Universiteit Utrecht

Methodology & Statistics for the Behavioural, Biomedical and Social Sciences

Research Report

Causal Inference with Observational Data: Confidence Intervals in Sequential Target Trial Emulation, Coverage and Computational Efficiency

Author:

Florian Metwaly

Student-Nr.: 0778265

Supervisor:

Oisín Ryan

Wouter van Amsterdam

Ethical Approval Case Nr.: 24-2059

Journal: *BMC Medical Research Methodology*

December 22, 2024

Wordcount: 2478

Contents

1 Introduction

Randomized Controlled Trials (RCTs) are the gold standard for causal inference but are often infeasible due to practical or ethical constraints (M. A. Hernán & Robins, 2016; Sanson-Fisher, Bonevski, Green, & D'Este, 2007). In such cases, large-scale observational datasets, like electronic health records (EHRs), are increasingly used for causal analysis (Bakker, Goossens, O'Kane, Uyl-de Groot, & Redekop, 2021). Target Trial Emulation (TTE; M. A. Hernán and Robins 2016) is a robust framework that mimics RCTs to prevent biases, such as immortal time bias (M. A. Hernán, Sauer, Hernández-Díaz, Platt, & Shrier, 2016). TTE has become the standard for causal analysis in epidemiology and biomedicine, especially for assessing medical treatments (Bakker et al., 2021).

A key concept in TTE is "time-zero," the point at which eligibility is assessed and follow-up begins (M. A. Hernán & Robins, 2016). While RCTs align time-zero by design, observational studies often require sequential TTE (M. A. Hernán & Robins, 2016), which emulates a series of trials starting on different calendar dates, reusing individual data across relevant trials.

Sequential TTE, though methodologically sound (Fu, 2023), presents practical challenges. Constructing datasets by duplicating individual records for each trial iteration can strain computational resources when applied to population-scale EHRs (Dickerman, García-Albéniz, Logan, Denaxas, & Hernán, 2023; Xie, Bowe, & Al-Aly, 2023). Additionally, the same individuals are included in multiple trials. As a result, observations are no longer independent. This necessitates the use of methods like non-parametric bootstrapping for inference (Maringe et al., 2020). However, these methods can be computationally expensive.

This project aims to develop computationally efficient and statistically valid methods for sequential TTE. First, we will implement sequential TTE in Julia, which offers advantages in speed and memory efficiency over R (Stefan Karpinski, 2012). Benchmarking against the R package *TrialEmulation* (Su, Rezvani, Seaman, Starr, & Gravestock, 2024) will evaluate these advantages. Second, a simulation study will compare the non-parametric bootstrap with naïve and sandwich estimators for variance estimation (Danaei, Rodríguez, Cantero, Logan, & Hernán, 2013; M. Hernán, Brumback, & Robins, 2000). We will assess coverage, confidence interval width, power, and Type-I error rates under varying dataset sizes and

reuse levels of individual data.

We expect bootstrapped standard errors to provide robust inference, with sandwich estimators performing well in some conditions. However, the computational cost of bootstrapping requires evaluating whether its statistical advantages justify the runtime, particularly in large-scale trials. This trade-off between statistical performance and computational efficiency is central to determining the practicality of bootstrapped confidence intervals.

2 Target Trial Emulation

Randomized control trials are the preferred method in science to gain causal inference. The main assumption behind RCTs is that through the randomization of treatment assignment, the different treatment groups are comparable. Unobserved confounding, therefore, does not interfere with the results from statistical analysis, and the effects can be attributed to the assigned treatment (M. A. Hernán, Wang, & Leaf, 2022). Additional blinding further excludes the occurrence of placebo effects or differences in treatment due to the expectations of the given treatment (Meldrum, 2000). However, randomized control trials often are unethical or not practical, for example due to limited availability of the population or the length of follow-up time (Sanson-Fisher et al., 2007). Therefore gaining causal insights from observational data has been a long prevailing topic of research (Nichols, 2007; Wold, 1956).

