

# Transporting survival of an HIV clinical trial to the external target populations

Dasom Lee, Chenyin Gao, Sujit Ghosh & Shu Yang

To cite this article: Dasom Lee, Chenyin Gao, Sujit Ghosh & Shu Yang (23 Mar 2024): Transporting survival of an HIV clinical trial to the external target populations, Journal of Biopharmaceutical Statistics, DOI: [10.1080/10543406.2024.2330216](https://doi.org/10.1080/10543406.2024.2330216)

To link to this article: <https://doi.org/10.1080/10543406.2024.2330216>



Published online: 23 Mar 2024.



Submit your article to this journal 



Article views: 60



View related articles 



View Crossmark data 



# Transporting survival of an HIV clinical trial to the external target populations

Dasom Lee, Chenyin Gao, Sujit Ghosh, and Shu Yang

Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA

## ABSTRACT

Due to the heterogeneity of the randomized controlled trial (RCT) and external target populations, the estimated treatment effect from the RCT is not directly applicable to the target population. For example, the patient characteristics of the ACTG 175 HIV trial are significantly different from that of the three external target populations of interest: US early-stage HIV patients, Thailand HIV patients, and southern Ethiopia HIV patients. This paper considers several methods to transport the treatment effect from the ACTG 175 HIV trial to the target populations beyond the trial population. Most transport methods focus on continuous and binary outcomes; on the contrary, we derive and discuss several transport methods for survival outcomes: an outcome regression method based on a Cox proportional hazard (PH) model, an inverse probability weighting method based on the models for treatment assignment, sampling score, and censoring, and a doubly robust method that combines both methods, called the augmented calibration weighting (ACW) method. However, as the PH assumption was found to be incorrect for the ACTG 175 trial, the methods that depend on the PH assumption may lead to the biased quantification of the treatment effect. To account for the violation of the PH assumption, we extend the ACW method with the linear spline-based hazard regression model that does not require the PH assumption. Applying the aforementioned methods for transportability, we explore the effect of PH assumption, or the violation thereof, on transporting the survival results from the ACTG 175 trial to various external populations.

## ARTICLE HISTORY

Received 4 February 2024

Accepted 20 February 2024

## KEYWORDS

Data integration;  
transportability; hazard  
regression; HIV trials

## 1. Introduction

In biomedical research, the distribution of baseline covariates in the randomized controlled trials (RCTs) sample, which is drawn from the *trial population*, is often different from that of the *target population*, the population that we want to make inferences about. Consequently, the estimated average treatment effect in the trial is not directly applicable to the target population beyond the trial population. As the findings from the RCTs often suffer from a lack of external validity (Rothwell 2005), observational studies that include large samples that are representative of the target population have been widely used as a complement to the RCTs.

Several recent works have proposed methods of extending RCT findings to the target population by balancing the distribution of the baseline covariates between the RCT and the observational data. A popular approach for covariate balancing is through direct modeling of the probability of participating in the trial, known as inverse probability of sampling weighting (IPSW) (Cole and Stuart 2010; Stuart et al. 2011; Westreich et al. 2017). However, the IPSW approach is unstable under extreme sampling scores. Alternatively, the entropy balancing or calibration weighting approach has been used to enforce the covariate balance between

the RCT and observational study without explicit modeling the sampling scores (Josey et al. 2021; Lee et al. 2021, 2022). There have been several terms for describing the process of extending the trial findings. The problem of transporting the results to the external target population beyond the trial has been termed *transportability* (Dahabreh and Hernán 2019; Josey et al. 2021; Pearl and Bareinboim 2011; Rudolph and van der Laan 2017; Westreich et al. 2017). Here the external target population is defined as the population that consists of at least some trial-ineligible individuals (Dahabreh et al. 2020). A similar problem that aims to generalize the findings from the RCT to its larger population has been termed *generalizability* (Cole and Stuart 2010; Dahabreh et al. 2019; Lee et al. 2021; Tipton 2013). Both the generalizability and transportability try to balance the distribution of the baseline covariates between the RCT and the observational data. The subtle differences between transportability and generalizability have been discussed in Lee et al (Lee et al. 2021). and Colnet et al (Colnet et al. 2020).

In this paper, we are interested in transporting the treatment effect for survival in the AIDS Clinical Trials Group (ACTG) 175 trial of intermediate-stage disease patients in the US. We consider three external target populations – US early-stage HIV patients, Thailand HIV patients, and southern Ethiopia HIV patients – whose patient characteristics differ significantly from that of the ACTG 175 trial. As a result, the treatment effect in the ACTG 175 HIV trial and that in the target populations are likely to be different. To evaluate the treatment effect in the target populations, we consider several methods for transportability. Specifically, we consider two intuitive approaches, the outcome regression approach for survival outcomes that is similar to the one proposed by Chen and Tsiatis (2001) under the PH assumption and the inverse probability weighting (IPW) approach that involves treatment propensity score, calibration weighting, and censoring model. We also consider a semiparametric efficient estimator, called the augmented calibration weighting (ACW) estimator, proposed by Lee et al. (2022). The ACW estimator combines the two intuitive approaches under a semiparametric theory (Tsiatis 2006). It accounts for the heterogeneity between the RCT and the observational study for a broad class of survival estimands that are functionals of treatment-specific survival functions, such as a difference in survival probabilities and restricted mean survival times. The ACW estimator is doubly robust in the sense that it is a consistent estimator if the Cox PH model for the survival outcome is correctly specified or the weighting models are correctly specified, and is the most efficient estimator when all models are correctly specified.

However, the proportionality assumption on which the ACW estimator depends may be questionable in many clinical trials in practice, particularly when there is a delayed effect or a crossing of the survival functions. Sheng and Ghosh (2020) explored various PH and non-PH models, including the linear spline-based hazard regression (HARE) model. They found that the PH assumption may be violated in the ACTG 175 trial data and showed that the violation of the PH assumption results in the biased variable section for the PH models but not for the non-PH models. Consequently, the methods that depend on the PH assumption may lead to the biased quantification of the treatment effect in the target populations. To account for this possible violation of the PH assumption, we extend the ACW estimator with the HARE model that does not require the PH assumption. This extension enhances the robustness of the ACW estimator compared to the original counterpart, as shown in our simulation studies. We apply the aforementioned methods for transportability to explore the effect of the PH assumption, or the violation thereof, on transporting the treatment effect from the ACTG 175 trial to the target populations.

