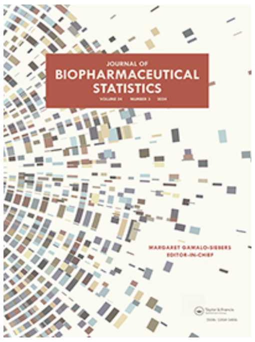


Journal of Biopharmaceutical Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/lbps20

genRCT: a statistical analysis framework for generalizing RCT findings to real-world population

Dasom Lee, Shu Yang, Mark Berry, Tom Stinchcombe, Harvey Jay Cohen & Xiaofei Wang

Taylor & Francis
Taylor & Francis Group

genRCT: a statistical analysis framework for generalizing RCT findings to real-world population

Dasom Lee[a], Shu Yang[a], Mark Berry[b], Tom Stinchcombe[c], Harvey Jay Cohen[c], and Xiaofei Wang[d]

[a]Department of Statistics, North Carolina State University, Elk Grove, USA; [b]Department of Cardiothoracic Surgery, Stanford University, Stanford, USA; [c]Department of Medicine, Duke University, Durham, USA; [d]Department of Biostatistics & Bioinformatics, Duke University, Durham, USA

ABSTRACT

When evaluating the real-world treatment effect, the analysis based on randomized clinical trials (RCTs) often introduces generalizability bias due to the difference in risk factors between the trial participants and the real-world patient population. This problem of lack of generalizability associated with the RCT-only analysis can be addressed by leveraging observational studies with large sample sizes that are representative of the real-world population. A set of novel statistical methods, termed "genRCT", for improving the generalizability of the trial has been developed using calibration weighting, which enforces the covariates balance between the RCT and observational study. This paper aims to review statistical methods for generalizing the RCT findings by harnessing information from large observational studies that represent real-world patients. Specifically, we discuss the choices of data sources and variables to meet key theoretical assumptions and principles. We introduce and compare estimation methods for continuous, binary, and survival endpoints. We showcase the use of the *R* package *genRCT* through a case study that estimates the average treatment effect of adjuvant chemotherapy for the stage 1B non-small cell lung patients represented by a large cancer registry.

ARTICLE HISTORY
Received 1 April 2024
Accepted 1 April 2024

KEYWORDS
Causal inference; covariates balance; genRCT; generalizability; Randomized clinical trials; real-world data; treatment effect heterogeneity

1. Introduction

Randomized clinical trials (RCTs) offer the highest level of evidence of treatment safety and efficacy in medical and pharmaceutical product developments, as randomization eliminates both measured and unmeasured confounders. However, patients enrolled in randomized clinical trials are conveniently ascertained and often represent a more restrictive patient group of the target population to which the new treatment will be given. Therefore, the treatment effects estimated by standard methods are found to lack external validity for the target real-world population (Kennedy-Martin et al. 2015; Rothwell 2005). On the other hand, real-world evidence studies (RWEs), including large population-based observational studies, registries, electronic health records, and medical claim databases, often contain a much larger number of patients of the same disease and represent a random sample of the target population.

Due to the lack of treatment randomization, there are always concerns about whether or not confounding has been addressed adequately in the analyses of the RWE data. In cancer research, there is an in-depth discussion on the strengths and limitations of utilizing data from RCTs and RWEs for comparative effectiveness analyses (Korn and Freidlin 2012; Visvanathan et al. 2017). The evaluation of the treatment effect in the RCT population and the RWE population is both crucial, as they offer

CONTACT Xiaofei Wang ✉ xiaofei.wang@duke.edu Department of Biostatistics & Bioinformatics, Duke University, 2424 Erwin Road, Suite 11020, Durham, NC 27705, USA

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10543406.2024.2333136

insights into different aspects of treatment effects, thereby assisting regulatory agencies in evaluating a drug's efficacy and safety during the approval process. Our method facilitates the generalization of treatment effect estimates from the RCT population to the RWE population. Analyzing both the similarities and differences in treatment effects between these populations provides complementary evidence about the drug's performance in both the RCT and RWE settings.

Statistical methods that allow generalization of RCT findings to a target population are in great need for informing better policy decision-making and countering mis-understanding in drug and device development. The problem of extending findings from RCT to a target population has been termed as generalizability (e.g., Cole and Stuart 2010; Dahabreh et al. 2019; Tipton 2013) and transportability (e.g., Pearl and Bareinboim 2011; Westreich et al. 2017). Most existing methods rely on direct modeling of the sampling score, the sampling analog of the propensity score. The subsequent sampling score adjustments include inverse probability of sampling weighting (IPSW; e.g., Cole and Stuart 2010), stratification (Tipton 2013), and augmented IPSW (AIPSW; Dahabreh et al. 2019). Most sampling score adjustment approaches require the sampling score model to be correctly specified. Moreover, weighting estimators involving inverse probability weighting are unstable if the sampling score is too extreme.

In this article, we consider the information contained in an RCT sample and an RWE sample, where the RCT sample is subject to patient selection bias and the RWE sample is representative of the target population with a known sampling mechanism. We review the theory and finite sample properties of new calibration weighting (CW) methods, reported in Lee et al. (2022, 2023). for improving the generalizability of the average treatment effect (ATE) of the trial. In contrast to the dominant approaches that focus on predicting sample selection probabilities, the CW methods estimate the sampling score weights directly by calibrating the covariates balance between the RCT sample and the design-weighted RWE sample to address the selection bias of the RCT sample.

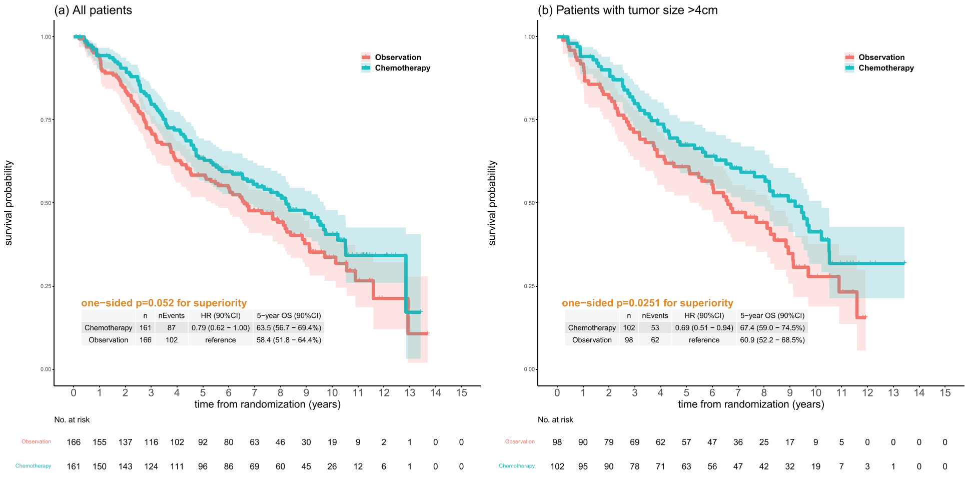
In this article, we will focus on the genRCT analysis. Given the complementarity of RCTs and observational studies, integrated analysis approaches are called for to efficiently exploit the relative strengths of the data from both RCTs and observational studies. Novel methods, such as pretesting elastic poolability of RCT and observational studies for estimating treatment effect heterogeneity over modifiers without or with hidden confounders and learning targeted, optimal, and interpretable individualized treatment regimes, will not be discussed. Systematic reviews of these methods can be found in Colnet et al. (2020) and Yang and Wang (2022).

In the rest of the article, we at first illustrate the problem of generalizing RCT results to a target patient population represented by real-world evidence studies (RWEs) with a real example of making inference on the average treatment effect of adjuvant chemotherapy in stage 1B NSCLC patients. In Section 3, we will review the theory and the finite sample properties of the CW estimators and their competitors, including both standard and augmented variants for binary, continuous and survival endpoints. Section 4 summarizes the workflow and key considerations of conducting the genRCT analysis to ensure its validity and completeness in practice. Section 5 provides a summary of the *R* package *genRCT*. In Section 6, we illustrate the genRCT analysis with the data from the motivating example. We will conclude the article with discussions in Section 7.

