This includes:

- Negative control set
- Function for creating outcome and nesting cohorts
- Function for synthesizing positive controls
- Functions for computing MDRR and metrics

Real negative controls (n = 200)

- Pick 4 outcomes and 4 exposures of interest

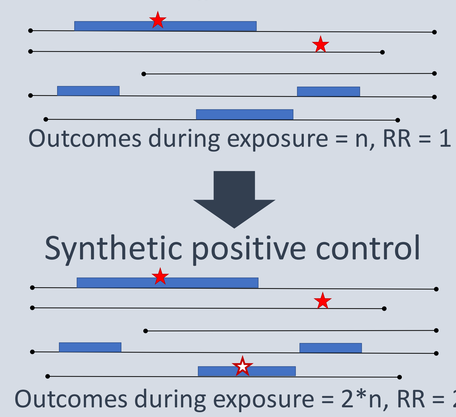
Acute pancreatitis	Diclofenac
GI bleeding	Ciprofloxacin
Stroke	Metformin
IBD	Sertraline

- Use LAERTES to identify potential negative controls
- Use clinicaltrials.gov + ATC to find potential comparator exposures
- Rank by prevalence
- Manual review, up to 25 per outcome or exposure of interest

Synthesized positive controls (n = 600)

- Based on real negative controls (where true RR = 1)
- Fit predictive models for each outcome in exposed population
- Sample simulated additional outcomes during exposure based on predicted probability until True RR = desired RR (1.5, 2, and 4)

Synthesis process



Gold standard (n = 800)

Target	Comparator	Nesting	Outcome	True effect size
Eszopiclone	Triazolam	Insomnia	Acute pancreatitis	1
Eszopiclone	Triazolam	Insomnia	Acute pancreatitis, RR=1.5	1.5
Eszopiclone	Triazolam	Insomnia	Acute pancreatitis, RR=2	2
Eszopiclone	Triazolam	Insomnia	Acute pancreatitis, RR=4	4
Ciprofloxacin	Azithromycin	Otitis media	Alcohol abuse	1
Ciprofloxacin	Azithromycin	Otitis media	Alcohol abuse, RR=1.5	1.5
Ciprofloxacin	Azithromycin	Otitis media	Alcohol abuse, RR=2	2

Compute power in database

- Compute minimum detectable relative risk (MDRR)
- Filter controls (e.g. MDRR < 1.25)

Evaluate *effect estimation* method

- Compute effect of exposure to *Target* on risk of *Outcome*
- Optionally nest in *Nesting* cohort (e.g. Nested Case-Control)
- Compare to true effect size

Evaluate *comparative effect estimation* method

- Compute effect of *Target* compared to *Comparator* on risk of *Outcome*
- Optionally nest in *Nesting* cohort
- Compare to true effect size

Compute performance metrics

- **AUC**: Area under the ROC curve for classifying positive controls vs. negative controls
- **Coverage**: Coverage of the 95% confidence interval
- **Mean precision**: Precision = 1/SE2; higher precision means narrower confidence intervals
- **MSE**: Mean squared error between effect size (point) estimate and the true effect size
- **Type 1 error**: For negative controls, how often was the null rejected (at alpha = 0.05)
- **Type 2 error**: For positive controls, how often was the null not rejected (at alpha = 0.05)
- **Missing**: For how many of the controls was the method unable to produce an estimate

Method(s) to evaluate

Case-control

- Nested
- 10 controls per case

Database (CDM)