Adjustment for confounder bias can be done through the design of the model or study, as it is common, for example in econometrics (Igelström et al., 2022; Listl, Jürges, & Watt, 2016).

Increasingly, researchers in the health sciences have noticed that mimicking randomization and dealing with confounding is necessary but not sufficient to gain causal inferences from observational data. When analyzing observational data, the researcher is free to make design choices that may accidentally introduce different forms of selection biases or lead to a mismatch between the target quantity being estimated and the quantity of clinical interest (Fu, 2023).

Immortal time bias, a type of bias that occurs when a period of time during which the event of interest cannot occur is incorrectly classified in the analysis, is a specific form of selection

bias often seen in observational studies (Yadav & Lewis, 2021).

TTE is a methodological framework to help design the analysis of observational data to make it possible to gain causal inference from them. To perform TTE one first has to specify the ideal randomized control trial, which one would perform if it was feasible. The design of the TTE then mimics this idealized RCT as closely as possible (M. A. Hernán et al., 2008). For this, a trial protocol needs to be specified, which includes both, the ideal RCT and the adjustments that need to be made for handling observational data (M. A. Hernán et al., 2008). An example protocol can be found in Table 1.

Table 1: Comparison of Target Trial and Emulated Target Trial Protocol Components

Protocol Component	**Target Trial**	**Emulated Target Trial**
Eligibility Criteria	Individuals aged 18–65 with no prior history of COVID-19 infection, not currently pregnant, and no history of severe allergic reactions to vaccines.	Individuals from an observational database aged 18–65 with no record of prior COVID-19 diagnosis, no pregnancy recorded at baseline, and no known contraindications to vaccination.
Treatment Strategies	Vaccination with a two-dose mRNA COVID-19 vaccine schedule versus no vaccination.	Comparison between individuals receiving the mRNA COVID-19 vaccine (first and second doses) and those who remain unvaccinated throughout the follow-up period.
Assignment Procedures	Random assignment of eligible individuals to receive either the vaccine or placebo, with participants blinded to their treatment assignment.	Individuals self-select into vaccinated or unvaccinated groups based on personal or provider decision. Propensity scores are used to adjust for baseline differences between groups.

Protocol Component	**Target Trial**	**Emulated Target Trial**
Follow-up Period	Follow participants from the date of randomization until the earlier of (1) confirmed symptomatic COVID-19, (2) withdrawal from the study, or (3) 6 months.	Follow individuals from the first vaccination date (or an analogous index date for unvaccinated individuals) until the earlier of (1) symptomatic COVID-19 diagnosis, (2) loss to follow-up, or (3) 6 months.
Outcome	Incidence of symptomatic COVID-19 (confirmed by PCR test).	Incidence of symptomatic COVID-19 (based on confirmed diagnostic codes or PCR test results in the observational dataset).
Causal Constraints of Interest	Intention-to-treat.	Observational equivalent to intention-to-treat effect.
Analysis Plan	Kaplan-Meier survival analysis to estimate the cumulative incidence of symptomatic COVID-19 in each group. Cox proportional hazards model to estimate hazard ratios (HR) for COVID-19 in the vaccinated versus unvaccinated group.	Use sequential inverse probability weighting (IPW) to adjust for time-varying confounders, fit a Cox proportional hazards model stratified by age group, and estimate hazard ratios. Baseline covariates: age, sex, comorbidities, occupation (e.g., healthcare worker), and geographic location.

The TTE framework has been developed to prevent the introduction of these biases. The key concept of TTE is that of t_0 alignment. T_0 describes the moment at which eligibility is determined, treatment is assigned and initiated and the follow-up period starts (M. A. Hernán et al., 2008).

To align t_0, two design choices can be made. One can select a single t_0. However, this leads to potentially losing a large number of observations due to them not being eligible at the chosen t_0 and therefore only being represented in the control group. Alternatively, the researcher can select *all* time points. When opting for this option, we speak of sequential TTE.

2.1 Sequential Target Trial Emulation

Sequential Target Trial Emulation addresses the challenge of non-unique t_0 in observational data by iterating through all observed time points to initiate a new trial with eligible individuals and their follow-up observations at each point. This process creates N trials, where individuals may participate in multiple trials.

