Proposed Project title:

***“Evaluating current causal inference methods in data analytics, and demonstrating novel results in projects where they are added”***

Proposed research objectives

1. Design a series of metrics/ a framework for evaluating CIMs that prioritises interpretability, ease of use, and breadth of applicability

2.Compare and contrast different types of CIMs according to these metrics and compare the findings with other metrics for evaluating the same CIMs, in order to motivate choice of methods for RO3

3.Determine whether CIMs can be applied to otherwise completed test projects to provide novel insight

Literature review

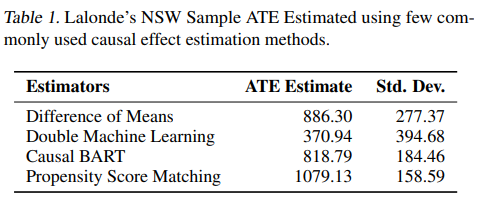
Structure of the Lit review

Justify the importance/significance/value of CIMS - “A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks Miguel A. Hernán, John Hsu, and Brian Healy”

“WHY ASK WHY? FORWARD CAUSAL INFERENCE AND REVERSE CAUSAL QUESTIONS Andrew Gelman Guido Imbens “

Look for some sources that cast doubt on their usefulness and add findings here

Introduce the background theory, especially area for evaluation methods like Average Treatment effect ATE. explain issues with differences between estimators like this graph from Parikh et al 2022:



Introduce the field as it stands with very brief description of how it’s evolved

Introduce the questions of the field as my topic relates to it- applicability & breadth of use

Slight explanation about theory and measures such as

Overview and review of various methods

Potential outcome framework - “A Survey on Causal Inference Yao 2021”

Y

Z

Overview of types of evaluations “face validity, placebo/negative control, synthetic data tests” Parikh et al 2022

These will be extra relevant to what I will be focussing on - “Validating Causal Inference Methods Harsh Parikh Carlos Vajao Louise Xu Eric Tchetgen Tchetgen 2022”

-”credence” created by the research group, which is a framework for creating synthetic tests

Review results of comparisons- “Automated versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition1 Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott and Dan Cervone 2019”

Review of real world testing and evaluation/ use cases / adoption - “Real-World Evidence in Medical Product Development Editors-Weili He Yixin Fang Hongwei Wang , Authors - Susan Gruber, Hana Lee, Rachael Phillips & Mark van der Laan 2023”

Conclusions about what methods/ types of methods are especially appropriate

Conclusions about what types of evaluation are comparable to what I am looking to do

Conclusions about how I’m going to use all this to answer my research questions

Proposed sampling strategy

Sampling for this research project will happen on two levels.

Appropriate CIMs to evaluate must be selected. The population in this case will be the set of all causal inference methods that have already been reviewed within the literature. This is a subset of the set of all causal inference methods in existence, but this subsetting is justified because tracking down all the methods would be extremely difficult, especially due to new ones constantly being created, and because unstudied CIMS have little guarantee of effectiveness. The existence of various types and subcategories of CIM leads itself to stratification of these groups. Within these strata, based on the literature review and the criteria of evaluation selected for the project, some methods will stand out, and will be chosen non-randomly as per the method of judgement or purposive sampling. Within other strata, there will not be any obvious choices, and random sampling within the strata will be used for those. The criteria for selection using judgement sampling here will be carefully justified and documented. While this selection will create sampling bias, this is the intention. It is more important with the selection of methods for the method to be appropriate than it is for it to be representative of the sample or subsample from where it was picked. For types of methods on which there is less information however, a random method(s) from within that group of methods will be selected. This will deal with cases where there isn’t enough information to determine the best method(s), and these randomly selected methods will represent the strata for the purposes of generalisation.

Separately, test projects must be selected for the addition of the CIMs. These can be sampled only from a population of projects whereby the addition of methods is possible, and has not already been done. For this, the projects themselves and the data used must be readily accessible, both must be licenced for use without explicit consent, and they must involve some form of concomitant variation.