2. Motivating example

CALGB 9633 is a randomized phase III trial to determine the efficacy of adjuvant chemotherapy compared with observation in stage 1B non-small cell lung cancer (Strauss et al. 2008). The primary endpoint is overall survival (OS), the time from randomization to deaths of all causes. The eligibility criteria of this trial are stage 1B NSCLC (T2N0M0), with tumor size >3 cm (T2) and negative N1 and N2 nodes. Other patient requirements are age $>$ over 18 years, histologically documented NSCLC, and the tumor was removed by lobectomy or pneumonectomy. Eligible patients were randomized with equal allocation to adjuvant chemotherapy (paclitaxel, 200 mg/m2 and carboplatin, AUC = 6 mg/ml × min 4 cycles over 12 weeks) versus observation within 4–8 weeks of surgical resection. After 12 years of patient recruitment and follow-up, the results of CALGB 9633 were published in the Journal of Clinical Oncology. The final analysis was done after 155 deaths were observed (74 Chemotherapy, 81 Observation), with a median follow-up of 74 months. Overall survival (OS) was not significantly different (hazard ratio [HR], 0.83; CI, 0.64 to 1.08; $p = 0.12$). Exploratory analysis demonstrated a significant survival difference in favor of adjuvant chemotherapy for patients who had tumors $\geq$ 4 cm in diameter (HR, 0.69; CI, 0.48 to 0.99; $p = 0.043$). The results are reproduced in Figure 1(a, b), and the numerical difference is because a few patients with missing tumor size were removed for the illustration of the genRCT analysis. Grades 3 to 4 neutropenia were the predominant toxicity; there were no treatment-related deaths. The current NCCN guidelines on treating stage 1B NSCLC patients after surgery with adjuvant chemotherapy are largely based on the findings from CALGB 9633 (Ettinger et al. 2018).

It is well known that the patients participating in randomized clinical trials tend to be different from the target population represented by large population-based databases or registries (Pang et al. 2016). National Cancer Data Base (NCDB) is an oncology outcomes database collecting information on 72% of all new invasive cancer diagnoses in U.S (Boffa et al. 2017). We identified a total 16,012 patients diagnosed with NSCLC between years 2004–2016 with stage 1B disease who first had surgery and then received either adjuvant chemotherapy or on observation (i.e. no chemotherapy). As seen in Table 1, there are considerable differences between the patient population represented by CALGB 9633 and the patient population represented by the NCDB sample, though we have used the eligibility criteria of the RCT to define the boundary of the target population. This raises an important question – can the adjuvant chemotherapy benefit observed in CALGB 9633 be observed in the target population represented by the NCDB patients, especially for those patients with larger tumor size? Analysis based on NCDB data has been published (e.g. Morgensztern et al. 2016) supports the benefit of adjuvant chemotherapy for early stage NSCLC, including stage 1B, as compared to observation. However, the analysis is subject to hidden confounders, which cannot be fully addressed by multivariable Cox proportional hazards (PH) modelling and propensity score-based methods. Another concern is about a possible underpowered trial. Originally, the trial was designed to randomize 500 patients. Due to slow accrual, the trial was amended in 2000. The sample size was reduced to 384 patients, and two-sided test was changed to one-sided test. The trial has 80% power to detect a HR of 0.67 with 155 observed deaths at one-sided type I error rate of 0.05. Because of this, Katz and Saad (2009) have criticized CALGB 9633 as an underpowered trial with a methodologically questionable conclusion.



one−sided p=0.052 for superiority

	n	nEvents	HR (90%CI)	5−year OS (90%CI)
Chemotherapy	161	87	0.79 (0.62 − 1.00)	63.5 (56.7 − 69.4%)
Observation	166	102	reference	58.4 (51.8 − 64.4%)

No. at risk

Observation	166	155	137	116	102	92	80	63	46	30	19	9	2	1	0	0
Chemotherapy	161	150	143	124	111	96	86	69	60	45	26	12	6	1	0	0

one−sided p=0.0251 for superiority

	n	nEvents	HR (90%CI)	5−year OS (90%CI)
Chemotherapy	102	53	0.69 (0.51 − 0.94)	67.4 (59.0 − 74.5%)
Observation	98	62	reference	60.9 (52.2 − 68.5%)

No. at risk

Observation	98	90	79	69	62	57	47	36	25	17	9	5	0	0	0	0
Chemotherapy	102	95	90	78	71	63	56	47	42	32	19	7	3	1	0	0

Figure 1. Overall survivals for (a) all patients and (b) the patients with >4 cm tumor in CALGB 9633.

Table 1. Summary of CALGB 9633 and NCDB samples.

	CALGB 9633 (*N* = 327)	NCDB* (*N* = 16008)	Overall (*N* = 16335)
Treatment			
Observation	166 (50.8%)	11542 (72.1%)	11708 (71.7%)
Chemotherapy	161 (49.2%)	4466 (27.9%)	4627 (28.3%)
Sex			
Female	118 (36.1%)	7258 (45.3%)	7376 (45.2%)
Male	209 (63.9%)	8750 (54.7%)	8959 (54.8%)
age			
Mean (SD)	60.7 (9.79)	67.8 (10.3)	67.6 (10.3)
Median [Min, Max]	61.0 [34.0, 81.0]	69.0 [20.0, 90.0]	69.0 [20.0, 90.0]
Histology			
Squamous	130 (39.8%)	6235 (38.9%)	6365 (39.0%)
Other	197 (60.2%)	9773 (61.1%)	9970 (61.0%)
Tumor size			
Mean (SD)	4.59 (2.06)	4.81 (1.70)	4.81 (1.71)
Median [Min, Max]	4.00 [1.00, 12.0]	4.40 [3.10, 25.0]	4.40 [1.00, 25.0]

*A few NCDB patients from the extracted sample were removed from this summary and the genRCT analysis due to extreme outliers.

3. Methods for genRCT analysis

Novel statistical methods have been developed to address the question whether the findings from the RCT data are *generalizable* to the underlying target population, which is represented by the NCDB sample in the motivating example, and whether it is possible to increase statistical inference *efficiency* by utilizing all data of both sources.

3.1. Notations and data structure

To facilitate the discussion, we let A denote the treatment indicator with 0=control and 1=treatment, and Y a continuous or binary variable. X denotes a p-dimensional vector of covariates. δ is an indicator for the RCT participation, and $\tilde{\delta}$ is the complimentary indicator for the RWE participation. We would like to frame the generalizability question in a counterfactual framework (Imbens and Rubin 2015). The potential (counterfactual) outcome for $a = 1, 0$ is denoted as $Y(a), a \in \{0, 1\}$. Define the conditional average treatment effect (CATE) as $\tau(X) = E\{Y(1) - Y(0)|X\}$. To make an inference on the causal treatment effect, we are interested in estimating the population ATE $\tau_0 = E\{Y(1) - Y(0)\} = E\{\tau(X)\}$, where E is taken w.r.t the target population. The data structure for both RCT and RWE samples is presented in Figure 2, where we assume the RCT sample and the RWE sample are independent. In other words, we assume that δ_i and $\tilde{\delta}_i$ cannot be both equal to 1 for any subject i in the genRCT analysis.

3.2. Inverse probability of sampling weighting method

Inverse probability of sampling weighting (IPSW) methods were discussed in Cole and Stuart (2010); Buchanan et al. (2018); Dahabreh and Hernán (2019). Estimate the sampling score $\pi_\delta(X) = P(\delta = 1|X)$ using a logistic regression model $\pi_\delta(X; \hat{\eta})$. Inversely weight $\pi_\delta(X)$ to account for the shift of the covariate distribution.

$$\hat{\tau}^{\text{IPSW}} = \frac{\sum_{i=1}^{n} \pi_\delta(X_i; \hat{\eta})^{-1} A_i Y_i}{\sum_{i=1}^{n} \pi_\delta(X_i; \hat{\eta})^{-1} A_i} - \frac{\sum_{i=1}^{n} \pi_\delta(X_i; \hat{\eta})^{-1} (1 - A_i) Y_i}{\sum_{i=1}^{n} \pi_\delta(X_i; \hat{\eta})^{-1} (1 - A_i)},$$

where n is the sample size of the RCT sample. However, this estimator has a few limitations. Its estimation requires the availability of the baseline covariates of the population, and $\pi_\delta(X) = P(\delta = 1|X)$ is correctly modelled. Inverting the estimated sampling probability often leads

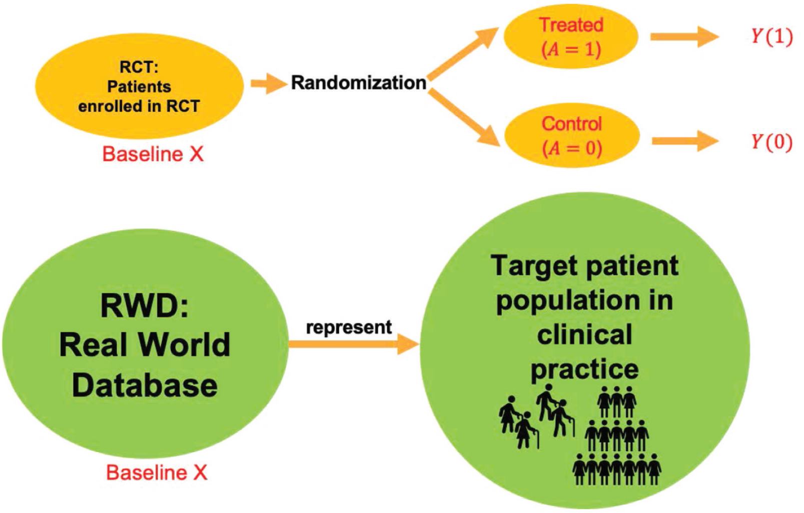


Figure 2. Data structure for a genRCT analysis.

to extreme weights that result in highly variable estimates. The model for the expected outcomes for the RCT sample and the model for the probability of trial participation can be combined to form an augmented IPSW estimator (AIPSW). Colnet et al. (2022) and Lee et al. (2023) have shown that AIPSW is doubly robust, i.e., it is consistent when either one of the two models is correctly specified and shown to be asymptotically normal when both models are consistently estimated at least at rate $n^{1/4}$.