The remainder of the paper is organized as follows. Section 2 provides the details of the motivating example, the ACTG-175 trial and three external datasets. Section 3 formalizes the basic causal inference framework and discusses identifiability conditions for transportability. Section 4 presents an overview of PH and non-PH models for transportability. Section 5 presents a set of simulations demonstrating the robustness of the ACW estimator with the HARE model against violations of the PH assumption. In Section 6, we present the result of transporting the treatment effect to three target populations. In Section 7, we provide discussion and concluding remarks.



## 2. A motivating example: ACTG-175 trial and external data

To explore the effect of PH assumption on transportability problem, we consider an HIV clinical trial. The ACTG 175 trial enrolled HIV-infected patients with 200–500 cells/mm<sup>3</sup> CD4 cell counts who were randomized to four antiretroviral therapies: Zidovudine monotherapy (ZDV), Zidovudine plus Didanosine (ZDV + ddI), Zidovudine plus Zalcitabine (ZDV + ZAL), or Didanosine monotherapy (ddI) (Hammer et al. 1996). For illustration purposes, we choose ZDV + ddI and ZDV as binary treatment, which consists of 522 ZDV + ddI patients and 532 ZDV monotherapy patients. The analyses for the effect of two other treatments, ZDV + ZAL and ddI, over ZDV, are provided in [Appendix B](#). The primary endpoint of the study is the progression of HIV disease, defined as a more than 50% decline in the CD4 cell count or development of the acquired immunodeficiency syndrome, or death. The causal estimand of interest is a 2-year event-free survival difference between ZDV + ddI and ZDV monotherapy, as at least 24 months of follow-up is required for the ACTG 175 trial. About 73% of the survival times were right-censored.

The ACTG 175 trial data are available from the R package `speff2trial` and were previously used in Yang et al. (2021) and Sheng and Ghosh (2020); all their analyses focus on comparing the effect of combination therapy and the ZDV monotherapy in the trial population. Responses to treatment, however, may vary according to the patient characteristics such as disease stage and the history of prior drug exposure (Hammer et al. 1996). Kennedy et al. (2021) have found that combination therapy may be more effective for the high-risk patients, but its utility for low-risk patients remains unclear. Therefore, our interest lies in transporting the results from the ACTG 175 trial with intermediate-stage HIV patients to the external target populations with the higher- and lower-risk compared to trial patients to evaluate the treatment effect in the target population. Specifically, we consider the following three target populations:

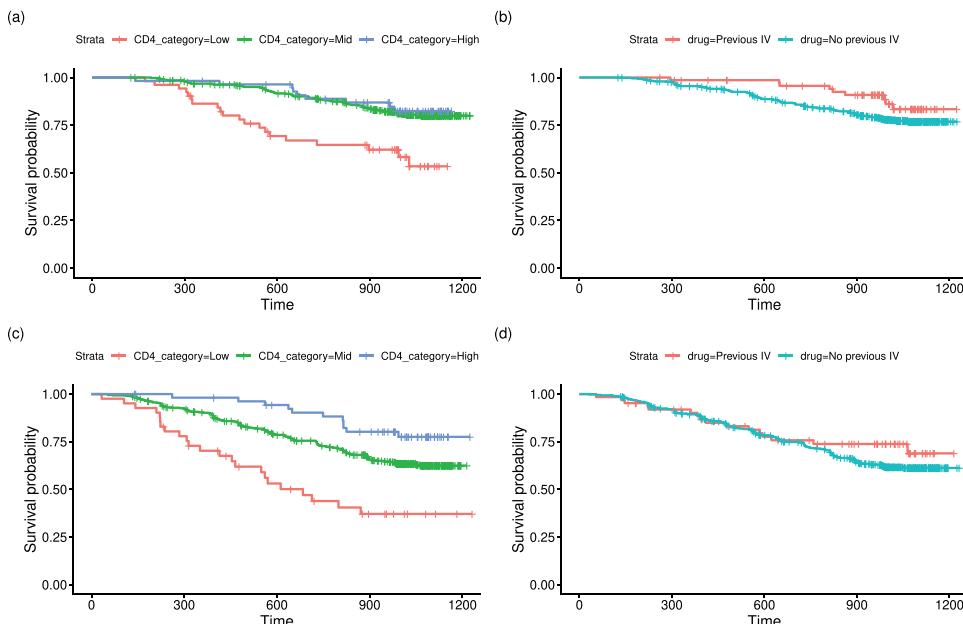
- (1) US early-stage HIV patients represented by the observational Acute Infection and Early Disease Research Program (AIEDRP) database (Hecht et al. 2006).
- (2) HIV patients in Thailand represented by the retrospective study by Manosuthi et al (Manosuthi et al. 2021).
- (3) HIV patients in southern Ethiopia represented by the retrospective study by Hailemariam et al (Hailemariam et al. 2016).

The baseline characteristics of the ACTG 175 trial and the three external datasets are summarized in [Table 1](#). We select seven covariates as these were considered prognostic factors for disease progression following the previous research (Sheng and Ghosh 2020) and also based on the external dataset availability. Even though not all factors are available in the external observational data as these were not collected for research purposes, some important covariates that are predictive biomarkers, e.g., CD4 count and drug, have been captured in external datasets. As seen in [Table 1](#), patients in the ACTG 175 trial are significantly different from those in the external datasets. Specifically, AIEDRP data consists of patients with higher CD4 counts and less history of previous intravenous drugs. On the other hand, patients in Thailand and southern Ethiopia studies have significantly lower CD4 count compared to the ACTG 175 trial. Thus, transporting to the external populations will lead to different quantification of the treatment effect.

