Paper PP01

Innovation of Causal Inferential Frameworks for Biostatistics in Real-World Data and Evidence

Anwesha Roy, Novartis, Hyderabad, India

Abstract

Real-world data (RWD), such as electronic health records, reimbursement requests adjudicated by health insurance companies, and health survey data, is increasingly being utilized in drug development. To address issues in the use of real-world evidence (RWE), regulatory agencies have undertaken major initiatives. Since statistical research and considerations in the design, analysis, and interpretation of RWE studies are still a nascent domain, the causal model's methodological strategy addresses many of the major obstacles in the analysis of studies incorporating RWD, including non-randomization of the exposure, treatment noncompliance, time-varying confounding. While exhibiting a causal association is critical, research on improving data quality and representation will be necessary for high-quality pharmacovigilance studies. This poster delivers a statistical landscape assessment of relevant causal model frameworks on employing RWE to inform clinical trial design from which the pharmacovigilance domain may benefit by integrating machine learning and the causal inference paradigm.

Introduction

In biostatistics, many problems of interest are questions of causality. Causality implies relationships that are potentially exploitable for purposes of manipulation and control. Most scientific questions of interest in the health sciences are causal. For example, does the intervention reduce the risk of death in the target population? A randomized controlled trial is the gold standard for inferring causal relationships. When randomization is infeasible, researchers draw causal conclusions based on clinical trials or observational data. Learning the causal network structure from observational studies with or without additional experiments is one of the thrusts to explore biostatistics. Under a hypothetical directed graphical model framework with potential unmeasured confounding variables, causal inference is invoked to estimate the causal effect of a given exposure on a specific outcome or the causal effect of a particular pathway. Creating a sensitivity analysis of untestable identification assumptions and suggesting novel study designs enables a better assessment of causal effects. This domain has a wide range of scientific applications, including learning gene regulatory networks, estimating drug or vaccine efficacy when there is noncompliance and interference, evaluating biomarker surrogacy, learning optimal dynamic treatment regimes, and determining whether a specific pathway mediates treatment effects. In this poster, the speaker intends to demonstrate new advances in methodological research on causal inference, which will analyze clinical trials and observational studies with RWE from routinely acquired administrative healthcare data.

Causal model

A causal model is a statistical portrait of the process by which data is generated in a real-life phenomenon. It explicitly outlines the relationships between variables, the mechanisms by which they are assigned, and any uncertainties due to unmeasured factors. Neyman (1923) introduced an early causal model that described the potential outcomes of a group of units in a randomized experiment. Rubin (1973) expanded on this concept to develop causal inference models in randomized experiments and observational studies. These models are now known as Neyman-Rubin causal models (NRCMs). An alternative class of causal models, known as nonparametric structural equations with independent error models (NPSEM-IE) or structural causal models (PSCMs), was proposed by (Pearl, 2009). These models are based on graphical models and provide a different framework for modeling causality.

Neyman-Rubin causal model (NRCM)

Neyman-Rubin causal models (NRCMs) (Rubin, 1973) consist of three fundamental elements: units (such as patients), treatments (active treatments or control), and potential outcomes. In an NRCM, the assignment mechanism must be explicitly formulated, which probabilistically determines how individual patients are assigned to treatments. The critical parameter in the model denoted as $\pi(A \mid W, Y_0, Y_1)$, specifies the conditional probability of a patient receiving treatment A (active or control) based on their covariates W and potential outcomes Y_0 and Y_1. In clinical studies, unconfounded and probabilistic assignment mechanisms are commonly used. It is essential to distinguish the "causal estimand" defined in the addendum to the ICH E9(R1) guideline (2019). The International Council for Harmonisation (ICH) states that the central questions for drug development and licensing are to establish the existence and to estimate the magnitude of treatment effects: how the outcome of treatment compares to what would have happened to the same subjects under alternative treatment (i.e., had they not received the treatment, or had they received a different treatment). Researchers have developed various methods within NRCMs by incorporating different causal assumptions. These methods aim to provide valid causal or treatment effects by mathematically controlling for the impact of observed confounding variables by selecting a subset of the study sample or using regression methods.