3.3. Calibration weighting method

Calibration weighting (CW) is widely used to integrate auxiliary information in survey sampling (Wu and Sitter 2001) and causal inference (Hainmueller 2012; Qin and Zhang 2007; Wang et al. 2019). Hartman et al. (2015) implemented CW to estimate the population ATEs by combining RCTs with RWEs. Here, we review the calibrating weight method in Lee et al. (2023) for binary and continuous endpoints. The following assumptions are needed for the CW estimation, and similar assumptions are required for the IPSW estimator.

A1: (Consistency) $Y = AY(1) + (1 - A)Y(0)$

A2: (i) $\{Y(0), Y(1)\} \perp\!\!\!\perp A|(X, \delta = 1)$; and (ii) $0 < P(A = 1|X, \delta = 1) < 1$ with probability 1.

A3: $E\{Y(1)\text{-}Y(0)|X, \delta = 1\}=\tau(X)$; and (ii) $\pi_\delta(X)>0$ with probability 1.

Under these assumptions, the ATE $\tau_0 = E\{\tau(X)\}$is identifiable.

To balance the covariate distribution between RCT and RWE

$$E\left\{\frac{\delta}{\pi_\delta(X)}g(X)\right\} = E\left\{\tilde{\delta}dg(X)\right\} = E\{g(X)\}$$

where $d = 1/P(\tilde{\delta} = 1|X)$ is the design weight of RWE sample, we consider the following balancing constraint:

$$\sum_{i=1}^{N} \delta_i \omega_i g(X_i) = \sum_{i=1}^{N} \tilde{\delta}_i d_i g(X_i) / \sum_{i=1}^{N} \tilde{\delta}_i d_i \equiv \tilde{g},$$

where N is the total sample size of the RCT sample and the RWE sample, and $g(X)$ is vector-valued function, which can be moment functions of X, i.e. $\{X, X^2, X^3, \cdots\}$, or any transformations of X. We can estimate $\{\omega_i : \delta_i = 1\}$ by solving $\min \sum_{i=1}^{n} \omega_i \log \omega_i$ subject to the balancing constraint and $\omega_i \geq 0$ for all i, $\sum_{i=1}^{n} \omega_i = 1$. This optimization problem can be solved using convex optimization with Lagrange multiplier

$$L(\lambda; \mathcal{Q}) = \sum_{i=1}^{n} \omega_i \log \omega_i - \lambda^T \left\{ \sum_{i=1}^{n} \omega_i g(X_i) - \tilde{g} \right\}$$

The estimated weights are given by

$$\hat{\omega}_i = \frac{\exp\{\hat{\lambda}^T g(X_i)\}}{\sum_{i=1}^{n} \exp\{\hat{\lambda}^T g(X_i)\}}$$

and $\hat{\lambda}$ from solving

$$\sum_{i=1}^{n} \exp\{\lambda^T g(X_i)\}\{g(X_i) - \tilde{g}\} = 0.$$

Let $\pi_{A_i} = P(A_i = 1 | X_i; \delta_i = 1)$ be the treatment propensity score for subject i. We obtain the CW estimator

$$\hat{\tau}^{CW} = \sum_{i=1}^{n} \hat{\omega}_i \left\{ \frac{A_i Y_i}{\pi_{A_i}} - \frac{(1 - A_i) Y_i}{1 - \pi_{A_i}} \right\}$$

For RCTs, often $\pi_{A_i} = 0.5$ for all i, but it can be replaced by $\hat{\pi}_{A_i}$ for better efficiency. To establish the consistency and asymptotic normality for the CW estimator, we need the following assumptions

A4: Linearity of the CATE: $\tau(X) = \gamma_0^T g(X)$

A5: Log-linear sampling score: The sampling score of RCT participation follows a log-linear model, i.e.$\pi_\delta(X) = P(\delta = 1 | X) = \exp\{\eta_0^T g(X)\}$ for some η_0.

Under A5, we have the following connection between calibration weights and the sampling score: $\hat{\omega}_i - \{N\pi_\delta(X_i)\}^{-1} \xrightarrow{p} 0$. The asymptotic properties are followed. If either $\tau(X) = \gamma_0^T g(X)$ (A4 holds) or A5 holds, CW estimator is consistent for τ_0, and $N^{1/2}(\hat{\tau}^{CW} - \tau_0) \xrightarrow{d} \mathcal{N}(0, V)$, as $n \to \infty$. The CW estimator does not need parametric modelling on the sampling score or the outcome mean model. We can use a sandwich estimator or bootstrap to estimate the variance of the CW estimator.

3.4. Augmented calibration weighting method

Lee et al. (2023) proposed the augmented CW (ACW) estimator that is doubly robust and achieves the semiparametric efficiency bound when both nuisance models are correctly specified. However, the parametric approach is prone to model misspecification, especially when complex confounding exists. To cope with model misspecification, we adopted a method of sieves (Chen 2007), which allows flexible data-adaptive estimation of the nuisance functions while the ACW estimator retains the usual root-n consistency under regularity conditions.

In comparison with other nonparametric and machine learning methods, the ACW estimator with the sieve approximation is attractive: (1) unlike black-box machine learning methods, calibration weighting is straightforward and transparent; and (2) our framework allows for selecting important sieve basis terms that are related to the outcome to calibrate and enforcing the balance on these covariates for efficient estimation. In the presence of many covariates, variable or sieve basis selection for calibration becomes necessary. We classify covariates into three types: the covariates associated with trial participation and outcome as confounders, that affect outcome only through trial participation as instrumental variables (IVs), and that are predictive of the outcome as precision variables or outcome predictors. In other causal inference contexts, studies have shown that in addition to the confounding variables, including outcome predictors in the propensity score may improve efficiency, whereas including IVs may decrease efficiency (e.g., Tang et al. 2020). Despite the importance of proper basis selection for the efficient causal estimator, the current literature lacks a principled approach to guide basis selection for covariate balancing. Capitalizing on an explicit connection between calibration weighting and estimating equations under parametric models, we propose a penalized estimating equation approach for variable selection, emphasizing outcome predictors. Variable selection is conducted by applying a penalty for each variable included, similar to the least absolute shrinkage and selection operator (LASSO) method. This approach to variable selection can enhance both the predictive accuracy and interpretability of the resultant models.

An augmented CW estimator can be obtained (doubly robust), and it achieves the semiparametric efficiency bound if both nuisance models are correctly specified. A flexible data-adaptive estimation of the nuisance functions can be used to retain the usual root-n consistency under mild regularity conditions. By following the semiparametric theory (Tsiatis 2006), Lee et al. (2023) derived the semiparametric efficiency score $\phi(X, A, Y, \delta, \tilde{\delta})$ for τ_0:

$$\frac{\delta}{\pi_\delta(X)}\left[\frac{A\{Y-\mu_1(X)\}}{\pi_A}-\frac{(1-A)\{Y-\mu_0(X)\}}{1-\pi_A}\right]+\tilde{\delta}d\{\tau(X)-\tau_0\}$$

The score $\phi(X, A, Y, \delta, \tilde{\delta})$ involves unknown nuisance functions about the sampling score $\pi_\delta(X;\eta)$ and the outcome mean $\mu_a(X) = E(Y|X, A = a)$, which can be estimated from the RCT sample. The ACW estimator is given by

$$\hat{\tau}^{\mathrm{ACW}} = \sum\nolimits_{i=1}^{N} \delta_i\hat{\omega}_i\left[\frac{A_i\{Y_i-\mu_1(X_i;\hat{\beta}_1)\}}{\pi_{A_i}}-\frac{(1-A_i)\{Y_i-\mu_0(X_i;\hat{\beta}_0)\}}{1-\pi_{A_i}}\right]$$

$$+\left\{\sum\nolimits_{i=1}^{N}\tilde{\delta}_i d_i\right\}^{-1}\sum\nolimits_{i=1}^{N}\tilde{\delta}_i d_i\left\{\mu_1(X_i;\hat{\beta}_1)-\mu_0(X_i;\hat{\beta}_0)\right\}$$

Under A1-A3 and if either A4 or A5 holds, $\hat{\tau}^{\mathrm{ACW}}$ is consistent for τ_0. When both A4 and A5 hold, $N^{1/2}(\hat{\tau}^{\mathrm{ACW}} - \tau_0) \xrightarrow{d} \mathcal{N}(0, V_{\mathrm{eff}})$ in distribution, as $n \to \infty$, where V_{eff} is the semiparametric efficiency bound. The variance estimator can be calculated empirically using bootstrap.