The approach maximizes data use and addresses time-dependent confounding by dynamically capturing changes in exposure, eligibility, and covariates over time. It is particularly beneficial in scenarios where treatment probability or outcome risk varies over time (Keogh, Gran, Seaman, Davies, & Vansteelandt, 2023).

2.2 Challenges in sequential Target Trial Emulation

Sequential Target Trial Emulation, while a powerful method, faces several challenges that require further research. These challenges primarily arise from the complexity of the approach and the computational demands associated with its implementation.

2.2.1 Variance Estimation

Two problems arise when estimating the variance of the parameters in sequential Target Trial Emulation. Due to the *emulation* of the different trials, an individual can be included in multiple target trials (Pearl, 2009). Further, IP weighting is used to adjust for selection bias introduced by censoring and for confounder adjustment.

Including one observation in multiple trials violates the independence assumption of naïve variance estimation approaches. Furthermore, IP weighting introduces variability into the estimates, as weights are often highly variable, especially when probabilities of treatment or censoring are close to 0 or 1. This variability can lead to underestimated standard errors and overconfident conclusions if not addressed properly (Austin, 2016).

This inflation of the apparent sample size and the impact of the weights on the final model necessitates an adjusted variance estimator that accounts for both the correlated observations and the weighting process. A common solution to these issues is the use of sandwich variance estimators, as proposed by Lin and Wei 1989, which provide robust standard error estimates that remain valid under such complexities. Due to its common integration in statistical software (Hernan & Robins, 2010), it is a frequently used method in the literature (Caniglia et al., 2018; Dickerman, García-Albéniz, Logan, Denaxas, & Hernán, 2020). This sandwich variance estimator is also integrated into the *TrialEmulation* *R*-package (Su et al., 2024).

Although the sandwich variance estimator is commonly used in practice, the TTE literature typically recommends to use a non-parametric bootstrap approach to estimate the variance (Bakker et al., 2021; M. A. Hernán & Robins, 2016; Maringe et al., 2020). This is due to previous findings for example in the survival analysis literature, where bootstrapping confidence intervals in comparison with naïve approaches or sandwich estimators, were found to produce approximately correct standard errors and correct coverage rates for confidence intervals, while naïve and sandwich estimators produced biased standard errors and incorrect coverage rates (Austin, 2016).

The non-parametric bootstrap is a powerful tool for inference in case of the violation of various assumptions. While point estimates are accurate in the naïve model after performing sequential TTE, this is not the case for the variance estimates, as explained earlier. Bootstrapping is a resampling technique that involves repeatedly sampling with replacement from the observed data to generate new datasets, which are then used to estimate the sampling distribution of a statistic.

In sequential TTE, this is done by resampling from the baseline observations, creating a new observational dataset, including the baseline observations and their follow-up time, and

Step	Description
1	Sample n observations randomly with replacement from Y_0 to obtain a bootstrap dataset, denoted Y^*.
2	Calculate the bootstrap version of the statistic of interest, $h_K^* = h_K(Y^*)$.
3	Repeat steps 1 and 2 a large number of times, say B, to obtain an estimate of the bootstrap distribution.

Table 2: Procedure for Non-Parametric Bootstrap from Carpenter and Bithell 2000

finally performing sequential TTE on this new dataset.

Optimizations for the bootstrap exist at the cost of additional assumptions but are not further considered here (see, e.g. Binder, Kovacevic, and Roberts 2004; Li and Lawson 2024). Due to the resampling process, bootstrapping confidence intervals is a computationally intensive procedure. However, based on prior research, it is expected to outperform sandwich variance estimators.

The goals, therefore, are twofold. First, to compare the bootstrapped confidence intervals with the sandwich variance estimators in terms of the width of confidence intervals, coverage, Type-I error rate, and power. Following, the expected gains in statistical performance have to be weighted against the higher computational demands of the bootstrapping procedure. The goal is to propose a recommendation about the variance estimation method when performing sequential TTE in practice.