Pearl's structural causal Model (PSCM)

A structural causal model (PSCM) (Pearl, 2009) is a widely used one that represents a system of equations describing the data generation process of interest. The model is entirely specified by endogenous variables (dependent variables), exogenous variables (independent variables), and deterministic functions. For example, in a PSCM representing a patient's status in a clinical study, the hierarchical model can be expressed as:

$$\begin{aligned} W|U_w &= g_W(U_W) \\ A|W, U_a &= g_a(W, U_a) \\ \boldsymbol{Y}|W, A, U_Y &= g_Y(W, A, U_Y) \end{aligned} \tag{1}$$

The endogenous variables in this example are the observed variables $\boldsymbol{O}$, which include the patient's baseline covariates W, assigned treatment A, and outcome Y. In other models, the endogenous

variables may also include unobserved variables of interest. The exogenous variables $\boldsymbol{U} = (U_W, U_a, U_Y)$ account for all other latent or unobserved sources that can potentially influence the endogenous variables (W, A, Y) via their respective structural functions $\boldsymbol{g} = (g_W, g_a, g_Y)$ which are assumed to be known. The parents of an endogenous variable, such as A, are the input endogenous variables of its corresponding structural function.

Causal assumptions in a structural causal model (PSCM) are defined by the specification and constraints placed on the deterministic functions $\boldsymbol{g}$ of the endogenous variables and the relationships between those variables and their parent(s). These assumptions also include the joint distribution of the exogenous variables $\boldsymbol{U}$. In PSCMs, the function vector $\boldsymbol{g}$ describing the relationship between variables is typically unrestricted. The exogenous variables (U_W, U_a, U_Y) are also assumed to be independently distributed. This independence property allows PSCMs to be referred to as nonparametric structural equations. By combining the functional relationships $\boldsymbol{g}$ and the distribution of the exogenous variables $\boldsymbol{U}$, the observed data $\boldsymbol{O}$ for a patient is determined in the PSCM framework.

In a structural causal model (PSCM), potential outcomes can be designed like Neyman-Rubin causal models (NRCMs). Specifically, the likely outcomes for treatment $A = 1$ and $A = 1$ can be defined as $Y_1 = g_Y(W, A = 1, U_Y)$ and $Y_0 = g_Y(W, A = 0, U_Y)$, respectively. This allows for the specification of causal quantities, such as the average treatment effect (ATE) as $\mathrm{ATE} = \mathbb{E}[Y_1 - Y_0]$. The assumption of an unconfounded mechanism in NRCMs corresponds to the assumption in PSCMs that the unmeasured error term (UA) is independent of the unmeasured error term U_Y when conditioning on W. This assumption is necessary to ensure valid causal inferences. Directed acyclic graphs (DAGs) are the heart of visually representing some of the assumptions made in PSCMs. These DAGs can visualize the relationships and dependencies between variables in the PSCM framework. In summary, the assumption of an unconfounded mechanism in NRCMs aligns with the assumption in PSCMs that the unmeasured error terms are independent U_A and U_Y, conditioned on the covariates W.

Causal inference pipeline

Jie Chen and White (2023); Martin Ho and White (2023) provided a causal framework of a comprehensive guide for designing, conducting, analyzing, and interpreting studies involving causal inference, such as clinical studies that use real-world data as external controls. This framework consists of six key steps:

1. Describing the observed data and the data-generating experiment: At the beginning of the roadmap, the researcher describes the observed data and the underlying data-generating process in the real world.

2. Specifying a realistic model for the probability distribution of the observed data: In this step, the researcher incorporates knowledge about the system, such as the treatment assignment mechanism, to construct a realistic model for the probability distribution of the observed data.

3. Specifying the target estimand of the observed data distribution: The researcher then moves into the causal world and explicitly defines the target estimand, representing the causal quantity of interest within the observed data distribution.

