PhD thesis: Causal inference for combining experimental and observational data

Julie Josse (Inria Montpellier, Premedical team)

April 2023

1 Context

Modern *evidence-based* medicine puts Randomized Controlled Trial (RCT) at the core of clinical evidence. In practice, almost all new drugs are authorized through such trials (after a pre-clinical study). Indeed, *randomization* enables to estimate the Average Treatment Effect by avoiding confounding effects of spurious or undesirable associated factors. In other words, RCTs are the current gold-standard to empirically measure a causal effect of a given intervention on an outcome. But more recently, concerns have been raised on the limited scope of RCTs: stringent eligibility criteria, unrealistic real-world compliance, short timeframe, limited sample size, etc. Such limitations threaten the external validity of RCT studies to other situations or populations [17]. The usage of complementary non-randomized data, referred to as *observational* or from *real world*, brings promises as additional sources of evidence, in particular combined to trials. **Transportability** (also known as **generalization**, *recoverability from sampling bias*, or *data-fusion* [20, 16]) allows to generalize or transport the trial findings toward a target population of interest, potentially subject to a covariate **distributional shift**.

As a recent extreme example, the Food and Drug Administration (FDA) has greenlighted the use of palbociclib to men with breast cancer, though clinical trials were performed only on women. Authorizing such extensions would help **reducing the time to approve a drug for patients who could benefit from it**. Hence, the societal impact of these methods to improve patient care but also to reduce costs (in France, the price of drugs depends among other things on their effectiveness) is huge. Yet, they raise some concerns, especially when the target population is very different from the RCT cohort. In addition, such methods are still in a prototype stage, with a wide **gap between theory and practice** and the confrontation with data give rise to many methodological challenges. Theoretical, methodological and applied developments and validations are needed to better understand and leverage **the speed-safety balance** and to design future RCTs.

2 Generalization of different causal measures

We use the *potential outcome* framework to characterize treatment (or causal) effects. This framework has been proposed by Neyman in 1923 [English translation in 19], and popularized by Donald Rubin in the 70's [8, 7]. It formalizes the concept of an intervention by studying two possible values $Y_i^{(1)}$ and $Y_i^{(0)}$ for the outcome of interest (say the pain level of headache) for the two different situations where the individual i has been exposed to the treatment ($A_i = 1$) or not ($A_i = 0$) –we will only consider binary exposure. The treatment has a causal effect if the potential outcomes are different, that is testing the assumption:

$$\mathbb{E}\left[Y^{(1)}\right] \stackrel{?}{=} \mathbb{E}\left[Y^{(0)}\right], \tag{1}$$

where $\mathbb{E}[Y^{(a)}]$ is the expected counterfactual outcome had all individuals in the *population* received the treatment level a. A common measure to test this assumption is the absolute difference (usually referred to as the Risk Difference - RD):

$$\tau_{\text{RD}} := \mathbb{E}\left[Y^{(1)}\right] - \mathbb{E}\left[Y^{(0)}\right].$$

This quantity depends on the population with respect to which the expectation is taken. More precisely, for a given set of covariates X, one can write

$$\tau_{\text{RD}} := \mathbb{E}\left[\mathbb{E}\left[Y^{(1)} - Y^{(0)}|X\right]\right],$$

which depends on the distribution of X, that is the considered population on which the conditional treatment effect $\mathbb{E}\left[Y^{(1)} - Y^{(0)}|X\right]$ is averaged.

Several estimators have been proposed to estimate this quantity in a target population, based on information resulting from an RCT. These generalization estimators use *weighting* (Inverse Propensity of Sampling Weighting, IPSW), *outcome modeling*, or combine the two in *doubly robust approaches* with Augmented IPSW (AIPSW) [see 4, 1, for a survey, and consistency proofs]. Depending on the output type (typically binary or continuous), other causal measures than Risk Difference, such as Risk Ratio, may be more appropriate [3]. However, there is little theory to adapt the previous classes of estimators to other metrics such as the Risk Ratio, and there is no theoretical guarantee to explain their empirical performance.

First axis of the PhD: Extend the different classes of estimators to classical causal measures (as the Risk Ratio) different from the Risk Difference and derive theoretical guarantees as in [1, 2], in particular finite-sample guarantees.

Moving beyond the consistency of such estimators (IPSW, outcome modeling and AIPSW) for the Risk Difference, we analyzed in a finite sample regime the IPSW estimator, which consists of re-weighting the trial so that it resembles the observational sample [see 2]. In particular, we established **finite sample bias and variance** (the literature mostly focuses on asymptotic results) and

upper bound on the risk of different versions of the estimator: oracle, semi-oracle, etc. We highlighted different regimes regarding the sample sizes of the trials and observational data, which can lead to practical recommendations in terms of data collection (doubling the size of the observational data leads to a smaller asymptotic variance than doubling the size of the trial). Extending these analyses to the case of binary outcomes could provide practical guidance when dealing with discrete output. A major improvement will also be to derive the variance of the estimators for different sets of variables (treatment effect modifiers, prognostic variables, etc.).

