An Evaluation Methodology to Guide Model Selection and Cohort Definition in Causal Inference

Yishai Shimoni[1], Ehud Karavani[1], Sivan Ravid[1], Peter Bak[1], Tan Hung Ng[2], Sharon Hensley Alford[3], Denise Meade[3] and Yaara Goldschmidt[1]

[1] IBM Research - Haifa, Israel [2] IBM Watson Health, San Jose, CA [3] IBM Watson Health, Cambridge, MA

Background.

Causal inference analysis leverages observational data to estimate the effect of an intervention on some outcome when randomized controlled trials are not available. It is, therefore, an essential tool for using data to guide intervention policies. Many methods for causal inference employ propensity (or weight) models, attempting to balance the covariate distributions between treatment groups, while others use outcome models that predict individual counterfactual outcomes directly. Additional methods, known as doubly-robust methods, combine models from both classes for a more robust estimation. All these methods employ underlying machine learning (ML) models to perform the estimation, and thus require performance evaluation.

Furthermore, for a derived effect to be considered truly causal, the analysis must also adhere to three basic assumptions – consistency, exchangeability and positivity. Therefore, when conducting causal analyses, we must verify two things. First, that the cohort is properly defined and that the causal assumptions hold; and second, that the underlying ML models, which can be arbitrarily complex, are well specified and generalizable.

Methods.

We present a toolkit to apply various methods of causal inference analysis, where the underlying ML model can be defined independently. This is augmented by a suite of evaluation methods that, unlike conventional benchmarking methods, operate on the data at hand in a cross-validated manner (or using held-out data). They are inspired by known ML evaluations, but their interpretations are adjusted to the causal inference context. Many of the evaluations apply only to propensity, weight, or outcome models. The toolkit includes both well-established evaluation methods in addition to novel ones.

It should be noted that since counterfactuals are never observed, outcome models can never be fully evaluated against ground-truth. However, for a causal model to be truly correct, it must be at least correct on the observed outcomes. Therefore, evaluating the ML models to ensures that the results of the analysis are not "obviously wrong".

The code is available as an open-source python package on GitHub, and is also installable directly from PyPI.

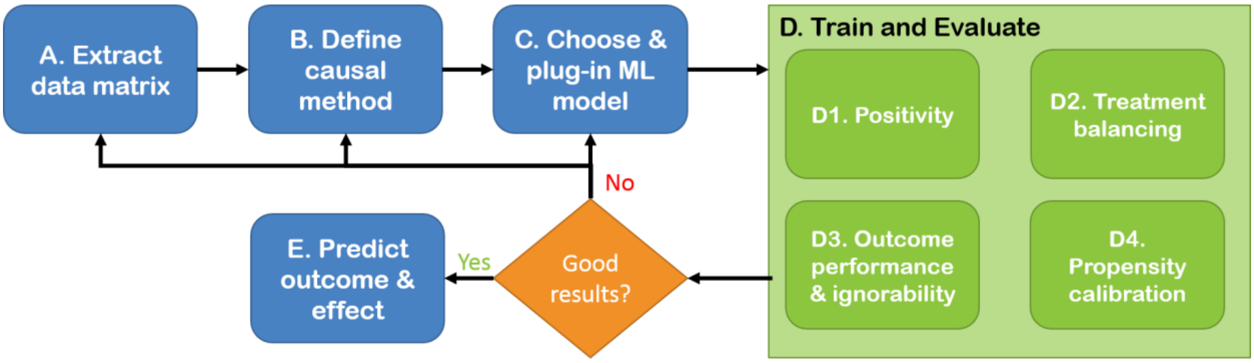
Results.

We present a novel interpretation for a weighted ROC curve to detect both positivity and overall balancing in a non-marginal approach; and include an expected ROC curve to evaluate the consistency of a propensity model. Additionally, we present a proxy for testing exchangeability by examining individual predictions of potential outcomes.

Conclusion. Evaluating in a cross-validation manner can hint at the source of violations. If poor evaluation performance is present only in the train set, it hints that the model might have overfitted, i.e., the core ML model is misspecified. Detecting misspecification can lead to model changes and thus the toolkit guide model selection. If bad performance is also present in the test set, it is likely to be caused by inherent structures in the data. Thus, we can detect when the data cannot support the causal question, and assume the causal assumptions are probably not met.

Additionally, cross validation can also help assess performance on sub-populations, which is essential for uses in precision medicine, by treating the sub-cohort as an out-of-sample set.

Finally, the ability to eliminate poor performing models and detect data misspecification allowed us to develop a workflow that can guide model selection and direct cohort definitions, as depicted in the figure below.


A. Extract data matrix
B. Define causal method
C. Choose & plug-in ML model
D. Train and Evaluate
D1. Positivity
D2. Treatment balancing
D3. Outcome performance & ignorability
D4. Propensity calibration
Good results?
No
Yes
E. Predict outcome & effect