2.2.2 Computational Complexity

Sequential TTE is essentially performed by copying observations from the original dataset into a sequence of new datasets, which then all together build the dataset used for the analysis according to the protocol. This results in sequential TTE itself already being a computationally demanding method that potentially requires a high amount of memory to store the data. The programming language R is commonly used for data analysis in the health sciences. The *TrialEmulation* R-package (Su et al., 2024) is a popular choice for performing analyses

containing sequential Target Trial Emulation. While *R* is a popular choice, especially due to its broad availability of packages, it has rather big shortcomings, especially in relation to the high memory usage that sequential TTE imposes. *R*'s inefficient memory handling is a known problem in the community. Further, *R* processes code rather slowly compared to other languages, which again leads to a longer runtime, especially as datasets get bigger.

Julia is a modern programming language designed for efficient memory handling and fast compilation of code. *Julia* is built on the principles of running as fast as a low-level programming language, with an efficient memory handling system, while keeping the syntax easy and comprehensive (Stefan Karpinski, 2012). *Julia*'s just-in-time (JIT) compiler ensures that code is compiled and optimized at runtime, allowing it to execute as efficiently as compiled languages like C or Fortran. Moreover, *Julia*'s efficient memory management system minimizes memory overhead and fragmentation (Bezanson, Edelman, Karpinski, & Shah, 2017a), addressing the inefficiencies often encountered in *R*.

These features make *Julia* an ideal candidate for computationally demanding tasks such as sequential TTE. By leveraging *Julia*'s speed and memory efficiency, it is possible to scale analyses to larger datasets without compromising performance. *Julia*'s potential performance improvements over *R* and other programming languages have previously been shown, for example, by van Amsterdam 2024. This potential to reduce runtime and memory usage while maintaining clarity in code is the reason for developing a sequential TTE package for *Julia*.

2.3 TargetTrialEmulation.jl

TargetTrialEmulation.jl is developed for this project and aims to implement a computationally efficient software package to perform sequential Target Trial Emulation in the *Julia* programming language. During the development, a focus is laid on memory efficiency, fast run time, and an easy-to-use user interface while maintaining flexibility with respect to the estimated treatment effect. A further goal is to ensure executability on high-performance computers to be able to handle large-scale datasets with potentially millions of observations, like EHR. Finally, variance is estimated through the non-parametric bootstrap.

3 Simulation Study

3.1 Defining the Estimand and Variance Estimation Methods

The simulation study assumes a setting in which the individuals are observed regularly over a time period. The start of follow-up time is denoted by $k = 0$. Information about treatment status A_k, eligibility E_k, time-varying confounders L_k, and outcome of interest Y_k is collected at each timepoint $k = 0, 1, 2, ...K$. Time invariant confounders are denoted by V. Treatment status A_k is binary with treatment levels $a = (0, 1)$, equal to treated or untreated respectively. A marginal structural model (MSM) is fitted in the form of a pooled logistic regression to compare the variance estimation procedures. The estimand is defined as the average treatment effect (ATE)

$$\text{ATE} = \Pr\left(Y_k^{a=1} = 1\right) - \Pr\left(Y_k^{a=0} = 1\right). \tag{1}$$

The ATE is constructed as a marginal risk difference (MRD)

$$\begin{aligned}\widehat{\text{MRD}}_m(k) =& \frac{1}{n_m}\sum_{i=1}^{n} E_{m,i}\prod_{j=0}^{k}\left\{1 - \text{logit}^{-1}\left\{\mu\left(j, m, a = 0, V_i, L_{m,0,i}; \hat{\boldsymbol{\beta}}\right)\right\}\right\} \\ & - \frac{1}{n_m}\sum_{i=1}^{n} E_{m,i}\prod_{j=0}^{k}\left\{1 - \text{logit}^{-1}\left\{\mu\left(j, m, a = 1, V_i, L_{m,0,i}; \hat{\boldsymbol{\beta}}\right)\right\}\right\},\end{aligned} \tag{2}$$