Although the trial and external datasets provide a unique opportunity to understand how treatment works in large target populations, they also present two major challenges that need to be addressed, one intrinsic to the trial data and the other one with respect to the external data. First, for the ACTG 175 trial, there is a possibility that the common PH assumption is violated. [Figure 1](#) depicts the treatment-specific Kaplan-Meier (KM) curves by two prognostic covariates, i.e., CD4 category and drug. Instead of CD4 count, we use CD4 category defined in [Table 1](#) to illustrate via KM curves. [Figure 1\(a,d\)](#) depict the possible violation of the PH assumption by the delayed effect and the crossing curves. [Figure 1\(b,c\)](#)

**Table 1.** Summary of baseline characteristics of the ACTG 175 trial and the three external datasets.

Variable	ACTG 175 trial (n = 1054)	External datasets		
		US early-stage (n = 1762)	Thailand (n = 11911)	Southern Ethiopia (n = 2579)
<b>Male, %</b>	82.16	95.46	67.7	44.5
<b>Age (year), Mean ± SD</b>	35.23 ± 8.77	34.99 ± 8.48	32 ± 11	32.5 ± 9.1
<b>CD4 count (cells/mm<sup>3</sup>), Mean ± SD</b>	351 ± 122.3	545.7 ± 228.3		164.02 ± 117.84
<b>CD4 category, %</b>				
Low (< 200)	8.82	2.33	52.1	
Mid (201–500)	80.65	45.63	13.7	
High (> 500)	10.53	52.04	3.6	
<b>Race (White), %</b>	72.11	67.14		
<b>Gray Drug, %</b>	12.9	3.92		
<b>Weight (kg), Mean ± SD</b>	75.47 ± 13.43			52.22 ± 9.13

**Figure 1.** Treatment-specific kaplan-meier (KM) curves by two prognostic covariates; (a) KM curves of ZDV + ddI treatment group by CD4 category; (b) KM curves of ZDV + ddI treatment group by drug; (c) KM curves of ZDV treatment group by CD4 category; (d) KM curves of ZDV treatment group by drug. The CD4 category is defined as in Table 1.

also depict the possible violation by the delayed and decreasing effects. To formally test the validity of the PH assumption, similar to Sheng and Ghosh (2020), we checked the relationship between the Schoenfeld residuals and the time (Grambsch and Therneau 1994) using the R function *cox.zph*. We use the KM transformed time, i.e.,  $\widehat{S}(t) = \prod_{i:t_i \leq t} (1 - d_i/n_i)$ , where  $t_i$  is a time with at least one event happened,  $d_i$  is the number of events happened at  $t_i$ , and  $n_i$  is the number of subjects survived up to  $t_i$ , which is a default option to avoid the extreme outlier and for efficiency. For the ZDV + ddI treatment group, the global  $p$ -value is 0.028, implying that the null hypothesis of no departure from proportionality is rejected. In particular, covariates that show significant violation of proportionality are CD4 count ( $p = 0.0092$ ) and drug ( $p = 0.033$ ). For the ZDV monotherapy group, the global test is not significant ( $p = 0.194$ ), but there is a possibility of departure from proportionality for CD4 count ( $p = 0.049$ ). Figure 1 and the test using the Schoenfeld residuals suggest the possible PH violation in the ACTG 175 trial. Thus, using a Cox PH model without adjusting for such violation could lead to

a biased estimation of the treatment effect in the trial as well as in the target populations. This also suggests the necessity of non-PH models that do not require the PH assumption.

Second, for the external observational data for Thailand and southern Ethiopia, we were only able to find the summary statistics. As individual-level data were not publicly available, we emulated the external datasets based on the summary statistics incorporating the correlation structure of the trial data. The details of how we emulated the data are described in [Section 6.2](#). Moreover, some important covariates were captured only in a few external datasets or in different data types. For example, the covariate drug was collected only in the US early-stage data, and the summary of the baseline covariate CD4 count is provided in categories for the Thailand study.

### 3. Notation and identification for transportability

We now introduce some basic notations and identifiability conditions to transport the findings from the RCT, i.e., the ACTG 175 trial, to the target populations that are represented by three external datasets. Let  $X$  be a vector of  $p$ -dimensional baseline covariates and  $A$  be the binary treatment,  $A = \{0, 1\}$ . Employing the potential outcomes framework (Rubin [1974, 1986](#)), let  $T^a$  be the potential survival time if a subject receives the treatment  $A = a$ . The survival time  $T$  is defined as  $T = T^1 A + T^0(1 - A)$ . Let  $C$  be the censoring time. We assume noninformative censoring conditional on covariates and treatment throughout the paper. Due to the censoring, the survival time  $T$  is not always observable for all subjects. Instead, we observe  $U = \min(T, C)$  and  $\Delta = I(T \geq C)$  where  $I(\cdot)$  is an indicator function.

In transportability, we consider a super-population framework in which the trial and target populations are subpopulations. The RCT is sampled from a trial population with an unknown mechanism, and the external observational sample is randomly sampled from the external target population with some known mechanism. Let  $\delta = 1$  denote the binary indicator of trial participation and  $\tilde{\delta} = 1$  denote the binary indicator of external observational study participation. From the RCT sample, we observe  $\{U_i, \Delta_i, A_i, X_i, \delta_i = 1, \tilde{\delta}_i = 0\}_{i=1}^n$  from  $n$  independent and identically distributed subjects. For the external data, we consider a common setting that only the covariates information is available, i.e.,  $\{X_i, \delta_i = 0, \tilde{\delta}_i = 1\}_{i=n+1}^{n+m}$ , from  $m$  independent and identically distributed subjects drawn from the target population (Dahabreh and Hernán [2019](#); Lee et al. [2022](#)).

When comparing the survival of the two treatments, a popular estimand of interest is the difference in survival probability at landmark times. Let  $S_a(t) = P(T^a \geq t)$  be the treatment-specific survival probability curve,  $a \in \{0, 1\}$ . Accordingly, we define the target average treatment effect (TATE) (Josey et al. [2021](#)):  $\tau = E\{I(T^1 \geq t) - I(T^0 \geq t) | \tilde{\delta} = 1\} = S_1(t | \tilde{\delta} = 1) - S_0(t | \tilde{\delta} = 1)$ .

Define the treatment propensity score  $\pi_A(X) = P(A = 1 | X, \delta = 1)$  and the sampling score  $\pi_\delta(X) = P(\delta = 1 | X)$ . We now discuss the conditions to identify the TATE.

**Assumption 3.1** (Ignorability and positivity of trial treatment assignment)

(i)  $\{T^0, T^1\} \perp\!\!\!\perp A | (X, \delta = 1)$ ; and (ii)  $0 < \pi_A(X) < 1$  with probability 1.