To overcome the model misspecification issue inherent to parametric models, we consider flexible models for $\pi_\delta(X)$ and $\mu_a(X)$. We approximate $\pi_\delta(X)$ and $\mu_a(X)$ by the generalized sieves functions. To estimate $\mu_a(X)$, we apply the penalization technique for regression models with the pre-specified basis functions based on the RCT sample. The sieve method with power series as basis functions can be used.

3.5. *when* Y *and* A *are available in the RWE sample*

The covariate distribution of the RWE sample was used to adjust/calibrate the selection shift of the RCT sample. With the following assumption

A6: For $a \in \{0,1\}, E(Y|X, A = a, \tilde{\delta} = 1) = \mu_a(X)$,

the efficiency of the ACW estimator can be further improved. The nuisance functions $\mu_a(X)$ can be estimated using the RWE sample. A6 is testable with data. Further discussion on this approach can be found in Yang et al. (2023) for a test-and-pool approach and Yang, Zeng, and Wang (2020) for an approach relying on estimating the confounding function associated with hidden confounders in observational studies, such as the NCDB sample in the motivating example, in which treatment selection is not randomly assigned but decided by the preference of surgeons and patients.

3.6. Extension to survival endpoints

Restricted mean survival time (RMST) $\mu(\tau)$ measures the averaged event time up to a pre-specified time horizon τ and is defined as the area under the survival curve up to τ, i.e.,

$$\mu(\tau) = E[\min(T, \tau)] = \int_0^\tau S(t)dt,$$

where T is the time to event and $S(t)$ is the survival function. As illustrated in Figure 3, RMST difference $\theta = \mu_1(\tau) - \mu_0(\tau)$ can be interpreted as the difference of event-free time (e.g., in months, years) between treatment group and control group up to τ. Unlike conventional estimand for treatment effect for survival endpoints, e.g., hazard ratio, RMST has valid interpretation when proportional hazards (PH) assumption has been violated. Violation of PH assumption is common in cancer clinical trials, as seen in Figure 1(a, b) for CALGB 9633, and cancer immunotherapies trials due to delayed treatment effect.

$S_a(t|x) = S(t|X, A = a, \delta = 1)$ is the treatment-specific conditional survival curves for a and $\delta \in \{0,1\}$. $\pi_A(X) = P(A = 1|X, \delta = 1)$ is the treatment propensity score. $\pi_\delta(X) = P(\delta = 1|X)$ is the sampling score. Estimands based on $S_a(t)$ are of interest: (1) survival rate difference at a fixed time τ: $\theta_\tau = S_1(\tau) - S_0(\tau)$; and (2) restricted mean survival time (RMST) difference over $[0, \tau]$

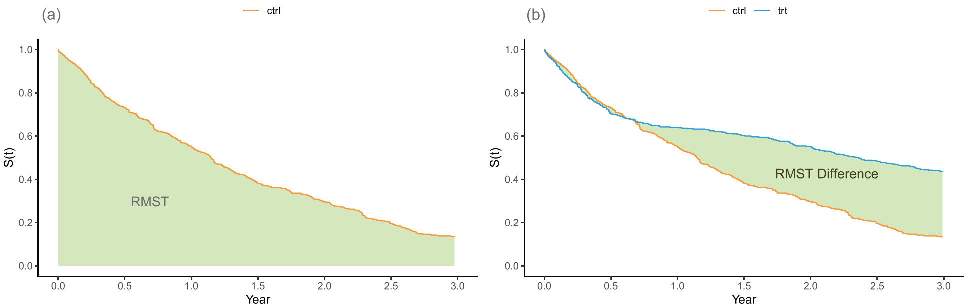


Figure 3. (a) Restricted mean survival time (RMST), and (b) RMST difference between treatment groups.

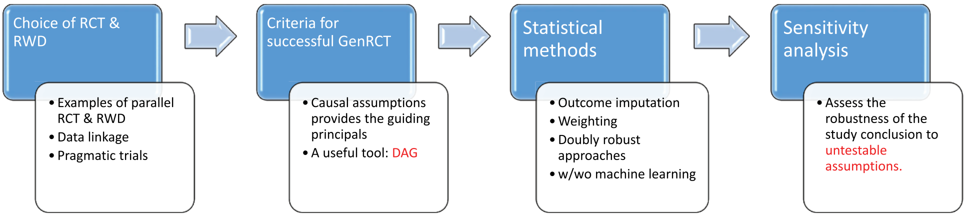

Choice of RCT & RWD
• Examples of parallel RCT & RWD
• Data linkage
• Pragmatic trials
Criteria for successful GenRCT
• Causal assumptions provides the guiding principals
• A useful tool: DAG
Statistical methods
• Outcome imputation
• Weighting
• Doubly robust approaches
• w/wo machine learning
Sensitivity analysis
• Assess the robustness of the study conclusion to untestable assumptions.

Figure 4. Workflow and key considerations for genRCT analysis.

$$\theta_\tau = \int_0^\tau \{S_1(t) - S_0(t)\}dt$$

$S_a(t)$ identification is achieved by noticing that $S_a(t) = E\{\tilde{\delta} dS_a(t|X)\}$, where

$$S_a(t|X) = E\{I(T \geq t)|X, A = a, \delta = 1\},$$

and

$$S_a(t) = E[\frac{\delta}{\pi_\delta(X)} \frac{I(A = a)}{\pi_A(X)^a\{1 - \pi_A(X)\}^{1-a}} \frac{Y(t)}{S^C(t|X, A)}],$$

where $S^C(t|X, A) = P(C > t|X, A, \delta = 1)$ is the conditional censoring model and C is the censoring time. The estimated calibration weights are obtained by $\widehat{w}_i = w(X_i; \widehat{\lambda}) = \exp\{\widehat{\lambda}^t g(X_i)\}/[\sum_{i=1}^n \exp\{\widehat{\lambda}^t g(X_i)\}]$, where $\widehat{\lambda}$ solves $U(\lambda) = \sum_{i=1}^n \exp\{\lambda^t g(X_i)\}\{g(X_i) - \tilde{g}\} = 0$. The estimated treatment propensity score is known for RCTs, but one can estimate it for better efficiency by $\pi_A(X) = [1 + \exp\{-\rho^t g(x)\}]^{-1}$, where ρ is the vector of regression coefficients of $g(x)$. We posit Cox PH model with conditional hazard $\lambda^C(t|X, A = a) = \lambda_{a0}^C(t)\exp(\gamma_a^t X)$ for $a \in \{0, 1\}$ to estimate γ_a and $\Lambda_{a0}^C = \int_0^t \lambda_{a0}^C(u)du$. Combining $\widehat{w}_i$, $\widehat{\pi}_{ai}$, and $\widehat{\Lambda}_{ai}^C(t)$, we have the CW estimator for the marginal treatment-specific survival curves as

$$\widehat{S}_a^{\mathrm{CW}}(t) = \sum_{i=1}^N \delta_i \widehat{w}_i \frac{A_{ai}}{\widehat{\pi}_{ai}} e^{\widehat{\Lambda_{ai}^C}(t)} Y_i(t)$$