where i is the patient index of the original observational dataset, the total number of patients enrolled in trial m is given by $n_m = \sum_{i=1}^{n} E_{m,i}$, $\text{logit}^{-1}(\cdot) = \exp(\cdot)/\{1 + \exp(\cdot)\}$ and $\mu\left(j, m, a, L_{m,0}, V; \hat{\boldsymbol{\beta}}\right) = \hat{\beta}_0(m) + \hat{\beta}_1(j) + \hat{\beta}_2 \cdot a + \hat{\boldsymbol{\beta}}_3^{\mathrm{T}} V + \hat{\boldsymbol{\beta}}_1^{\mathrm{T}} L_{m,0}$ for $j = 0, ..., k$ (Limozin, Seaman, & Su, 2024).

The model reflects the observational equivalence to an intention-to-treat effect and is defined by

$$\text{logit}\left\{\Pr\left(Y_k = 1 \mid Y_{k-1} = 0, A_0 = a, V, L_0\right)\right\} = \beta_0 + \beta_1 k + \beta_2 k^2 + \beta_3 a + \beta_4^{\top} V + \beta_5^{\top} L_0. \tag{3}$$

During estimation, the model is weighted using stabilized inverse propensity of treatment weights for confounder adjustment.

The sandwich-variance estimator, as implemented in *TrialEmulation*, is defined as

$$\hat{\Sigma} = \left\{\sum_{i=1}^{n} \frac{\partial \boldsymbol{U}_i(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right\}^{-1} \left\{\sum_{i=1}^{n} \boldsymbol{U}_i(\hat{\boldsymbol{\beta}})\boldsymbol{U}_i(\hat{\boldsymbol{\beta}})^{\mathrm{T}}\right\} \left\{\sum_{i=1}^{n} \frac{\partial \boldsymbol{U}_i(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right\}^{-1}, \tag{4}$$

with $\hat{\boldsymbol{\beta}}$ being the estimates of the $\boldsymbol{\beta}$ parameter of the model defined in Equation 3, and $\boldsymbol{U}_i(\hat{\boldsymbol{\beta}})$ being the weighted score function of the model evaluated at $\hat{\boldsymbol{\beta}}$ (Su et al., 2024). As a comparison, the bootstrap variance estimator is implemented as described in Table 2

3.2 Simulation setup

3.2.1 Data Generation

The data is simulated following an algorithm proposed by Young and Tchetgen Tchetgen 2014. The process is outlined in Table 3. The basic simulation setup follows Limozin et al.

Step	**Description**
1	For $m = 0, \ldots, K$: Draw $L_{m,i}$ from some choice of $f(L_m \vert \bar{A}_{m-1}, \bar{L}_{m-1}, Y_m = 0; \beta)$ evaluated at the previously generated $(\bar{A}_{m-1,i}, \bar{L}_{m-1,i})$.
2	Draw $A_{m,i}$ from some choice of $\Pr[A_m = 1 \vert \bar{L}_m, \bar{A}_{m-1}, Y_m = 0; \alpha]$ evaluated at the previously generated $(\bar{A}_{m-1,i}, \bar{L}_{m,i})$.
3	Draw $Y_{m+1,i}$ from some choice of $\Pr[Y_{m+1} = 1 \vert \bar{L}_m, \bar{A}_m, Y_m = 0; \theta]$ evaluated at the previously generated $(\bar{A}_{m,i}, \bar{L}_{m,i})$.
4	If $Y_{m+1,i} = 1$, then this is the last record in the dataset for observation i. Otherwise, generate another record for observation i (i.e., go to index $m + 1$).

Table 3: Outline of the data generating algorithm by Young and Tchetgen Tchetgen 2014. With $\Pr[A_m = 1 \vert \bar{L}_m = \bar{l}_m, \bar{A}_{m-1} = \bar{a}_{m-1}, Y_m = 0; \alpha]$ being a parametric model for the probability of receiving treatment in the interval m given the survival to m and history $(\bar{l}_m, \bar{a}_{m-1})$. For each of $i = 1, \ldots, n$ simulated observations, $\bar{L}_{-1,i} \equiv \bar{A}_{-1,i} \equiv Y_{0,i} = 0$ is implicitly defined (Young & Tchetgen Tchetgen, 2014).