**Assumption 3.2** (Conditional survival transportability)

$$P(T^a \geq t | X, \delta = 1) = P(T^a \geq t | X, \tilde{\delta} = 1) \text{ for } x \text{ s.t. } P(X = x | \tilde{\delta} = 1) > 0 \text{ and } a \in \{0, 1\}.$$

**Assumption 3.3** (Positivity of trial participation)  $0 < \pi_\delta(X) < 1$  for  $x$  s.t.  $P(X = x | \tilde{\delta} = 1) > 0$ .

Assumption 3.1 is a standard assumption and is likely to hold for well-defined RCTs in general. As the ACTG 175 trial implemented treatment randomization and had good patient compliance,

Assumption 3.1 is valid. Assumptions 3.2 and 3.3 are needed to transport findings from the RCT to the target population. Assumption 3.2 is formally weaker than the ignorability assumption on trial participation (Stuart et al. 2011; Westreich et al. 2017), i.e.,  $\{T^0, T^1\} \perp\!\!\!\perp \delta | X$ , but it suffices to identify the TATE as well as the potential survival curves in the target population,  $P(T^a \geq t | \tilde{\delta} = 1)$ . Several previous literature discussed the weaker version of Assumption 3.2,  $E\{I(T^1 \geq t) - I(T^0 \geq t) | X, \delta = 1\} = E\{I(T^1 \geq t) - I(T^0 \geq t) | X, \tilde{\delta} = 1\}$ , called exchangeability in measure (Dahabreh et al. 2020; Josey et al. 2021). However, the potential survival curves in the target population are not identifiable under the weaker version. Assumption 3.2 is not verifiable but plausible as some important covariates that are predictive of the disease progression, e.g., CD4 count and drug, were captured both in the ACTG 175 trial and external datasets. Assumption 3.3 implies the absence of patient characteristics in the target population that prevent them from participating in the trial; thus, covariates separating the trial and the target population should be excluded (Dahabreh et al. 2020; Tipton 2013), which is valid for the ACTG 175 trial and external data we are interested in.

In addition, we assume that the external observational data is drawn from the target population under a simple random sampling or more complex sampling designs. Accordingly, we define a known sampling weight  $d$ .

#### **Assumption 3.4** (The known design weight for the external observational sample) 1

*The observational sample design weight  $d = 1/P(\tilde{\delta} = 1 | X)$  is known.*

Under the above assumptions, the TATE, or  $S_a(t | \tilde{\delta} = 1)$ , are identified using only the observed data, by

$$S_a(t | \tilde{\delta} = 1) = \mathbb{E}\{S_a(t, X) | \tilde{\delta} = 1\}, \quad (1)$$

where  $S_a(t, X) = \mathbb{E}\{I(T \geq t) | X, A = a, \delta = 1, C \geq t\}$  is the conditional survival curves, for  $a \in \{0, 1\}$ . Alternatively,  $S_a(t | \tilde{\delta} = 1)$  can also be identified based on the IPW-based approach by

$$S_a(t | \tilde{\delta} = 1) = \frac{1}{P(\tilde{\delta} = 1)} \mathbb{E}\left[\frac{\delta}{\pi_{\delta}(X)} \frac{1}{d} \frac{I(A = a)}{\pi_A(X)^a \{1 - \pi_A(X)\}^{1-a}} \frac{Y(t)}{S_a^C(t, X)}\right], \quad (2)$$

where  $Y(t) = I(U \geq t)$  and  $S_a^C(t, X) = P(C > t | X, A = a, \delta = 1)$ , the conditional censoring probability model. By positivity assumptions, we have

$$P(\tilde{\delta} = 1) = \mathbb{E}\left[\frac{\delta}{\pi_{\delta}(X)} \frac{1}{d} \frac{I(A = a)}{\pi_A(X)^a \{1 - \pi_A(X)\}^{1-a}}\right].$$

Several methods for transportability motivated by each identification formula in (1) and (2) as well as jointly will be discussed in the following section.

## 4. Estimation methods for transportability

In this section, we present four different methods for transporting results from the ACTG 175 trial to estimate TATE under Assumptions 3.1–3.4. Due to the possible violation of the PH assumption in the ACTG 175 trial as described in Section 2, we consider both PH and non-PH models.

### 4.1. Inverse probability and Calibration Weighting (CW) model

A common approach for transportability is through the IPW-based approach following the identification formula in (2). As shown in (2), this approach can be viewed as a combination of IPSW, inverse

probability of treatment weighting (IPTW), and inverse probability of censoring weighting (IPCW). With respect to sampling score, a key idea to account for the different patient characteristics between the RCT sample and the target population represented by the external data is to weight the RCT sample with the inverse odds of sampling (Westreich et al. 2017). In particular, following IPSW, one can model  $\pi_\delta(X) = \{\omega_{\text{IPSW}}(X)\}^{-1}$  in (2), where  $\omega_{\text{IPSW}}(X) = P(\tilde{\delta} = 1 | \delta + \tilde{\delta} = 1, X) / P(\delta = 1 | \delta + \tilde{\delta} = 1, X)$  using the common logistic regression model. However, the IPSW method may not be stable if  $P(\delta = 1 | \delta + \tilde{\delta} = 1, X)$  is close to zero for some  $X$ , and it requires the sampling score model to be correctly specified.

Instead of estimating the sampling scores directly, the calibration weighting (CW) approach has been used in many recent works (Josey et al. 2021; Lee et al. 2021, 2022), which is known to be more stable than the IPSW approach. The CW approach is analogous to the entropy balancing method (Hainmueller 2012) and is based on the following identity,

$$E\left\{\frac{\delta}{\pi_\delta(X)} \mathbf{g}(X)\right\} = E\left\{\tilde{\delta} d\mathbf{g}(X)\right\} = E\{\mathbf{g}(X) | \tilde{\delta} = 1\}, \quad (3)$$

where  $\mathbf{g}(X)$  is a function of  $X$  to be calibrated, such as the moment functions or nonlinear transformations. That is, the calibrated RCT sample empirically matches the external data that represents the target population. Empirically, the calibration weights  $\omega_i$  are the solutions to the following optimization problem

$$\min_{\mathcal{W}} \sum_{i=1}^n \omega_i \log \omega_i, \quad (4)$$

$$\text{subject to } \omega_i \geq 0, \quad i, \sum_{i=1}^n \omega_i = 1, \text{ and } \sum_{i=1}^N \delta_i \omega_i \mathbf{g}(X_i) = \tilde{\mathbf{g}},$$