4. Specifying an estimator of the target estimand, respecting the statistical model: Here, the researcher identifies and specifies an appropriate estimator that respects the chosen statistical model and can estimate the target estimand.

5. Specifying an estimator of the sampling distribution of the estimator of the target estimand for statistical inference: In this step, the researcher defines an estimator for the sampling distribution of the estimator of the target estimand. This allows for statistical inference and quantification of uncertainties associated with the estimated causal effect.

6. Interpreting results and performing sensitivity analysis: The final step involves interpreting the results obtained from the study and conducting sensitivity analysis to assess the robustness of the findings to different assumptions and modeling choices.

By following this roadmap, researchers can ensure a systematic approach to causal inference studies, starting from describing the observed data and progressing through the specification of target estimands, estimation, inference, and interpretation of results. We choose the roadmap of Statistical Learning proposed by Van der Laan, Rose, et al. (2011) and Petersen and van der Laan (2014) as an illustrative example of a causal inference framework.

The data-generating experiment using real-world data (RWD) in regulatory clinical studies involves randomly sampling patients from a target population and observing them over time. This experiment may be independently repeated to obtain multiple sets of data. The target population is patients in a collection of hospitals with a confirmed disease diagnosis who are eligible to participate in a clinical study. This represents the baseline of the longitudinal data. Alternatively, the experiment can involve case-control sampling or a two-stage sampling approach. In the case of a clinical study using an external control, the target population is a superset of eligible patients in clinical research and external control. The observed data structure, denoted as O, for a patient is represented as $\boldsymbol{O} = (W, A, \Delta, Y)$ where:

- W represents the patient's baseline covariates.
- A represents the patient's treatment assigned at baseline.
- Δ indicates the patient's outcome of interest.
- Y being the observed random variable.

For a study with multiple time points, the data structure of the outcome on a patient is denoted as:

- $\boldsymbol{L}(t)$ is a vector of time-dependent variables measured at time t.
- $\Delta(t)$ is an indicator of $\boldsymbol{L}(t)$ being measured at time t.

This may include the current status of time till events such as death and other time-dependent covariates. The researcher observes the covariates W and treatment A, along with $(t, \boldsymbol{L}(t))$, across multiple time points $t = 1, 2, \cdots, K$. The outcome indicator $\Delta(K+1)$ and the outcome measured after time K, $\Delta(K+1)$, are also recorded. Therefore, the observed data for a patient denoted as $\boldsymbol{O}$ is $\boldsymbol{O} = (W, A, \Delta(1), \Delta(1)\boldsymbol{L}(1), \cdots, \Delta(K)\boldsymbol{L}(K), \Delta(K+1), \Delta(K+1)Y)$. This data structure allows researchers to simulate data on a computer compatible with the actual experiment, enabling statistical analyses and validation of the methods used in regulatory clinical studies. One example of causal inference in a Real-World Evidence (RWE) study is studying the impact of a medication on patient outcomes. The following sub-algorithms can outline the probabilistic modeling and the practitioner's guidelines:

Selecting a class of probabilistic model from EDA

I: Propose a probability distribution model and simulate $Y_n \overset{iid}{\sim} P_0$.

II: $O = (W, A, \Delta, \Delta Y)$: observed random vector.

III: Augment $\mathcal{M}^{\text{obs}}$ (the class of all possible probability distribution on $\boldsymbol{O}$ with knowledge about data generating process.

IV: $O \sim P_0 \Rightarrow P_0 \in \mathcal{M}^{obs}$.

V: Problem: Find $\Psi^{obs} : \mathcal{M}^{obs} \to \mathbb{R}^d$.

Specification of the target estimand

I: Define patient i 's Casual Effect (C) e.g., $Y_{1i} - Y_{0i}$

II: $\mathcal{M}^C$: all possible distributions of (X, U) in the causal model.