Second axis of the PhD: Handling missing values and unmeasured covariates

The problematic of missing values is ubiquitous in data analysis practices and it is exacerbated when aggregating data of different sources. Naive approaches such as complete-case analysis which can lead to important bias, cannot be applied in **high-dimensional settings** when almost all data can be deleted (with only 300 features, and a probability to be missing for each individual and feature of 0.01%, complete case analysis would result in keeping around 5% of the rows). There exists an abundant literature on the topic [12, 21, 10, 14] and many methods available either to estimate some parameters (EM, multiple imputation) or to do supervised learning with missing values [9]. However, in the context of causal inference the literature is scarce, [13, 18, 11, 15] and these works only consider the case of a single dataset — or potentially multiple datasets with the same data distribution, i.e., sampled from the same population of interest – and do not treat the case of generalizing a treatment effect from an RCT to a target distribution defined with an observational dataset. *So as far as we know, this issue, although predominant in practice, has never been addressed.*

We will develop methods to handle both **sporadic missing data** (missing data for some individuals on some features) in the RCT and in the observational data, but also the so-called **systematic missing data when a variable is not available in either the RCT or the observational data**. The first case already requires establishing new conditions of identifiability with missing data and deriving estimators that handle missing values in the spirit of [15], who suggested Augmented IPW estimators using two random forests adapted to missing data. As for the second case, depending on which variables are missing, it may be necessary to turn to **sensitivity analyses** because the hypotheses of ignorability will no longer be verified. This problem is reminiscent of the highly challenging problem of **unobserved confounding** in classical observational studies.

Third axis of the PhD: Meta-Analyses to provide a better estimation of the causal measure of interest.

Combining information from different data sources is an intrinsic difficult task, notably due to the variability in the collected information (different vari-

ables, missing information due to merging...). Combining different studies (observational or clinical) in order to obtain a better estimation of the causal measure of interest (Risk Difference or Risk Ratio) is definitely a promising avenue. A proper understanding of the causal measure properties would allow us to aggregate directly the causal measure computed on each study, instead of needing to access the data from each study separately. Whereas this last axis is a long-term project with many technical challenges, we believe it constitutes an important direction for the future.

3 Application context and objective: decisions in medical emergencies

The Premedical team has different medical collaborations. One of the oldest collaboration is with the Traumabase group of APHP (Public Assistance - Hospitals of Paris) on polytraumatized patients. This project is mature in that we are testing real-time cell phone applications in the ambulance to help clinicians make decisions.

Major trauma denotes injuries that endanger the life or the functional integrity of a person. The WHO has recently shown that major trauma, –including road-traffic accidents, interpersonal violence, falls– remains a world-wide public-health challenge and a major source of mortality and handicap. An effective and timely management of trauma is crucial to improve outcomes, as delays or errors entail high risks for the patient.

A motivating application for this work is the generalization to a French target population – represented by the Traumabase registry – of the CRASH-3 trial [5], evaluating Tranexamic Acide (TXA) to prevent death from Traumatic Brain Injury (TBI).

CRASH-3 A total of 175 hospitals in 29 different countries participated to the randomized and placebo-controlled trial, called CRASH-3 [6], where adults with TBI suffering from intracranial bleeding were randomly administrated TXA [5]. Primary outcome was mortality (binary) after 28 days and secondary outcome is the Disability Rating Scale after 28 days of injury (ordinal indicator ranging from 0 to 29)

Traumabase To improve decisions and patient care in emergency departments, 30 French Trauma centers are collecting detailed clinical data from the scene of the accident to the exit of the hospital. The resulting database, the Traumabase, comprises to date 30 000 trauma admissions, and is permanently updated. The data are of unique granularity and size in Europe. However, they are highly heterogeneous, with both categorical (sex, type of illness...) and quantitative (blood pressure, hemoglobin level...) features, multiple sources, and many missing data. In fact 98% of the individuals have missing values. The cause of missing information is also coded, such as technical hurdles with

the measurement, or impossibility due to the severity of the patient's state. The Traumabase currently comprises around 8,270 patients suffering from TBI.

We have in mind this particular application when defining the different axis of the PhD thesis. However the methods are generic and can be applied for many other questions whether in the medical domain or in other domains such as economy, etc. In addition, even if the project is motivated by practical questions, the project requires strong methodological and theoritical contributions. Each contribution could help to have a better understanding of the treatment effect, as other causal measures than the Risk Difference are reported in the literature.