where  $\mathcal{W} = \{\omega_i : \delta_i = 1\}$  and  $\tilde{\mathbf{g}} = \sum_{i=1}^N \tilde{\delta}_i d_i \mathbf{g}(X_i) / \sum_{i=1}^N \tilde{\delta}_i d_i$ , a consistent estimator of  $E\{\mathbf{g}(X) | \tilde{\delta} = 1\}$ . The calibration weights  $\omega_i$  are also the solution to the Lagrangian dual problem  $L(\lambda, \mathcal{W}) = \sum_{i=1}^n \omega_i \log \omega_i - \lambda^T \{\sum_{i=1}^n \omega_i \mathbf{g}(X_i) - \tilde{\mathbf{g}}\}$  where  $\hat{\lambda}$  solves  $U(\lambda) = \sum_{i=1}^n \exp\{\lambda^T \mathbf{g}(X_i)\} \{\mathbf{g}(X_i) - \tilde{\mathbf{g}}\} = 0$ . The estimated calibration weights are  $\hat{\omega}_i = \omega(X_i; \hat{\lambda}) = \exp\{\hat{\lambda}^T \mathbf{g}(X_i)\} / [\sum_{i=1}^n \exp\{\hat{\lambda}^T \mathbf{g}(X_i)\}]$ . For the calibration weighting, Lee et al. (2021) posited the loglinear sampling score model

$$\pi_\delta(X) = \exp\{\eta_0^T \mathbf{g}(X)\}, \text{ for some } \eta_0. \quad (5)$$

This is because the objective function in (4) has a solution that has the same functional form as inverse probability of sampling score under the loglinear model. It is known that  $\hat{\lambda}$  is equivalent to  $-\hat{\eta}$  (Lee et al. 2021), thus  $\hat{\pi}_\delta(X)$  and  $\hat{\omega}_i$  can be expressed using  $\hat{\eta}$ , i.e.,  $\hat{\pi}_\delta(X) = \pi_\delta(X; \hat{\eta}) = \exp\{\hat{\eta}^T \mathbf{g}(X)\}$  and  $\hat{\omega}_i = \omega(X_i; \hat{\eta}) = \exp\{-\hat{\eta}^T \mathbf{g}(X_i)\} / [\sum_{i=1}^n \exp\{-\hat{\eta}^T \mathbf{g}(X_i)\}]$ .

In addition to the sampling score model, one can model the treatment propensity scores using a common logistic regression model,

$$\pi_A(X) = [1 + \exp\{-\rho_0^T \mathbf{g}(X)\}]^{-1}, \text{ for some } \rho_0. \quad (6)$$

Even though  $\pi_A(X)$  is generally known for RCTs, many researchers have suggested estimating the treatment propensity scores which can increase the efficiency and account for the chance of imbalance of prognostic variables [e.g., Lee et al. 2022; Williamson et al. 2014]. Moreover, the right censoring needs to be accounted for. We assume the noninformative censoring assumption conditional on covariates and treatment, i.e.,  $\{T^1, T^0\} \perp\!\!\!\perp C | (X, A, \delta = 1)$ , which also implies  $T \perp\!\!\!\perp C | (X, A, \delta = 1)$ . A widely used Cox proportional hazard model can be used, with the conditional hazard

$$\lambda^C(t|X, A = a) = \lambda_{a0}^C(t) \exp(\gamma_a^T X), \text{ for } a \in \{0, 1\}. \quad (7)$$

Define  $\widehat{\pi}_{ai} = A_i \pi_A(X_i; \widehat{\rho}) + (1 - A_i)\{1 - \pi_A(X_i; \widehat{\rho})\}$ ,  $A_{ai} = I(A_i = a)$ , and  $\widehat{\Lambda}_{ai}^C(t) = \widehat{\Lambda}_{a0}^C(t) \exp(\widehat{\gamma}_a^T X_i)$  where  $\Lambda_{a0}^C(t) \equiv \int_0^t \lambda_{a0}^C(u) du$ . The CW estimator of the treatment-specific survival curve is

$$\widehat{S}_a^{\text{CW}}(t) = \sum_{i=1}^N \delta_i \widehat{\omega}_i \frac{A_{ai}}{\widehat{\pi}_{ai}} e^{\widehat{\Lambda}_{ai}^C(t)} Y_i(t). \quad (8)$$

The IPSW estimator can be defined by replacing the calibration weights  $\widehat{\omega}_i$  in (8) with  $\widehat{\omega}_{\text{IPSW}}(X_i)$ .

#### 4.2. Outcome regression with a PH assumption

Another method for transporting beyond the trial population relies on the conditional survival curves in the identification formula (1), known as the outcome regression (OR) approach. A widely used Cox PH model assumes the treatment-specific conditional hazard function as

$$\lambda_{ai}(t) = \lambda_a(t | X_i) = \lambda_{a0}(t) \exp(\beta_a^T X_i). \quad (9)$$

Following Chen and Tsiatis (2001), one can adjust for the imbalances between the RCT sample and the external data by averaging the treatment effect in the trial sample over the external data. Specifically, first estimating the treatment-specific survival curve conditional on  $\delta = 1$  under model (9), i.e.,  $\widehat{S}_a(t, X_i) = \exp\{-\widehat{\Lambda}_{ai}(t)\} = \exp\{-\widehat{\Lambda}_{a0}(t) \exp(\widehat{\beta}_a^T X_i)\}$  for the RCT participants, then applying the design-weighted averaging over  $f(X|\delta = 1)$  thus transporting the RCT results to the target population. The resulting OR estimator of the treatment-specific survival curve is

$$\widehat{S}_a^{\text{OR}}(t) = \left( \sum_{i=1}^N \widetilde{\delta}_i \right)^{-1} \sum_{i=1}^N \widetilde{\delta}_i e^{-\widehat{\Lambda}_{ai}(t)}, \quad (10)$$

which is consistent under the PH assumption.