Additionally another set of criteria will be added to select a population of projects that are appropriate. The projects must be at an appropriate level of complexity, they must not be in a particularly niche domain, they must not have significant barriers to understanding. They must not introduce ethical concerns that will incur significant additional ethics proposals. Other criteria will be added to this list as they are encountered during research.

For the population of projects where the addition of CIMS is possible and appropriate, A combination of convenience sampling and purposive/judgement sampling will be used, selecting test projects from convenient sources that are within the population and attempting to represent the sample by selecting projects that are maximally different and yet still typical in order to best represent the sample.

The convenience sampling will be used to justify narrowing down the population even further into projects that can be easily found as well as accessed, especially as there are still innumerable projects that fit the criteria.

For the purposive sampling, there will be a significant amount of bias involved towards selecting test projects for which the addition of causal inference methods is more likely to produce positive results. However this is both intentional and acceptable, as this project is not attempting to make a generalised statement about applying the CIMs to the entire population of appropriate projects to find novel results. Instead it will try to show that their addition can be used to find new results at all.

However it is still desirable to represent the breadth of the population, hence the decision to choose significantly different typical test projects.

Proposed primary research methodology

Experimental research has been chosen as the primary research methodology for this project.

Experiments will be performed to test the hypothesis “Does the addition of Causality Inference Methods to this test project produce novel results?” To test this hypothesis, several otherwise completed analytics test projects will be selected, one or more causal inference methods will be applied to each, and the results of the methods will be compared with the original results of the test project. Specific criteria for success, and a definition of ‘novel results’ will be essential for evaluating this hypothesis for each of them. For a control, the original test project presents itself as a convenient option. For these experiments, all four factors will be satisfied. There will be concomitant variation between the independent variable describing applied/ not applied and the dependent variable describing novel results found/not found, there will also be temporal sequence of the states prior to and after the methods being applied. All the research provided by the literature review will comprise the theoretical support. Finally, there will be complete control of the system where nothing will be introduced that could influence the dependent variable aside from the treatment.

Overall this research method satisfies the requirements of primary research, and will provide a robust basis to generate the desired results for the project

Ethical and risk considerations

While there are no expectations to encounter ethical considerations in the evaluation of methods, there are several that will be encountered when sourcing the test projects that get chosen for the addition of methods. For each of these, all ethical concerns will need to be addressed. As a result, it is desirable to avoid including any projects that involve special categories of personal data from higher risk groups, as this may entail additional ethics proposals and applications.

Assuming the test projects are readily available online, licensing will also be sourced and reviewed to ensure they allow for use without permission. Otherwise, explicit consent from the test project creators will be needed showing that they understand and agree to everything their project will be used for.

An inherent consideration to be addressed will be the secondary use of data. Details for all sources will be included. Data being open sourced will be a prerequisite for use in this research project, but the ethical considerations of the use of the data must still be considered, especially in the case where personal data will be used. There are legal regulations for dealing with personal data as laid out in the GDPR, as well as many countries having national data protection laws. It may not be possible to get the explicit informed consent of data subjects due to the nature of the project, and so one step that may need to be taken is the complete irreversible anonymisation of all personal data so as to relax requirements that might make their use difficult. Otherwise, details must be found and provided on the initial collection of the data, and protection of the rights of the data subject must be ensured. Data minimisation will be undertaken during both collection and processing where appropriate so as to avoid using unnecessary personal information.

In the case personal data is used at all, ensuring data security will be important, and measures will need to be taken that are proportional to the risks faced by the subjects in the case of breach or destruction of the data in compliance with GDPR. For this project, any personal data collected will be stored locally only on a password protected personal laptop, and any software or applications used for the processing of data will be checked to identify risks to the data subjects.

Depending on the nature of the test projects selected, it may be of concern that the findings produced by this project have the propensity to mislead, or be misused. It will be essential to address any ethical questions related to this prior to the release of results.