2024. Baseline hazards are defined to result in low, medium, and high percentages of patients experiencing the outcome of interest during follow-up. Confounding strength and treatment prevalence are set to either low or high. These parameters will be evaluated across a range of sample sizes, varying from very small datasets ($n = 200$) to extremely large datasets ($n = 1,000,000$), reflecting the case of extensive EHR datasets.

Additionally, the effect of varying the number of times an observation is copied to a new trial is investigated, as this has implications for the violation of independence assumptions and the robustness of statistical estimators.
For each scenario, 1,000 datasets are simulated to ensure robust statistical evaluation of the performance metrics.

3.2.2 Performance Measures

The newly developed *Julia* package's performance is compared with that of the existing *TrialEmulation* *R*-package in terms of speed and memory usage. This comparison will only include the sequential TTE and the estimation of a naïve model to have comparable cases in both packages. In *R* the *bench* package is used to benchmark time and memory usage (Hester & Vaughan, 2023). In *Julia* the *BenchmarkTools.jl* package is used (Chen & Revels, 2016).
Multiple measures are used to compare the confidence intervals of the sandwich estimator and the non-parametric bootstrap. The empirical coverage probability measures the proportion of simulated datasets where the true parameter value lies within the confidence interval, assessing the accuracy of the interval estimation. The Type-I error rate indicates the frequency of incorrectly rejecting a true null hypothesis, while power reflects the probability of correctly rejecting a false null hypothesis.
The width of the confidence interval provides insight into the precision of the estimator, with narrower intervals typically preferred. Standard error quantifies the estimate's variability, and Monte Carlo standard error estimates the uncertainty introduced by the simulation process.
The entire sequential TTE process will also be timed. This evaluation aims to determine whether the improved variance estimates justify the additional computational expense of using the non-parametric bootstrap.
All analyses are done using *R* version 4.4.1 (R Core Team, 2024) or *Julia* version 1.11.2 (Bezanson, Edelman, Karpinski, & Shah, 2017b).

3.2.3 Expected Results

The proposed scenarios are chosen due to different behavior expectations. Concerning the computational comparison, the *Julia* package is expected to be consistently faster than *TrialEmulation*. Further, the *Julia* package is expected to be able to handle larger datasets, while *TrialEmulation* is expected to crash when the dataset becomes large.

Due to copying observations, the assumption of independence is violated. The non-parametric bootstrap estimation is expected to produce approximately correct coverage rates and Type-I error rates, as it accounts for the structure of the resampled data. In contrast, the naïve model is expected to underestimate the variability of the estimates, leading to overly narrow confidence intervals, incorrect coverage rates, and an inflated Type-I error rate.

Further, the sandwich estimator is expected to partially correct for the dependencies introduced by copying observations, improving the standard error estimates compared to the naïve model. However, it may still fall short of fully accounting for the complex correlation structure, leading to slight underestimation of variability and suboptimal coverage rates, especially in scenarios with high levels of observation reuse.

Bibliography

Austin, P. C. (2016). Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Statistics in Medicine*, *35*(30), 5642–5655. doi: 10.1002/sim.7084

Bakker, L. J., Goossens, L. M. A., O'Kane, M. J., Uyl-de Groot, C. A., & Redekop, W. K. (2021, September). Analysing electronic health records: The benefits of target trial emulation. *Health Policy and Technology*, *10*(3), 100545. doi: 10.1016/j.hlpt.2021.100545

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017a). Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, *59*(1), 65–98. (Publisher: Society for Industrial and Applied Mathematics)

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017b). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*(1), 65–98. Retrieved from `https://epubs.siam.org/doi/10.1137/141000671` doi: 10.1137/141000671

Binder, D., Kovacevic, M., & Roberts, G. (2004). Design-based methods for survey data: Alternative uses of estimating functions. In *Proceedings of the section on survey research methods* (pp. 3301–3312).