Lee et al. (2022) improved the CW estimator following the semiparametric theory (Tsiatis 2006). Based on the efficient influence function, the improved ACW estimator is driven as

$$\widehat{S}_a^{\mathrm{ACW}}(t) = \exp\left\{-\int_0^t \frac{-\mathrm{d}\widehat{S}_a^{\mathrm{denom}}(u)}{\widehat{S}_a^{\mathrm{denom}}(u)}\right\},$$

where

$$\widehat{S}_a^{\mathrm{denom}}(t) = \sum_{i=1}^N \delta_i \widehat{\omega}_i \frac{A_{ai}}{\widehat{\pi}_{ai}} e^{\widehat{\Lambda_{ai}^C}(t)} Y_i(t)$$

$$+ \sum_{i=1}^N e^{-\widehat{\Lambda}_{ai}(t)}\left[\left(\sum_{i=1}^N \widetilde{\delta}_i d_i\right)^{-1} \widetilde{\delta}_i d_i - \delta_i \widehat{\omega}_i \frac{A_{ai}}{\widehat{\pi}_{ai}}\left\{1 - \int_0^t \left\{e^{\widehat{\Lambda_{ai}^C}(u) + \widehat{\Lambda}_{ai}(u)}\right\} \mathrm{d}\widehat{M}_{ai}^C(u)\right\}\right]$$

and

$$-\mathrm{d}\widehat{S}_a^{\mathrm{denom}}(u) = \sum_{i=1}^N \delta_i \widehat{\omega}_i \frac{A_{ai}}{\widehat{\pi}_{ai}} + e^{\widehat{\Lambda_{ai}^C}(u)} \mathrm{d}N_i(u)$$

$$+ \sum_{i=1}^N e^{-\widehat{\Lambda}_{ai}(u)} \mathrm{d}\widehat{\Lambda}_{ai}(u)\left[\left(\sum_{i=1}^N \widetilde{\delta}_i d_i\right)^{-1} \widetilde{\delta}_i d_i - \delta_i \widehat{\omega}_i \frac{A_{ai}}{\widehat{\pi}_{ai}}\left\{1 - \int_0^u \left\{e^{\widehat{\Lambda_{ai}^C}(s) + \widehat{\Lambda}_{ai}(s)}\right\} \mathrm{d}\widehat{M}_{ai}^C(s)\right\}\right],$$

where $M_a^C(u) = N_a^C(u) - \int_0^u \Lambda_a^C(s)ds$ is a martingale with $N_a^C = A_a I(U \leq u, \Delta = 0)$. The ACW estimator can be viewed as an augmentation of the survival model and the weighting model, which combines the sampling score model, the treatment propensity score model, and the censoring model.

The ACW estimator is consistent if either the survival model or the weighting model is correctly specified, and achieves local efficiency when both are correct. Note that even though $\widehat{S}_a^{\text{denom}}(t)$ itself is a survival estimator and is asymptotically equivalent to $\widehat{S}_a^{\text{ACW}}(t)$, the latter was found to show better finite-sample performance. Additionally, as an alternative to parametric estimation, Lee et al. (2022) employed the nonparametric method of sieves to add flexibility and robustness to the ACW estimator, meanwhile retaining the root-n consistency and efficiency.

3.7. *Finite sample performance for genRCT methods*

Extensive simulation studies were conducted to compare the finite sample performance of the genRCT estimators, including

Naive: difference in sample means of the two treatment groups in the RCT sample to demonstrate the degree of selection bias;

IPSW: inverse probability of sampling weighting estimator, where the sampling weights are estimated by logistic regression;

AIPSW: augmented inverse probability of sampling weighting estimator, where the sampling weights and the outcome mean are estimated by logistic regression;

CW: calibration weighting estimator with $g(X) = g_1(X)$;

ACW: augmented calibration weighting estimator with $g(X) = g_1(X)$ and the nuisance functions $\mu_1(X)$ and $\mu_0(X)$ are estimated based on the RWE sample;

ACW(S): penalized augmented calibration weighting estimator using the method of sieves with $g(X) = g_2(X)$.

We have used $g_1(X) = (X_1, X_2, X_3, X_4)^T$ in all four scenarios, and $g_2(X) = (X_1, \cdots, X_p, X_1X_2, \cdots, X_{p-1}X_p, X_1^2, \cdots, X_p^2)^T$ for ACW(S). Four scenarios were considered to evaluate the performance of these estimators.

Scenario 1 (O:C/S:C): both outcome and sampling score models are correctly specified;

Scenario 2 (O:C/S:W): the outcome model is correctly specified; the sampling score model is incorrectly specified;

Scenario 3 (O:W/S:C): the outcome model is incorrectly specified; the sampling score model is correctly specified;

Scenario 4 (O:W/S:W): both outcome model and sampling score models are incorrectly specified.

For details, please refer to Lee et al. (2022, 2023). Overall, IPSW and AIPSW have wider variation compared to their CW and ACW counterparts. When both outcome and sampling score models are misspecified, the ACW(S) estimator is still unbiased and efficient. The empirical coverage rates for the unbiased ACW estimators are close to the nominal level. Moreover, ACW-b(S) has smaller variance than ACW-t(S) by exploiting the predictive power from the observational sample.

4. Workflow and key considerations for genRCT analysis

Figure 4 displays the major steps to conduct a comprehensive genRCT analysis. These steps will help verify the assumptions required for valid and efficient genRCT analysis.

4.1. Identify the need for generalizability

One way to gauge the need for generalizability is by examining the consistency in the distribution of essential baseline risk factors. As evident from Table1 and Table A1 in Appendix, which provide a snapshot of the motivating example, there are evident inconsistencies in the distribution of some important factors. This becomes a preliminary condition, warranting the exploration of a generalizable randomized controlled trial (genRCT) analysis.

A systematic method to ascertain the necessity for such an analysis is by evaluating the standardized differences of key baseline covariates between two distinct data sources. If these standardized differences are closely aligned or are relatively similar, it negates the requirement for a genRCT analysis. This suggests that the baseline factors are balanced, thereby making generalization less critical in this context.

However, if the need arises to integrate results from RCTs and RWE, particularly to achieve heightened efficiency or for other objectives, it is essential to harness the specialized methodologies tailored for this integration. Such integrative analysis offers a broader perspective and ensures the resultant findings reflect the diverse patient populations, thus enhancing the external validity of the study findings.

4.2. Generate comparable patient populations, i.e., RCT sample and RWE-target

In the genRCT analysis of CALGB 9633 and NCDB, we have chosen comparable patients from the NCDB database by defining the target patient population using the eligibility criteria of CALGB 9633. In contrast, the real NCDB patients have more diverse constituents, such as early-stage NSCLC with a tumor size of less than 3 cm. Even though the distribution of the two data sources is comparable, the distribution of the two data sources may still be different. One key observation is that RCTs, such as CALGB 9633, often have a propensity to enroll patients who are generally younger and have smaller tumor size. Furthermore, these trials might showcase a demographic tilt, leaning towards younger and white participants. Various factors can contribute to this bias, such as the stringent eligibility criteria for RCTs, socioeconomic determinants, or institutional preferences.

In general, comparing an RCT to an external dataset aims to bridge the gap between controlled experimental conditions and diverse, real-world patient scenarios. External data sources can be diverse and encompass population-based disease registries or comprehensive national health studies. Such comparisons not only enhance the generalizability of RCT findings but also provide a more nuanced understanding of how treatment works in everyday clinical practice.

Additionally, it is worth highlighting the unique position of pragmatic clinical trials conducted within Electronic Health Record (EHR) systems. For such trials, the population captured in the EHR system inherently represents the real-world population. This setup offers a more organic, day-to-day representation of patients, emphasizing the practicality and applicability of the data within real-world medical settings.

4.3. Selection and calibration of variables

For the genRCT analysis, selecting variables for calibration is important. Our primary objective in this selection process is to ensure a consistent and efficient analysis. The variables used for calibration have been identified based on specific criteria.

Firstly, confounders play a crucial role. A confounder is a variable related to the treatment effect and selection into the RCT study. The presence of confounders can introduce bias, as they might inadvertently suggest a relationship between the treatment effect and the selection. Hence, when identifying variables for calibration, it's critical to consider these confounders to enhance the validity of our analysis. Secondly, it is essential to incorporate all prognostic factors. These factors have a direct correlation with the outcomes under consideration. Including these ensures that our analysis is efficient (Cho and Yang 2023). Thirdly, there is a possibility of the presence of instrumental variables that are associated solely with the selection and not the outcomes. Incorporating these instrumental variables does not compromise the consistency of the estimator, though it may increase the variance. Nonetheless, as indicated by Yang, Kim, and Song (2020), including these instrumental variables reduces the likelihood of omitting potential confounders. Finally, there might be scenarios where certain covariates, suspected to be confounders, are not available in any of the data sources at our disposal. In such situations, it is advisable to identify surrogate variables for these unobserved confounders.

Once the covariates are determined, the subsequent step ensures the distributional balance between the RCT and RWE studies. Toward this end, one can include first-order, second-order, and even higher-order moments of these covariates for calibration. Such an approach strengthens the overall validity of the genRCT analysis. However, as the number of calibration constraints increases, over-calibration or improper application of calibration weighting on too many variables, can lead to variance inflation. In such cases, one can use regularization (Tan 2020) for selecting important calibration constraints or soft calibration (Gao et al. 2022) for relaxing the constraints.

4.4. Selection of methods for the primary analysis

To determine the most suitable method for the primary analysis, a comprehensive review was undertaken based on both theoretical foundations and extensive simulation studies. In situations where the RWE study provides only the covariate data, we recommend using ACW(s)-t as the primary analysis method. The rationale behind this recommendation is the double robustness feature of the ACW(s)-t method and its ability to provide stable weighting, making it especially reliable in such cases.