III: Translate question into a causal quantity Ψ^c in terms of counterfactual or potential outcomes, e.g. select ATE as causal quantity then $\Psi^c = \mathbb{E}Y_1 - \mathbb{E}Y_0$.

Identifiability of the causal quantity Ψ^c from P_0

I: Express $\Psi^c\,(= \mathbb{E}Y_1 - \mathbb{E}Y_0)$ as a function of Ψ^{obs}.

II: Find a target estimand $\Psi^{obs} : \Psi^C = \Psi^{obs}$ under appropriate causal assumptions.

III: Find $\Psi^{obs}\,(P_0) = \mathbb{E}_{w,0}\left[\mathbb{E}_0(Y \mid A = 1, \Delta = 1, W) - \mathbb{E}_0(Y \mid A = 0, \Delta = 1, W)\right]$.

IV: Checking assumptions of randomization, missing at random and positivity on Ψ^C and commit to Ψ^{obs}.

Estimator for the target estimand Ψ^{obs}

I: Find estimator $\widehat{\Psi}$ for $\mathbb{E}_0(Y = 1 \mid A = 1, \Delta = 1, W)$ and $\mathbb{E}_0(Y = 1 \mid A = 0, \Delta = 1, W)$.

II: Specify a method for statistical inference of the chosen estimator using parametric g-computation, matching estimators, doubly robust estimation, inverse probability weighting, TMLE, and Bayesian methods.

III: Quantify the uncertainty using confidence intervals and sensitivity analysis.

Aligned with the algorithmic workflow proposed by Martin Ho and White (2023), we recommend using nonparametric Bayesian methods, which have a long history of flexibility in choosing the class of probability distributions from $\mathcal{M}_{obs}$ (Dunson, 2010; Hjort, Holmes, Müller, & Walker, 2010). The posterior analysis can be done using Markov chain Monte Carlo methods, variational Bayes, or approximate Bayesian inference (Foster et al., 2019; Ryan, Drovandi, McGree, & Pettitt, 2016).

Workflow for practitioners

Let us say one wants to investigate whether a specific medication for hypertension is effective in reducing the risk of stroke in patients. To perform the causal inference analysis, they can follow these

steps in R:

Stage 1: **Study design**

- Identify a cohort of patients diagnosed with hypertension from a real-world database.
- Define an exposed group (patients prescribed the medication of interest) and a control group (patients not prescribed the medication).

Stage 2: **Data Acquisition and preprocessing**

- Obtain access to a relevant dataset with variables such as medication use, patient demographics, comorbidities, and outcomes.
- Clean and preprocess the data, handling missing values, outliers, and potential data quality issues.

Stage 3: **Covariate adjustment**

- Identify potential confounding factors influencing medication prescription and stroke risk (e.g., age, sex, comorbidities, prior stroke history).
- Utilize propensity score matching or regression models to balance the covariates across the exposed and control groups.

Stage 4: **Outcome analysis**

- Assess the primary outcome of interest (e.g., incidence of stroke) between the exposed and control groups.
- Utilize appropriate statistical methods such as logistic regression or Kaplan-Meier survival analysis to estimate the treatment effect while accounting for residual confounding.

Stage 5: **Sensitivity analysis**

- Conduct sensitivity analyses to evaluate the robustness of the findings to potential biases or unmeasured confounders. Examples could include inverse probability weighting, instrumental variable analysis, or multiple imputations, among many of the available literature.

Throughout the analysis, researchers must be cautious about limitations and consider biases commonly associated with RWE studies, such as selection bias, confounding, and information bias.

Conclusion

We conclude the article by highlighting the current landscape of using causal inference methods and real-world data in regulatory decision-making for drugs, biologics, and medical devices. However, there are specific challenges and opportunities.

1. **Opportunities**

   (a) **Informed benefit-risk assessment:** Causal inference methods and RWE studies provide valuable evidence to inform benefit-risk assessments of drugs and biologics in premarket applications, which can lead to better decision-making and improved patient safety.