#### 4.3. Augmented Calibration Weighing (ACW) model with a PH assumption

The CW estimator specified in (8) and the OR estimator specified in (10) are singly robust in that the former is consistent only under the correct weighting models (5) – (7) and the latter is consistent under the correct survival outcome regression model (9). Lee et al. (2022) proposed an improved estimator, named the ACW estimator, that combines the CW and the OR estimators employing the semiparametric theory (Tsiatis 2006). The ACW estimator is in the form of

$$\widehat{S}_a^{\text{ACW}}(t) = \exp\left\{-\int_0^t \frac{\text{num}}{\text{denom}}\right\}, \quad (11)$$

where

$$\begin{aligned} \text{denom} &= \sum_{i=1}^N \delta_i \widehat{\omega}_i \frac{A_{ai}}{\widehat{\pi}_{ai}} e^{\widehat{\Lambda}_{ai}^C(t)} Y_i(t) \\ &+ \sum_{i=1}^N e^{-\widehat{\Lambda}_{ai}(t)} \left[ \left( \sum_{i=1}^N \widetilde{\delta}_i \right)^{-1} \widetilde{\delta}_i - \delta_i \widehat{\omega}_i \frac{A_{ai}}{\widehat{\pi}_{ai}} \left\{ 1 - \int_0^t \left\{ e^{\widehat{\Lambda}_{ai}^C(u) + \widehat{\Lambda}_{ai}(u)} \right\} d\widehat{M}_{ai}^C(u) \right\} \right], \end{aligned} \quad (12)$$

and

$$\begin{aligned}
num &= \sum_{i=1}^N \delta_i \widehat{\omega}_i \frac{A_{ai}}{\widehat{\pi}_{ai}} e^{\widehat{\Lambda}_{ai}^C(u)} dN_i(u) \\
&+ \sum_{i=1}^N e^{-\widehat{\Lambda}_{ai}(u)} d\widehat{\Lambda}_{ai}(u) \left[ \left( \sum_{i=1}^N \widetilde{\delta}_i \right)^{-1} \widetilde{\delta}_i - \delta_i \widehat{\omega}_i \frac{A_{ai}}{\widehat{\pi}_{ai}} \left\{ 1 - \int_0^u \left\{ e^{\widehat{\Lambda}_{ai}^C(s) + \widehat{\Lambda}_{ai}(s)} \right\} d\widehat{M}_{ai}^C(s) \right\} \right].
\end{aligned}$$

This approach is based on the representation  $\Lambda_a(t|\tilde{\delta} = 1) = \frac{t}{0} - \{S_a(u|\tilde{\delta} = 1)\}^{-1} dS_a(u|\tilde{\delta} = 1)$  where *num* estimates  $-dS_a(u|\tilde{\delta} = 1)$  and *denom* estimates  $S_a(u|\tilde{\delta} = 1)$  separately (Zhang and Schaubel 2012).

**Theorem 4.1.** Under Assumptions 3.1–3.4, the ACW estimator in (11) is doubly robust in a sense that its consistency is ascertained if either the conditional survival curve  $S_a(t, X)$  is correctly specified or the weight models (5)–(7), i.e., the sampling score model  $\pi_\delta(X)$ , the treatment propensity score model  $\pi_A(X)$ , and the censoring model  $\lambda^C(t|X, A = a)$ , are correctly specified. In addition, if all working models are correctly specified, the ACW estimator achieves the semiparametric efficiency.

Theorem 4.1 claims that the ACW estimator has two nice properties, namely its double robustness and local efficiency. Note that the denominator itself is a doubly robust and locally efficient estimator of  $S_a(t|\tilde{\delta} = 1)$ , but it was found to show worse performance than the  $\widehat{S}_a^{\text{ACW}}(t)$  in a finite sample (Lee et al. 2022; Zhang and Schaubel 2012). A nonparametric bootstrap method can be used for a straightforward estimation of the variance of the ACW estimator.

#### 4.4. ACW with a hazard regression (HARE) model

The ACW estimator in (11) is a consistent estimator under the PH assumption, which, in many case studies, may not be correct. In particular, as shown in Figure 1 and the formal PH assumption tests, the PH assumption may be violated in the ACTG 175 data. In this situation, the ACW estimator in (11) forfeits its double robustness property, as its consistency solely depends on the correct specification of the weight functions. Furthermore, it lacks semiparametric efficiency even when the weight functions are correctly specified. The ACW estimator can be extended to account for such non-PH data using the HARE for the outcome model. The HARE model is based on the linear splines and their tensor products and does not depend on the PH assumption (Kooperberg et al. 1995).

Let the vector of  $p$ -dimensional covariates  $X$  lies in  $\mathcal{X} \subseteq \mathbb{R}^p$  and let  $\mathcal{G}$  be a  $p$ -dimensional linear space of functions on  $[0, \infty) \times \mathcal{X}$  such that  $g(\cdot|X)$  is bounded on  $[0, \infty)$  for  $g \in \mathcal{G}$  and  $x \in \mathcal{X}$ . The treatment-specific conditional log-hazard function for the HARE model is

$$\log \lambda_a^H(t|X) = \sum_{k=1}^p \beta_{ak}^H B_k(t|X), \quad (13)$$

where  $B_1(\cdot), B_p(\cdot)$  are the basis of  $\mathcal{G}$ . The coefficients  $\beta_a^H = (\beta_{a1}^H, \dots, \beta_{ap}^H)^T$  are estimated using the maximum likelihood estimation. We use a superscript  $H$  to denote the HARE model.

The HARE model is implemented in the R function *hare* in the R package *polyspline*. It considers knots in the covariates and time and the pairwise interaction between covariates and time, and performs the variable selection simultaneously. For example, an  $l$ th knot of the  $j$ th covariate  $x_j$  is represented by the basis function  $(x_j - x_{jl})_+$ , and an  $l$ th knot of time  $t$  is represented by the basis function  $(t_l - t)_+$ , where  $x_+ = \max(x, 0)$ . Then the variable selection for the pairwise covariate interactions as well as covariates-time interactions are conducted using stepwise addition, stepwise deletion, and the Akaike Information Criterion. If the interaction between the covariate and time is

selected in the model, then the HARE model becomes a non-PH model, which still correctly specifies the survival outcome model  $S_a(t, X)$ . We can easily combine the HARE model with the ACW estimator, by substituting  $\widehat{\Lambda}_a^H(t|X_i) = \int_0^t \widehat{\lambda}_a^H(u|X_i)du$  in (11) for  $\widehat{\Lambda}_{ai}(u)$ , and the resulting ACW estimator remains consistent under Theorem 4.1. However, the Cox PH model is prone to misspecification when facing such a non-PH pattern in the data. With the HARE model, the ACW estimator gains robustness to the violation of the PH assumption, and our simulation studies in Section 4 provide more empirical evidence.