On the other hand, when the RWE study provides a more comprehensive dataset including covariates, treatment, and outcome data, our recommendation shifts to the ACW(s)-b method. ACW(s)-b is designed to utilize the outcome mean information in the RWE study, ensuring a more informed and accurate analysis in scenarios where complete data is accessible.

4.5. Sensitivity analysis

The success of genRCT relies on the key assumption, labeled A3. This key assumption, however, cannot be verified directly using the data available. As a consequence, the robustness of the study's conclusions may be vulnerable to any deviations from this assumption (e.g. due to missing important confounders), making a sensitivity analysis critical. Sensitivity analysis serves as an essential tool to assess how much the conclusions might change under potential violations of this assumption.

In recent times, there has been a growing recognition of the challenges posed by missing confounders, both within the RCT and RWE studies. This has culminated in the proposal of various sensitivity analyses aimed at counteracting the impact of unmeasured confounders in both of these types of data. Notable contributions in this area have been made by Nguyen et al. (2017), Nie et al. (2021), Huang (2022) and many others. These analyses provide valuable frameworks to evaluate the robustness of conclusions derived from datasets where key confounders might be missing.

5. Software

The *R* package for conducting genRCT analysis is available for download from the *github* and installed in the local *R* environment by typing

$$devtools :: install\_github("idasomm/genRCT")$$

at *R* command prompt. The function *genRCT* is for the analysis of binary and continuous endpoints. Its arguments can be found on the package help page and are listed below.

$Y.trial$ Observed outcome from a trial; vector of size n. (trial sample size)

$A.trial$ Treatment received from a trial; vector of size n.

$X.trial$ Matrix of p baseline covariates from a trial; dimension n by p.

$Y.rwe$ Observed outcome from RWE; if obtained, vector of size m (RWE sample size); otherwise, set $Y.rwe = NULL$.

$A.rwe$ Treatment received from RWE; if obtained, vector of size m; otherwise, set $A.rwe = NULL$.

$X.rwe$ Matrix of p baseline covariates from RWE; dimension m by p.

family The type of outcome, "gaussian" for Gaussian regression or "binomial" for logistic regression Default is "gaussian".

estimators A vector of one or multiple methods to estimate the ATE. Allowed values are "Naive", "IPSW", "AIPSW", "CW", "ACW-t", "ACW-b". The "ACW-b" is allowed only when both "Y.rwe" and "A.rwe" are obtained. Default specifies all six methods.

sieve A logical value indicates whether the method of sieves is used for estimating sampling score and outcome models. Used only if estimators = "AIPSW" or "ACW-t" or "ACW-b". The default is *TRUE*.

inference A logical value indicating whether inference for the ATE via bootstrap should be provided. The default is TRUE.

$n.boot$ A numeric value indicating the number of bootstrap samples used. This is only relevant if $inference = TRUE$. The default is 100.

$conf.level$ The level of bootstrap confidence interval; Default is 0.05.

seed An optional integer specifying an initial randomization seed for reproducibility. The default is *NULL*, corresponding to no seed.

$plot.boot$ A logical value indicating whether histograms of the bootstrap samples should be produced. The default is *TRUE*.

verbose A logical value indicating whether intermediate progress messages should be printed. The default is *TRUE*.

The outputs of the *R* function are

fit A table of estimated ATEs with bootstrap SE and confidence interval.

plot A set of histograms displaying the distribution of the bootstrapped estimates. The red vertical reference lines represent the estimated ATEs from each method.

The *R* function $genRCT.surv$ is for conducting the genRCT analysis for survival endpoints subject to right censoring. The treatment effect is characterized as survival rates and its difference at landmark follow-up time or RMST and its difference up to a time horizon τ. The arguments of $genRCT.surv$ can be found on the package help page and are listed below. At this moment, $genRCT.surv$ does not support incorporating the outcome and treatment data from RWE into the generalizability analysis.

$Y.trial$ Observed outcome from a trial; vector of size n. (trial sample size)

$d.trial$ The event indicator, normally $1 = event$, $0 = censored$; vector of size n.

A.trial Treatment received from a trial; vector of size n.

X.trial Matrix of p baseline covariates from a trial; dimension n by p.

X.rwe Matrix of p baseline covariates from RWE; dimension m by p.

tau A vector of truncation time for defining restricted mean survival time; e.g., seq (10, 50, by = 10)

n.boot A numeric value indicating the number of bootstrap samples used. This is only relevant if $inference = TRUE$. The default is 100.

conf.level The level of bootstrap confidence interval; Default is 0.05.

seed An optional integer specifying an initial randomization seed for reproducibility. The default is *NULL*, corresponding to no seed.

verbose A logical value indicating whether intermediate progress messages should be printed. The default is *TRUE*.

The outputs of the *R* function are

rmst A list of estimated RMSTs with bootstrap SE and confidence interval.

surv A list of estimated treatment-specific survival functions.

6. Case study of genRCT analysis

6.1. genRCT analysis with binary endpoint

The *R* scripts for this analysis can be found in the supplementary material. The following variables are available for the genRCT analysis:

recurrence Y: 1=overall survival time <3 years, 0=otherwise

arm=A: treatment indicator with 1 for Chemotherapy and 0 for Observation

male=X_1: 1 for male and 0 for female

age=X_2: age at randomization

squam=X_3: histology with 1=squamous, 0=non-squamous

tsize=X_4: tumor size measured at baseline

cohort: trial="CALGB 9633", rwe="NCDB"

We applied the following estimators to the pooled data of CALGB 9633 and NCDB samples.

- Naive: difference in sample means of the two treatment groups in the RCT sample to demonstrate the degree of selection bias;
- IPSW: inverse probability of sampling weighting estimator, where the sampling weights are estimated by logistic regression;
- AIPSW(S): augmented IPSW estimator with sieves method for sampling score and outcome models;
- CW: calibration weighting estimator with $g(X) = g_1(X)$;
- ACW-t(S): augmented CW estimator with sieves method for sampling score and outcome models, and the outcome and treatment data from RWE are not used;

- ACW-b(S): augmented CW estimator with sieves method for sampling score and outcome models, and the outcome and treatment data from both RCT and RWE are used. See Lee et al. (2023) for more discussion on ACW-b(S) and its difference from ACW-t(S).

Y is the indicator of cancer recurrence within 3 years after the surgery, i.e. $Y = 1$ if recurrence occurred and $Y = 0$ otherwise. As a few patients dropped out or withdrew consent for being followed up before 3 years, they were excluded from this analysis due to indefinite recurrence status. Figure 5 is a Love plot showing the standardized difference of key baseline covariates before and after balancing the covariates of the RCT sample against the RWE-target. Figure 6 shows the averaged risk difference between adjuvant chemotherapy and observation based on CALGB 9633 and NCDB. ACW-t(S) denotes the ACW estimator with sieves method using the outcome and treatment data only from the trial CALGB 9633, and it suggests a marginal benefit of adjuvant chemotherapy in risk reduction by 12% with the upper 95%CI barely exceeding 0. ACW-b(S) denotes the ACW estimator with sieves method using the outcome and treatment data from both CALGB 9633 and NCDB samples, and it shows a 17% risk decrease ($p < 0.05$), supporting a more profound benefit of adjuvant chemotherapy over observation in the real-world population.

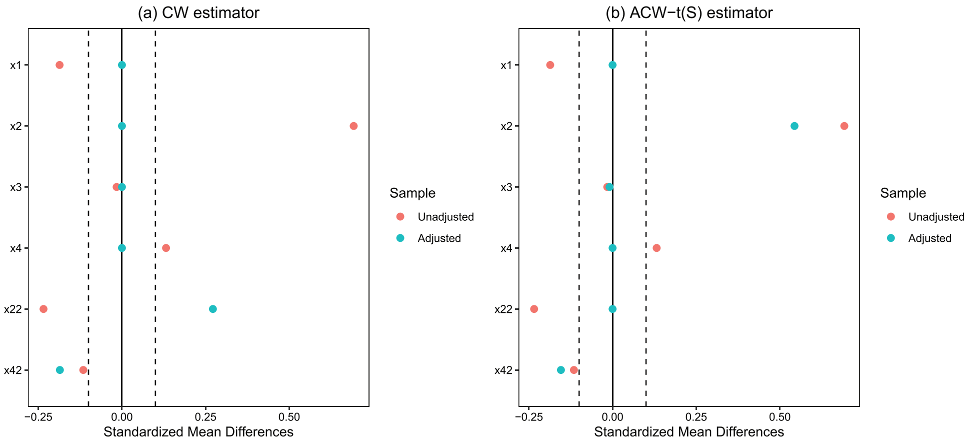


Figure 5. Love plot before and after balancing the covariates of CALGB 9633 sample against the NCDB sample using (a) CW and (b) ACW-t(S) estimators.