(b) **Use of single-arm studies for medical devices:** Single-arm studies have been used to support premarket approval applications, providing evidence against external controls such as registries that allow for faster approval and availability of new medical devices.

(c) **Regulatory adherence:** Approaches like the two-stage paradigm of propensity score methods can align with traditional regulatory pathways and satisfy requirements such as prospective specification and outcome-blinding, which facilitates the integration of causal inference methods into the regulatory framework.

1. **Challenges**

(a) **Limited use for efficacy claims:** Clinical studies with external control or borrowing strength from external data rarely support efficacy claims for drugs and biologics. There may be a misperception that these methods do not conform to regulatory requirements. This hinders the wider adoption of causal inference methods in generating evidence for regulatory considerations.

(b) **Concerns about adherence to regulatory requirements:** The use of forward and backward selection to identify confounders in causal inference studies, although widely applied outside the regulatory setting, violates the outcome-blinding principle of the two-stage paradigm. This raises concerns about regulatory compliance when applying causal inference methods.

(c) **Limited use of advanced methods:** Advanced methods like the potential outcomes model for causal inference and targeted maximum likelihood estimation (TMLE) have not been widely used in regulatory settings. This may be due to clinicians' difficulty interpreting estimates from complex ML-based models and preferring more straightforward, interpretable approaches.

(d) **Statistical challenges:** The nonparametric Bayesian approach for choosing the class of probability distributions (P_0) and functional estimation in structural equations in the presence of latent (exogenous) variables could be an open research challenge. Also, feature selection and testing procedures are critically challenging problems when the dimension of covariates is polynomial or exponential in order of sample size $p \gg n$.

References

Dunson, D. B. (2010). Nonparametric bayes applications to biostatistics. *Bayesian nonparametrics*, *28*, 223–273.

Foster, A., Jankowiak, M., Bingham, E., Horsfall, P., Teh, Y. W., Rainforth, T., & Goodman, N. (2019). Variational bayesian optimal experimental design. *Advances in Neural Information Processing Systems*, *32*.

Hjort, N. L., Holmes, C., Müller, P., & Walker, S. G. (2010). *Bayesian nonparametrics* (Vol. 28). Cambridge University Press.

Jie Chen, H. W., Martin Ho, & White, R. (2023). The current landscape in biostatistics of real-world data and evidence: Clinical study design and analysis. *Statistics in Biopharmaceutical Research*, *15*(1), 29–42. Retrieved from https://doi.org/10.1080/19466315.2021.1883474 doi: 10.1080/19466315.2021.1883474

Martin Ho, H. W., Mark van der Laan, & White, R. (2023). The current landscape in biostatistics of real-world data and evidence: Causal inference frameworks for study design and analysis.

Statistics in Biopharmaceutical Research, *15*(1), 43–56. Retrieved from https://doi.org/10.1080/19466315.2021.1883475 doi: 10.1080/19466315.2021.1883475

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. *Ann. Agricultural Sciences*, 1–51.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Petersen, M. L., & van der Laan, M. J. (2014). Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology*, *25*(3), 418–426.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, 159–183.

Ryan, E. G., Drovandi, C. C., McGree, J. M., & Pettitt, A. N. (2016). A review of modern computational algorithms for bayesian optimal design. *International Statistical Review*, *84*(1), 128–154.

Van der Laan, M. J., Rose, S., et al. (2011). *Targeted learning: causal inference for observational and experimental data* (Vol. 4). Springer.

Contact Information

Your comments and questions are valued and encouraged.

Contact the author at:
Name: Anwesha Roy
Position: Senior Statistical Programmer
Company: Novartis Healthcare Private Limited
Address: Salarpuria-Sattva Knowledge City, Raidurg, Rangareddy District, Hyderabad, 500081, India
Work Mail: anwesha.roy@novartis.com
Alternative Mail: anwesharoy19041997@gmail.com

Disclaimer

All views in the paper are the author's personal views and do not reflect Novartis's views.