## 5. Simulation

In this section, we present a set of simulation studies demonstrating that the ACW estimator, when applied with the HARE model, exhibits greater robustness against violations of the PH assumptions compared to its counterpart using the Cox PH model in finite samples. We consider a target population of size  $N = 200,000$  with covariates  $X = (X_1, X_2, X_3)$ , generated from  $N(0, 1)$  and truncated at  $-4$  and  $4$  to satisfy the regularity conditions. A RCT sample of size  $n \approx 1,300$  is selected from a hypothetical RCT population with a size of  $50,000$ . The binary indicator of trial participation, denoted as  $\delta$ , is generated by Bernoulli sampling with probabilities  $\pi_\delta(X) = \{1 + \exp(-3.4 - 0.5X_1 - 0.5X_2 - 0.3X_3)\}^{-1}$ . Treatment is randomly assigned within the selected RCT sample with a probability of  $0.5$ . The remaining is the external target population, from which a random sample of size  $m = 5,000$  is chosen.

To generate time-to-event data while allowing for the violation of the PH assumption, we consider treatment-specific conditional hazard functions  $\lambda_a(t|X)$  with time-varying covariate effects:

$$\lambda_0(t|X) = \exp(-3) \exp\{-1.8X_1 - 1.5X_2 - X_3 - t\alpha(X_1 + X_2 + X_3)\}$$

$$\lambda_1(t|X) = \exp(-3.7) \exp\{-X_1 - X_2 - 1.5X_3 - t\alpha(X_1 + X_2 + X_3)\}.$$

Here, the extent of violation of the PH assumption is measured by the parameter  $\alpha$ . We select values for  $\alpha$  as  $\log(1)$ ,  $\log(1.0006)$ , and  $\log(1.0012)$ , representing scenarios where the PH assumption holds, is slightly violated, and is strongly violated, respectively. Time-to-event outcomes  $T$  are generated by inverting the survival function induced by the hazard functions  $\lambda_a(t|X)$  (Austin 2012). Similarly, censoring times  $C$  are generated by considering the hazard  $\lambda^C(t|X) = \exp(-3.5) \exp(-0.5X_1 - X_2 - X_3)$  and inverting its corresponding survival function. We compare the ACW estimator with the HARE model ( $\text{ACW}_{\text{HARE}}$ ) to the one with the Cox PH model ( $\text{ACW}_{\text{PH}}$ ) to showcase the improved robustness of  $\text{ACW}_{\text{HARE}}$ .

The parameter of interest is the difference in survival probabilities between the treatment group and the control group at  $t = 20$ , denoted as  $\tau = P(T^1 \geq 20|\delta = 1) - P(T^0 \geq 20|\delta = 1)$ . To assess the performance of the proposed ACW estimator under model misspecification, we consider two scenarios: In the first scenario, we use the same set of covariates  $X$  to fit the weight models (5)-(7) (i.e., the weight models are correctly specified). In the second scenario, we use the transformed covariates  $X^* = (\exp(X_1/2), X_2/\{1 + \exp(X_1)\}, (X_1X_3/25 + 0.6)^2)$  to fit the weight models instead (i.e., the weight models are misspecified).

**Table 2** presents the bias, standard error, mean squared error and 95% coverage probability for estimating differences in survival probabilities using 500 replicated datasets. The results suggest that  $\text{ACW}_{\text{HARE}}$  and  $\text{ACW}_{\text{PH}}$  demonstrate comparable efficiency when the weight models are correctly specified for all values of  $\alpha$ . However, in the second scenario,  $\text{ACW}_{\text{PH}}$  exhibits inflated mean squared errors and unsatisfactory coverage probabilities, particularly when the PH assumption is strongly violated. This is reasonable as  $\text{ACW}_{\text{PH}}$  is unable to handle such non-PH data, and its consistency is forfeited when the weight functions are misspecified. In contrast,  $\text{ACW}_{\text{HARE}}$  adapts to the non-PH



**Table 2.** Bias, standard error, mean squared error and coverage probability of 95% confidence intervals for the estimation of the differences in survival probabilities at  $t = 20$ . Two methods are compared: ACW estimator with cox PH model ( $\text{ACW}_{\text{PH}}$ ) and ACW estimator with the HARE model ( $\text{ACW}_{\text{HARE}}$ ) when the PH assumption holds, is slightly violated and is strongly violated.

$\alpha$		Bias ( $\times 10^{-2}$ )	Std err ( $\times 10^{-2}$ )	MSE ( $\times 10^{-2}$ )	CP ( $\times 10^{-2}$ )
The weight models are correctly specified					
log(1)	$\text{ACW}_{\text{PH}}$	0.29	4.28	0.18	93.4
	$\text{ACW}_{\text{HARE}}$	0.19	4.30	0.18	92.7
log(1.0006)	$\text{ACW}_{\text{PH}}$	0.14	4.38	0.19	93.7
	$\text{ACW}_{\text{HARE}}$	0.06	4.34	0.19	92.8
log(1.0012)	$\text{ACW}_{\text{PH}}$	0.27	4.34	0.19	93.8
	$\text{ACW}_{\text{HARE}}$	0.18	4.39	0.19	93.4
The weight models are misspecified					
log(1)	$\text{ACW}_{\text{PH}}$	2.52	4.08	0.23	92.6
	$\text{ACW}_{\text{HARE}}$	1.54	3.84	0.18	93.6
log(1.0006)	$\text{ACW}_{\text{PH}}$	2.89	3.99	0.24	91.7
	$\text{ACW}_{\text{HARE}}$	1.56	4.36	0.21	92.8
log(1.0012)	$\text{ACW}_{\text{PH}}$	3.52	4.10	0.29	84.6
	$\text{ACW}_{\text{HARE}}$	1.88	4.13	0.23	93.0