Caniglia, E. C., Zash, R., Jacobson, D. L., Diseko, M., Mayondi, G., Lockman, S., ... others (2018). Emulating a target trial of antiretroviral therapy regimens started before conception and risk of adverse birth outcomes. *Aids*, *32*(1), 113–120.

Carpenter, J., & Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, *19*(9), 1141–1164. (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291097-0258%2820000515%2919%3A9%3C1141%3A%3AAID-SIM479%3E3.0.CO%3B2-F) doi: 10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F

Chen, J., & Revels, J. (2016, Aug). Robust benchmarking in noisy environments. *arXiv e-prints*.

Danaei, G., Rodríguez, L. A. G., Cantero, O. F., Logan, R., & Hernán, M. A. (2013, February). Observational data for comparative effectiveness research: An emulation of

randomised trials of statins and primary prevention of coronary heart disease. *Statistical Methods in Medical Research*, *22*(1), 70–96. doi: 10.1177/0962280211403603

Dickerman, B. A., García-Albéniz, X., Logan, R. W., Denaxas, S., & Hernán, M. A. (2020). Emulating a target trial in case-control designs: an application to statins and colorectal cancer. *International journal of epidemiology*, *49*(5), 1637–1646.

Dickerman, B. A., García-Albéniz, X., Logan, R. W., Denaxas, S., & Hernán, M. A. (2023, September). Evaluating Metformin Strategies for Cancer Prevention: A Target Trial Emulation Using Electronic Health Records. *Epidemiology*, *34*(5), 690. doi: 10.1097/EDE.0000000000001626

Fu, E. L. (2023, August). Target Trial Emulation to Improve Causal Inference from Observational Data: What, Why, and How? *Journal of the American Society of Nephrology*, *34*(8), 1305. doi: 10.1681/ASN.0000000000000152

Hernan, M. A., & Robins, J. M. (2010). Causal Inference: What If.

Hernán, M. A., Alonso, A., Logan, R., Grodstein, F., Michels, K. B., Willett, W. C., ... Robins, J. M. (2008, November). Observational Studies Analyzed Like Randomized Experiments: An Application to Postmenopausal Hormone Therapy and Coronary Heart Disease. *Epidemiology*, *19*(6), 766–779. doi: 10.1097/EDE.0b013e3181875e61

Hernán, M. A., & Robins, J. M. (2016, April). Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available: Table 1. *American Journal of Epidemiology*, *183*(8), 758–764. doi: 10.1093/aje/kwv254

Hernán, M. A., Sauer, B. C., Hernández-Díaz, S., Platt, R., & Shrier, I. (2016, November). Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of Clinical Epidemiology*, *79*, 70–75. doi: 10.1016/j.jclinepi.2016.04.014

Hernán, M. A., Wang, W., & Leaf, D. E. (2022, December). Target Trial Emulation: A Framework for Causal Inference From Observational Data. *JAMA*, *328*(24), 2446–2447. doi: 10.1001/jama.2022.21383

Hernán, M. , Brumback, B., & Robins, J. M. (2000, September). Marginal Structural Models to Estimate the Causal Effect of Zidovudine on the Survival of HIV-Positive Men. *Epidemiology*, *11*(5), 561.

Hester, J., & Vaughan, D. (2023). bench: High precision timing of r expressions [Computer software manual]. (https://bench.r-lib.org/, https://github.com/r-lib/bench)

Igelström, E., Craig, P., Lewsey, J., Lynch, J., Pearce, A., & Katikireddi, S. V. (2022, November). Causal inference and effect estimation using observational data. *Journal of Epidemiology and Community Health*, *76*(11), 960–966. doi: 10.1136/jech-2022-219267

Keogh, R. H., Gran, J. M., Seaman, S. R., Davies, G., & Vansteelandt, S. (2023, June). Causal inference in survival analysis using longitudinal observational data: Sequential trials and marginal structural models. *Statistics in Medicine*, *42*(13), 2191–2225. doi: 10.1002/sim.9718