4 Laboratory - contact

The candidate will be supervised by both Julie Josse (expert in Missing Values and Causal Inference) and Erwan Scornet (expert in Random Forest, Statistical Learning, Missing Values). Julie Josse has many international connections in causal inference (she was invited to the semester on causality in Berkeley, to the Rousseeuw prize in Belgium, etc.) and often sends her PhD students to do research internships abroad, in particular with the Department of Statistics at Stanford University with whom she has many connections.

Premedical Team - Inria Montpellier The Premedical (Precision Medicine by Data Integration and Causal Learning) team[1], is a recent Inria-Inserm team located in Montpellier. It is an interdisciplinary team composed of statisticians, biostatisticians, machine learners, and clinicians. Premedical develops methods for optimal treatment policy (drug efficacy, who gets treated and when, etc.) from heterogeneous data (clinical trials, observational data) that come with methodological challenges. In particular, Premedical develops methods for causal inference, statistical learning, management of missing data, federated learning, etc. Premedical holds the missing data and causality research group[2] and has created a taskview on causal inference methods. The candidate will also be able to participate in the activities of the Inserm team, Idesp, specialized among others on respiratory diseases such as asthma and also specialists in the exposome.

References

[1] Bénédicte Colnet, Julie Josse, Erwan Scornet, and Gaël Varoquaux. Generalizing a causal effect: sensitivity analysis and missing covariates. *Journal of Causal Inference*, 2021.

[2] Bénédicte Colnet, Julie Josse, Gaël Varoquaux, and Erwan Scornet. Reweighting the rct for generalization: finite sample error and variable selection. 2022.

[1] https://team.inria.fr/premedical/

[2] https://misscausal.gitlabpages.inria.fr/misscausal.gitlab.io/

[3] Bénédicte Colnet, Julie Josse, Gaël Varoquaux, and Erwan Scornet. Risk ratio, odds ratio, risk difference... which causal measure is easier to generalize? *arXiv preprint arXiv:2303.16008*, 2023.

[4] Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review. *Statistical Science*, 2023.

[5] CRASH-3. Effects of tranexamic acid on death, disability, vascular occlusive events and other morbidities in patients with acute traumatic brain injury (CRASH-3): a randomised, placebo-controlled trial. *The Lancet*, 394(10210):1713–1723, 2019.

[6] Yashbir Dewan, Edward Komolafe, Jorge Mejìa-Mantilla, Pablo Perel, Ian Roberts, and Haleema Shakur-Still. CRASH-3: Tranexamic acid for the treatment of significant traumatic brain injury: study protocol for an international randomized, double-blind, placebo-controlled trial. *Trials*, 13:87, 06 2012.

[7] MA Hernàn and JM Robins. *Causal Inference: What If*. 2020.

[8] Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, Cambridge UK, 2015.

[9] Julie Josse, Nicolas Prost, Erwan Scornet, and Gaël Varoquaux. On the consistency of supervised learning with missing values. *arXiv preprint*, 2019.

[10] Julie Josse and Jerome P. Reiter. Introduction to the special section on missing data. *Statist. Sci.*, 33(2):139–141, 05 2018.

[11] N. Kallus, X. Mao, and M. Udell. Causal inference with noisy and missing covariates via matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6921–6932, 2018.

[12] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002.

[13] Alessandra Mattei and Fabrizia Mealli. Estimating and using propensity score in presence of missing background data: an application to assess the impact of childbearing on wellbeing. *Statistical Methods and Applications*, 18(2):257–273, 2009.

[14] Imke Mayer, Aude Sportisse, Julie Josse, Nicholas Tierney, and Nathalie Vialaneix. R-miss-tastic: a unified platform for missing values methods and workflows, 2021.

[15] Imke Mayer, Erik Sverdrup, Tobias Gauss, Jean-Denis Moyer, Stefan Wager, and Julie Josse. Doubly robust treatment effect estimation with missing attributes. *Ann. Appl. Statist.*, 14(3):1409–1431, 2020.

[16] Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, AAAI'11, pages 247– 254. AAAI Press, 2011.

[17] Peter Rothwell. External validity of randomised controlled trials: To whom do the results of this trial apply? *Lancet*, 365:82–93, 01 2007.

[18] Shaun Seaman and Ian White. Inverse probability weighting with missing predictors of treatment assignment or missingness. *Communications in Statistics-Theory and Methods*, 43(16):3499–3515, 2014.

[19] Jerzy Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, 5(4):465 – 472, 1990.

[20] Elizabeth A Stuart, Stephen R Cole, Catherine P Bradshaw, and Philip J Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174:369–386, 2011.

[21] S. van Buuren. *Flexible Imputation of Missing Data*. Chapman and Hall/CRC, Boca Raton, FL, 2018.