A Clinical Trial Derived Reference Set for Evaluating Observational Study Methods

Ethan Steinberg[1], MCS, Steve Yadlowsky[2], PhD, and Nigam H. Shah[1], PhD

[1] Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA
[2] Electrical Engineering, Stanford University, Stanford, CA, USA

Background

- Observational studies are an essential tool for estimating treatment effects, but they rely on assumptions that cannot be tested with observational data alone.
- Reference sets are one technique for evaluating observational study methods and determining whether or not those assumptions hold in practice. The core idea behind a reference set is to construct a set of known truths and then run corresponding observational studies that in theory should provide similar results.
- Previous OHDSI affiliated researchers have constructed a series of reference sets based on FDA labels and other sources.
  - OMOP, EU-ADR, OHDSI are the most prominent examples.
- In this work, we introduce a new clinical trial derived reference set with much stronger empirical and statistical backing and show an example use case where we evaluate a suite of observational methods.

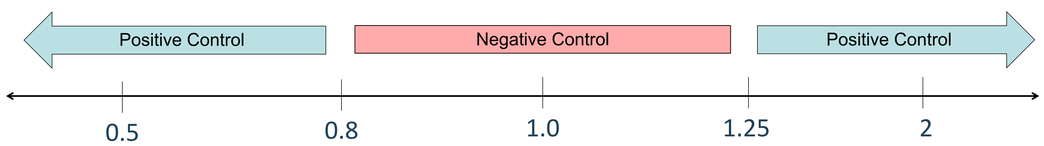
Methods

Definition Of Controls

Controls were defined as drug vs drug comparisons with respect to a particular outcome. Having active comparators is crucial for constructing corresponding observational studies.

Classifying Controls as Positive Or Negative

Controls were classified as positive or negative based on the magnitude of the odds ratio. Effects >1.25 or <0.8 were considered positive controls. Weaker effects were considered to be negative controls. Defining controls in terms of ranges is important as it is not possible to prove that an odds ratio is exactly 1.



Classifying Controls From Contingency Tables

Given our definition of positive and negative controls and a contingency table, we then classified particular controls as either positive, negative or undetermined. We ran Fisher exact tests to obtain separate P-values for a control being positive or negative. We corrected for multiple comparisons using the Benjamini-Hochberg procedure at $\alpha = 0.05$. Controls that failed both significance tests were discarded.

Example Controls

Drug A	Drug B	Outcome	Direction
Linagliptin	Glimepiride	Nasopharyngitis	=
Letrozole	Anastrozole	Arthralgia	=
Vemurafenib	Dacarbazine	Alopecia	<
Empa	Glimepiride	Hypoglycemia	<

Contact: ethanid@stanford.edu

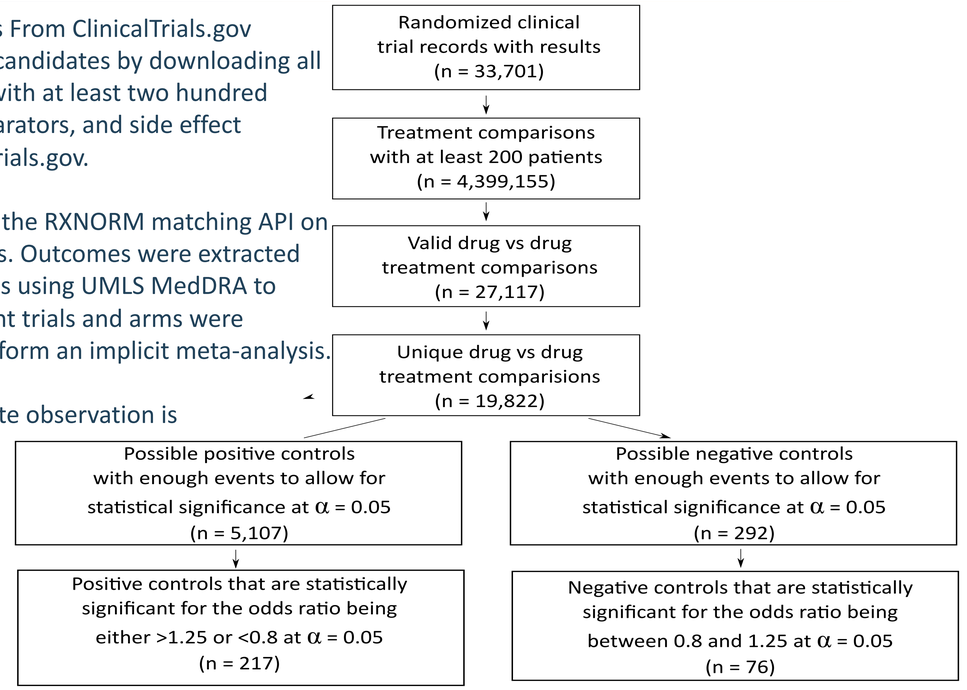
Methods (continued)

Extracting Controls From ClinicalTrials.gov

We extracted raw control candidates by downloading all randomized clinical trials with at least two hundred patients, two active comparators, and side effect information from ClinicalTrials.gov.

Drugs were mapped using the RXNORM matching API on the names of the trial arms. Outcomes were extracted from the side effect reports using UMLS MedDRA to ICD10 mappings. Equivalent trials and arms were aggregated in order to perform an implicit meta-analysis.

One important intermediate observation is that very few clinical trials had sufficient data to measure negative effects.



Our code is available at:
https://github.com/som-shahlab/clinical-public

Example Application

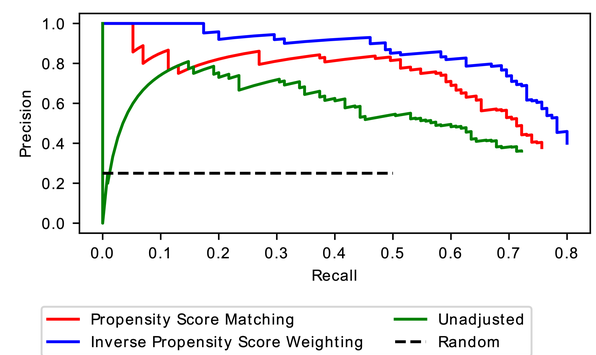
As an example use case, we ran an analysis comparing common observational study techniques on the Optum Clinformatics Data Mart dataset. We compared propensity score matching, inverse propensity score weighting, and unadjusted analyses.

We defined performance in terms of precision and recall of being able to recover effects at various hazard ratio thresholds.

Recall = Fraction of correctly recovered controls
Precision = Fraction of results which were correct

Sweeping all possible hazard ratio thresholds provides a precision recall plot.

Precision And Recall Of Various Methods On Optum



Conclusions

- Our new control set provides a promising alternative for evaluating observational study methods.
- Propensity score based adjustment methods perform significantly better than unadjusted analyses.
- Overall, observational studies appear to have reasonably high precision, but low recall.