Li, T., & Lawson, J. (2024, March). A Generalized Bootstrap Procedure of the Standard Error and Confidence Interval Estimation for Inverse Probability of Treatment Weighting. *Multivariate Behavioral Research*, *59*(2), 251–265. (Publisher: Routledge _eprint: https://doi.org/10.1080/00273171.2023.2254541) doi: 10.1080/00273171.2023.2254541

Limozin, J. M., Seaman, S. R., & Su, L. (2024). Inference procedures in sequential trial emulation with survival outcomes: comparing confidence intervals based on the sandwich variance estimator, bootstrap and jackknife. *arXiv preprint arXiv:2407.08317*.

Lin, D. Y., & Wei, L. J. (1989, December). The Robust Inference for the Cox Proportional Hazards Model. *Journal of the American Statistical Association*, *84*(408), 1074–1078. doi: 10.1080/01621459.1989.10478874

Listl, S., Jürges, H., & Watt, R. G. (2016). Causal inference from observational data. *Community Dentistry and Oral Epidemiology*, *44*(5), 409–415. (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cdoe.12231) doi: 10.1111/cdoe.12231

Maringe, C., Benitez Majano, S., Exarchakou, A., Smith, M., Rachet, B., Belot, A., & Leyrat, C. (2020, October). Reflection on modern methods: trial emulation in the presence of immortal-time bias. Assessing the benefit of major surgery for elderly lung cancer patients using observational data. *International Journal of Epidemiology*, *49*(5), 1719–1729. doi: 10.1093/ije/dyaa057

Meldrum, M. L. (2000, August). A BRIEF HISTORY OF THE RANDOMIZED CON-

TROLLED TRIAL: From Oranges and Lemons to the Gold Standard. *Hematology/Oncology Clinics of North America*, *14*(4), 745–760. doi: 10.1016/S0889-8588(05)70309-9

Nichols, A. (2007). Causal Inference with Observational Data. *The Stata Journal*, *7*(4).

Pearl, J. (2009). *Causality*. Cambridge university press.

R Core Team. (2024). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Sanson-Fisher, R. W., Bonevski, B., Green, L. W., & D'Este, C. (2007, August). Limitations of the Randomized Controlled Trial in Evaluating Population-Based Health Interventions. *American Journal of Preventive Medicine*, *33*(2), 155–161. doi: 10.1016/j.amepre.2007.04.007

Stefan Karpinski, A. E. J. B., Viral Shah. (2012). *Why We Created Julia.* Retrieved 2024-10-10, from `https://julialang.org/blog/2012/02/why-we-created-julia/`

Su, L., Rezvani, R., Seaman, S. R., Starr, C., & Gravestock, I. (2024, February). *TrialEmulation: An R Package to Emulate Target Trials for Causal Analysis of Observational Time-to-event Data.* arXiv. (arXiv:2402.12083 [stat])

van Amsterdam, W. (2024). *The need for speed, performing simulation studies in R, JAX and Julia.* Retrieved 2024-12-13, from `https://vanamsterdam.github.io/posts/240308-jaxopt-vs-r-vs-julia/`

Wold, H. (1956). Causal Inference from Observational Data: A Review of End and Means. *Journal of the Royal Statistical Society. Series A (General)*, *119*(1), 28–61. (Publisher: [Royal Statistical Society, Oxford University Press]) doi: 10.2307/2342961

Xie, Y., Bowe, B., & Al-Aly, Z. (2023, March). Molnupiravir and risk of hospital admission or death in adults with covid-19: emulation of a randomized target trial using electronic health records. *BMJ*, *380*, e072705. (Publisher: British Medical Journal Publishing Group Section: Research) doi: 10.1136/bmj-2022-072705

Yadav, K., & Lewis, R. J. (2021). Immortal time bias in observational studies. *Jama*, *325*(7), 686–687.

Young, J. G., & Tchetgen Tchetgen, E. J. (2014, March). Simulation from a known Cox MSM using standard parametric models for the g-formula. *Statistics in Medicine*,

33(6), 1001–1014. doi: 10.1002/sim.5994