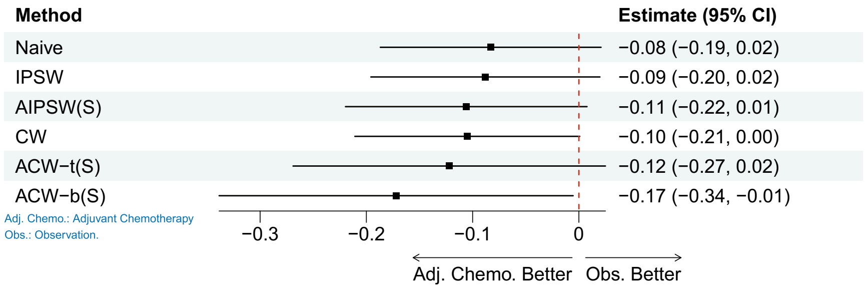


Figure 6. Case study of CALGB 9633 and NCDB samples with recurrence within 3 years (yes or no) as binary outcome.

6.2. genRCT analysis with survival endpoint

The R scripts for this analysis can be found in the supplementary material. The following variables are available for the genRCT analysis with overall survival (OS) as the primary outcome.

survtime: overall survival (OS)

survcens: OS censoring indicator, 1 for “death” and 0 for “censored”

arm=A: treatment indicator with 1 for Chemotherapy and 0 for Observation

male=X_1: 1 for male and 0 for female

age=X_2: age at randomization

squam=X_3: histology with 1=squamous, 0=non-squamous

tsize=X_4: tumor size measured at baseline

cohort: trial=“CALGB 9633”, rwe=“NCDB”

The estimates and 95% CIs of the RMST difference with $\tau = 5$ and $\tau = 10$ years between adjuvant chemotherapy and observation are given in Figure 7, supporting a strong benefit of adjuvant chemotherapy in the RWE target population, as defined by the NCDB sample, relative to observation by generalizing the treatment effect found in CALGB 9633.

7. Concluding remarks

In this article, we introduce a framework for conducting the genRCT analysis that generalizes the findings of a randomized clinical trial from the RCT population to its corresponding RWD population. This analysis offers a means to assess the disparity in treatment effects between populations in RCTs and RWD. This analysis becomes particularly relevant when treatment effects are not readily deducible from the Real-World Evidence (RWE) database. By applying the findings from RCTs to the RWE population, the generalizability analysis estimates the treatment effect for the RWE population. The discrepancy between the treatment effects observed in RCTs and those calculated through genRCT analysis serves as an indicator of the gap between the two populations. This difference is valuable for both the pharmaceutical industry and regulatory agencies, as it aids in evaluating the effectiveness of specific drugs or treatments during the development of pharmaceutical products. The

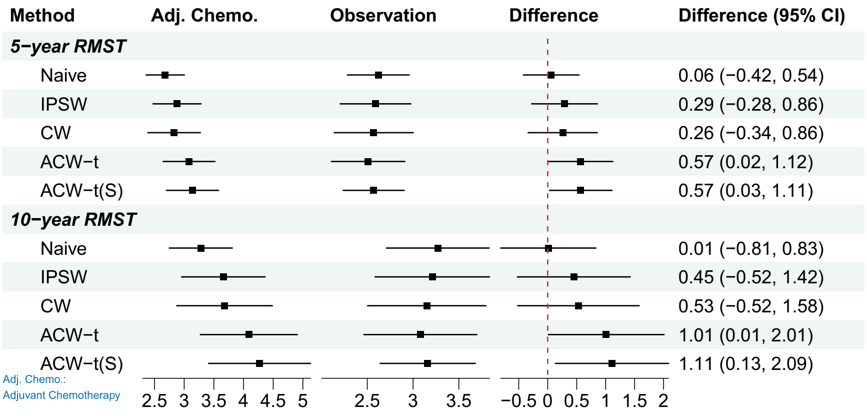


Figure 7. RMST difference as treatment effect for the case study of CALGB 9633 and NCDB samples with time horizon at 5 and 10 years.

genRCT framework leverages both the RCT (e.g., CALGB 9633) and the observational studies (e.g., NCDB), to estimate the average treatment effect (e.g., adjuvant chemotherapy on average survival times). Among a few competitors, the ACW estimator has been demonstrated to be doubly robust, surpassing the efficiency of both IPSW and CW estimators. Even with misspecified models, the ACW(S) estimator remains unbiased and efficient, with empirical coverage rates aligning closely with nominal values. This framework caters to continuous, binary, and survival outcomes. Our accompanying *R* package, "genRCT", facilitates straightforward implementation. We provide a comprehensive guideline for achieving generalizable treatment effect analyses pertinent to real-world patient demographics, detailing necessary criteria for data representation, variable selection, and estimator preferences.

To conclude the article, we would like to make a few remarks regarding the use of the genRCT analysis in practice. A specific quantitative metric for determining the need for a genRCT analysis does not exist. However, it's advisable to evaluate the differences in covariates between RCT and RWE groups. Conducting a genRCT analysis becomes necessary if the standardized difference in at least one covariate, identified as a moderate or strong confounder by subject matter experts, is observed. Moreover, even if there's pre-existing knowledge suggesting similarities between RCT and RWE populations, a genRCT analysis may still be beneficial. Consequently, it's recommended to perform a genRCT analysis on all completed RCTs, regardless of any prior understanding of potential discrepancies between the RCT and RWE cohorts. Drawing from semiparametric efficiency theory and the positive outcomes seen in extensive simulation studies, the ACW(S) estimator provides both consistency, efficiency, and double robustness. Furthermore, the associated variance estimator provides a close-to-nominal coverage rate for confidence intervals. Given its good performance relative to other estimators, its application in practical scenarios is highly recommended. Although the ACW(S) estimator outperforms others in certain respects, employing alternative estimators for comparative analysis remains beneficial. Such comparisons not only reinforce the reliability of ACW(S) results but also when significant discrepancies are noted, prompt further investigation into the underlying causes of these differences. The performance of various estimators, in terms of consistency and efficiency, hinges on different assumptions and the correctness of either the sampling model or the outcome model. Consequently, it is common in practice to observe numerical and quantitative differences among these methods, particularly when the treatment effect is only marginally significant. In such instances, it is crucial for researchers to delve into the possible causes behind these variances. However, overall, the results derived from the ACW(S) estimator should be considered more reliable than those from alternative estimators.

Acknowledgements

Dr. Wang was partially supported by NIA grant R01AG066883. Dr. Yang is partially supported by NSF grant SES2242776, NIA grant R01AG066883 and NIEHS grant R01ES031651. Drs. Berry, Stinchcombe, and Cohen are partially supported by NIA grant R01AG066883.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Boffa, D. J., J. E. Rosen, K. Mallin, A. Loomis, G. Gay, B. Palis, K. Thoburn, D. Gress, D. P. McKellar, L. N. Shulman, et al. 2017. Using the national cancer database for outcomes research: A review. *JAMA Oncology* 3 (12):1722–1728. doi:10.1001/jamaoncol.2016.6905.

Buchanan, A. L., M. G. Hudgens, S. R. Cole, K. R. Mollan, P. E. Sax, E. S. Daar, A. A. Adimora, J. J. Eron, and M. J. Mugavero. 2018. Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society Series A: Statistics in Society* 181 (4):1193–1209. doi:10.1111/rssa.12357.

Chen, X. 2007. Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics* 6:5549–5632.

Cho, E., and S. Yang. 2023. Variable selection for doubly robust causal inference. *arXiv preprint arXiv:2301.11094*.

Cole, S. R., and E. A. Stuart. 2010. Generalizing evidence from randomized clinical trials to target populations: The actg 320 trial. *American Journal of Epidemiology* 172 (1):107–115. doi:10.1093/aje/kwq084.

Colnet, B., J. Josse, G. Varoquaux, and E. Scornet. 2022. Causal effect on a target population: A sensitivity analysis to handle missing covariates. *Journal of Causal Inference* 10 (1):372–414. doi:10.1515/jci-2021-0059.

Colnet, B., I. Mayer, G. Chen, A. Dieng, R. Li, G. Varoquaux, J.-P. Vert, J. Josse, and S. Yang. 2020. Causal inference methods for combining randomized trials and observational studies: A review. *arXiv preprint arXiv:2011.08047*.

Dahabreh, I. J., and M. A. Hernán. 2019. Extending inferences from a randomized trial to a target population. *European Journal of Epidemiology* 34 (8):719–722. doi:10.1007/s10654-019-00533-2.

Dahabreh, I. J., S. E. Robertson, E. J. Tchetgen, E. A. Stuart, and M. A. Hernán. 2019. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics Bulletin* 75 (2):685–694. doi:10.1111/biom.13009.

Ettinger, D. S., D. L. Aisner, D. E. Wood, W. Akerley, J. Bauman, J. Y. Chang, L. R. Chirieac, T. A. D'Amico, T. J. Dilling, M. Dobelbower, et al. 2018. Nccn guidelines insights: non–small cell lung cancer, version 5.2018. *Journal of the National Comprehensive Cancer Network* 16(7):807–821. doi:10.6004/jnccn.2018.0062.

Gao, C., S. Yang, and J. K. Kim. 2022. Soft calibration for selection bias problems under mixed-effects models. *Biometrika* 110 (4):897–911. doi:10.1093/biomet/asad016.

Hainmueller, J. 2012. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* 20 (1):25–46. doi:10.1093/pan/mpr025.

Hartman, E., R. Grieve, R. Ramsahai, and J. S. Sekhon. 2015. From sample average treatment effect to population average treatment effect on the treated: Combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society Series A: Statistics in Society* 178 (3):757–778. doi:10.1111/rssa.12094.

Huang, M. 2022. Sensitivity analysis in the generalization of experimental results. *arXiv preprint arXiv:2202.03408*.

Imbens, G. W., and D. B. Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge, UK: Cambridge University Press.

Katz, A., and E. D. Saad. 2009. Calgb 9633: An underpowered trial with a methodologically questionable conclusion. *Journal of Clinical Oncology* 27 (13):2300–2301. doi:10.1200/JCO.2008.21.1565.

Kennedy-Martin, T., S. Curtis, D. Faries, S. Robinson, and J. Johnston. 2015. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials* 16 (1):1–14. doi:10.1186/s13063-015-1023-4.

Korn, E. L., and B. Freidlin. 2012. Methodology for comparative effectiveness research: Potential and limitations. *Journal of Clinical Oncology* 30 (34):4185–4187. doi:10.1200/JCO.2012.44.8233.

Lee, D., S. Yang, L. Dong, X. Wang, D. Zeng, and J. Cai. 2023. Improving trial generalizability using observational studies. *Biometrics Bulletin* 79 (2):1213–1225. doi:10.1111/biom.13609.

Lee, D., S. Yang, and X. Wang. 2022. Doubly robust estimators for generalizing treatment effects on survival outcomes from randomized controlled trials to a target population. *Journal of Causal Inference* 10 (1):415–440. doi:10.1515/jci-2022-0004.

Morgensztern, D., L. Du, S. N. Waqar, A. Patel, P. Samson, S. Devarakonda, F. Gao, C. G. Robinson, J. Bradley, M. Baggstrom, et al. 2016. Adjuvant chemotherapy for patients with t2n0m0 nsclc. *Journal of Thoracic Oncology* 11 (10):1729–1735. doi:10.1016/j.jtho.2016.05.022.

Nguyen, T. Q., C. Ebnesajjad, S. R. Cole, and E. A. Stuart. 2017. Sensitivity analysis for an unobserved moderator in rct-to-target-population generalization of treatment effects. *The Annals of Applied Statistics* 11 (1):225–247. doi:10.1214/16-AOAS1001.

Nie, X., G. Imbens, and S. Wager. 2021. Covariate balancing sensitivity analysis for extrapolating randomized trials across locations. *arXiv preprint arXiv:2112.04723*.

Pang, H. H., X. Wang, T. E. Stinchcombe, M. L. Wong, P. Cheng, A. K. Ganti, D. J. Sargent, Y. Zhang, C. Hu, S. J. Mandrekar, et al. 2016. Enrollment trends and disparity among patients with lung cancer in national clinical trials, 1990 to 2012. *Journal of Clinical Oncology* 34(33):3992. doi:10.1200/JCO.2016.67.7088.

Pearl, J., and E. Bareinboim (2011). Transportability of causal and statistical relations: A formal approach. *Proceedings of the AAAI Conference on Artificial Intelligence* 25:247–254.

Qin, J., and B. Zhang. 2007. Empirical-likelihood-based inference in missing response problems and its application in observational studies. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 69 (1):101–122. doi:10.1111/j.1467-9868.2007.00579.x.

Rothwell, P. M. 2005. External validity of randomised controlled trials:"to whom do the results of this trial apply?". *The Lancet* 365 (9453):82–93. doi:10.1016/S0140-6736(04)17670-8.

Strauss, G. M., J. E. Herndon, M. A. Maddaus, D. W. Johnstone, E. A. Johnson, D. H. Harpole, H. H. Gillenwater, D. M. Watson, D. J. Sugarbaker, R. L. Schilsky, et al. 2008. Adjuvant paclitaxel plus carboplatin compared with observation in stage ib non–small-cell lung cancer: Calgb 9633 with the cancer and leukemia group b, radiation therapy oncology group, and north central cancer treatment group study groups. *Journal of Clinical Oncology* 26 (31):5043. doi:10.1200/JCO.2008.16.4855.

Tan, Z. 2020. Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *The Annals of Statistics* 48 (2). doi:10.1214/19-AOS1824.

Tang, D., D. Kong, W. Pan, and L. Wang. 2020. Ultra-high dimensional variable selection for doubly robust causal inference. *arXiv preprint arXiv:2007.14190*.

Tipton, E. 2013. Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics* 38 (3):239–266. doi:10.3102/1076998612441947.

Tsiatis, A. A. 2006. *Semiparametric theory and missing data*. New York: Springer.

Visvanathan, K., L. A. Levit, D. Raghavan, C. A. Hudis, S. Wong, A. Dueck, and G. H. Lyman. 2017. Untapped potential of observational research to inform clinical decision making: American society of clinical oncology research statement. *Journal of Clinical Oncology* 35 (16):1845–1854. doi:10.1200/JCO.2017.72.6414.

Wang, X., F. Bai, H. Pang, and S. L. George. 2019. Bias-adjusted kaplan–meier survival curves for marginal treatment effect in observational studies. *Journal of Biopharmaceutical Statistics* 29 (4):592–605. doi:10.1080/10543406.2019.1633659.

Westreich, D., J. K. Edwards, C. R. Lesko, E. Stuart, and S. R. Cole. 2017. Transportability of trial results using inverse odds of sampling weights. *American Journal of Epidemiology* 186 (8):1010–1014. doi:10.1093/aje/kwx164.

Wu, C., and R. R. Sitter. 2001. A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association* 96 (453):185–193. doi:10.1198/016214501750333054.

Yang, S., C. Gao, D. Zeng, and X. Wang. 2023. "Elastic integrative analysis of randomized trial and real-world data for treatment heterogeneity estimation." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 85 (3): 575–596.

Yang, S., J. K. Kim, and R. Song. 2020. Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82 (2):445–465. doi:10.1111/rssb.12354.

Yang, S., and X. Wang. 2022. Rwd-integrated randomized clinical trial analysis. *Biopharmaceutical Report* 29 (2):15. doi:10.1111/biom.13609.

Yang, S., D. Zeng, and X. Wang. 2020. Improved inference for heterogeneous treatment effects using real-world data subject to hidden confounding. *arXiv preprint arXiv:2007.12922*.

Appendix

Data summary for CALGB 9633 and NCDB samples

Table A1 lists all key baseline characteristics of CALGB 9633 patients. A few patients in NCDB are removed from the genRCT analysis due to missing data.

Table A1. Baseline patient characteristics of CALGB 9633 and NCDB samples.

	CALGB 9633		NCDB	
	Observation (*N* = 166)	Adjuvant Chemo (*N* = 161)	Observation (*N* = 11544)	Adjuvant Chemo (*N* = 4468)
Sex				
Female	61 (36.7%)	57 (35.4%)	5219 (45.2%)	2040 (45.7%)
Male	105 (63.3%)	104 (64.6%)	6325 (54.8%)	2428 (54.3%)
Age				
Mean (SD)	61.0 (9.24)	60.3 (10.4)	69.3 (10.2)	63.9 (9.29)
Median [Min, Max]	62.0 [40.0, 81.0]	61.0 [34.0, 78.0]	70.0 [20.0, 90.0]	65.0 [29.0, 88.0]
Histology				
Other	101 (60.8%)	96 (59.6%)	6902 (59.8%)	2872 (64.3%)
Adenocarcinoma	65 (39.2%)	65 (40.4%)	4642 (40.2%)	1596 (35.7%)
Tumor Size				
Mean (SD)	4.56 (2.06)	4.63 (2.07)	4.67 (1.65)	5.19 (1.78)
Median [Min, Max]	4.00 [1.00, 12.0]	4.00 [1.00, 12.0]	4.20 [3.10, 25.0]	4.80 [3.10, 21.0]

R scripts for genRCT analysis for binary endpoint
R scripts for genRCT analysis for survival endpoint