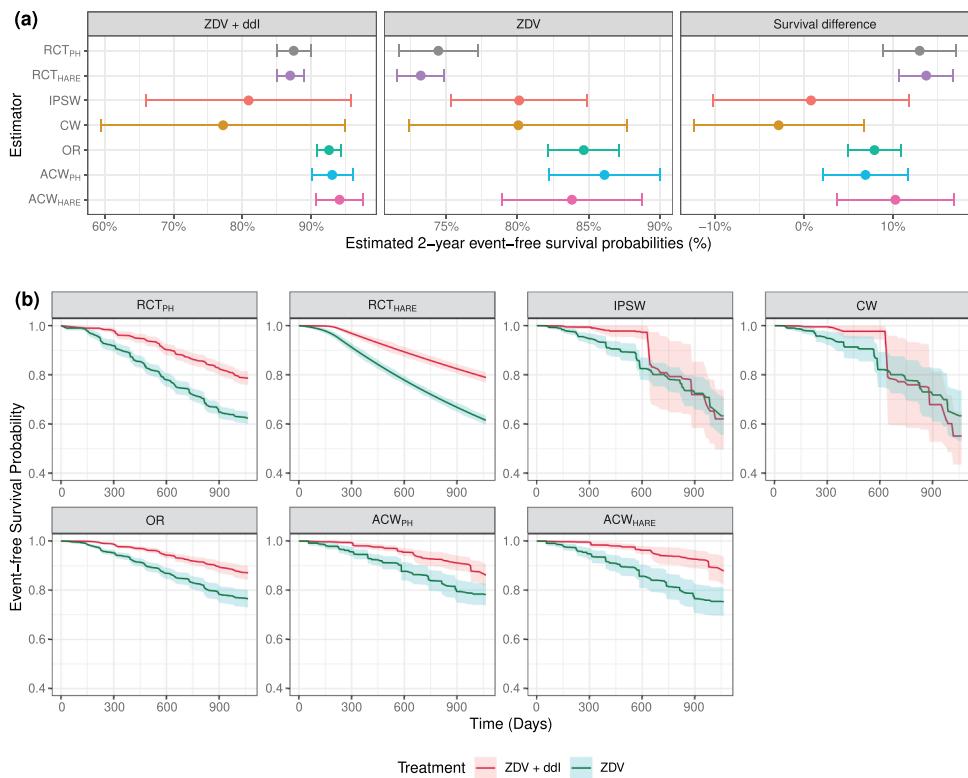
model and maintains its consistency even if the weight models are incorrectly specified. This empirical evidence supports the robustness property of  $\text{ACW}_{\text{HARE}}$ , aligning with Theorem 4.1.

## 6. Results of transporting the ACTG 175 trial

### 6.1. Transporting the treatment effect to US early-stage HIV patients

The AIEDRP cohort study is an observational cohort study that enrolled patients within 1 year of having HIV antibody seroconversion (Hecht et al. 2006). An important question could be an estimation of the effect of the combination therapy ZDV + ddI over ZDV monotherapy in the US early-stage HIV patient population represented by the AIEDRP study that consists of patients with higher CD4 count and less history of intravenous drug use, transporting from the ACTG 175 trial with intermediate-stage HIV patients.

Figure 2 plots estimated treatment-specific event-free survival probabilities in the US early-stage HIV patients. Figure 2(a) shows the estimated 2-year event-free survival probabilities by treatment and their differences in early-stage HIV patients in the US. The models based only on the ACTG 175 trial, i.e.,  $\text{RCT}_{\text{PH}}$  and  $\text{RCT}_{\text{HARE}}$ , give the estimated survival probabilities of about 84% for the ZDV + ddI and 74% for the ZDV. Thus, there is a 13% higher chance of survival for ZDV + ddI over ZDV monotherapy, and no notable differences between  $\text{RCT}_{\text{PH}}$  and  $\text{RCT}_{\text{HARE}}$ . On the other hand, after transporting the ACTG 175 trial, the estimated 2-year survival probabilities in the US early-stage HIV patients population are higher for both treatment groups in general. Specifically, the  $\text{OR}_{\text{PH}}$ ,  $\text{ACW}_{\text{PH}}$ , and  $\text{ACW}_{\text{HARE}}$  estimators give an estimate of about 93% and 84% 2-year survival probabilities for ZDV + ddI and ZDV, respectively. The estimated 2-year survival differences in the early-stage HIV patients are smaller than that in ACTG 175 patients. The  $\text{OR}_{\text{PH}}$  and  $\text{ACW}_{\text{PH}}$  estimators that are based on the Cox PH model show about a 7% increase in 2-year survival probabilities. On the other hand, the  $\text{ACW}_{\text{HARE}}$  estimator show about a 10% increase in 2-year survival probability for ZDV + ddI over ZDV monotherapy. Given the possible violation of the PH assumption in the ACTG 175 trial, the estimated treatment effect under the PH model could be underestimated. All these differences were significant at the 0.05 level. Figure 2(b) plots the estimated treatment-specific survival curves. After transporting to the US early-stage HIV patients, survival curves for both treatment groups are higher than the survival curves in the ACTG 175 trial patients across time. The combined therapy is more effective in treating the intermediate-stage HIV patients in the trial than early-stage patients. This may be because the combined therapy is more toxic and lower the compliance in the early-stage patients, which decreases the treatment effect of the combined therapy over the ZDV monotherapy.



**Figure 2.** Estimated treatment-specific event-free survival probabilities in the US early-stage HIV patients; (a) 2-year event-free survival probabilities; (b) treatment-specific survival curves.

The IPSW and CW estimators estimate treatment-specific survival curves that are in different shapes from the estimators with outcome models. Moreover, the estimated 2-year survival differences are not significant with larger variability. As IPSW and CW estimators are singly robust, these could be due to the misspecified sampling score model, resulting in the biased estimation of the treatment effect. In contrast, the ACW estimators are robust to such misspecification.

## 6.2. Transporting the treatment effect to HIV patients in Thailand and Southern Ethiopia

Next, we transport the ACTG 175 trial to the HIV patients in two countries outside of the US: Thailand and southern Ethiopia. We first estimate the treatment effect in Thailand HIV patients. With about 500,000 HIV patients and 12,000 patients died of HIV-related diseases in 2020, Thailand is considered to be the country with a high HIV burden in Asia/Pacific regions (Manosuthi et al. 2021). A retrospective study was conducted among the HIV patients registered in the national AIDS program database in eight provinces in Thailand. We are interested in evaluating the treatment effect in HIV patients in Thailand represented by the retrospective study. This observational study includes more than 10,000 patients with lower CD4 counts than the ACTG 175 patients.

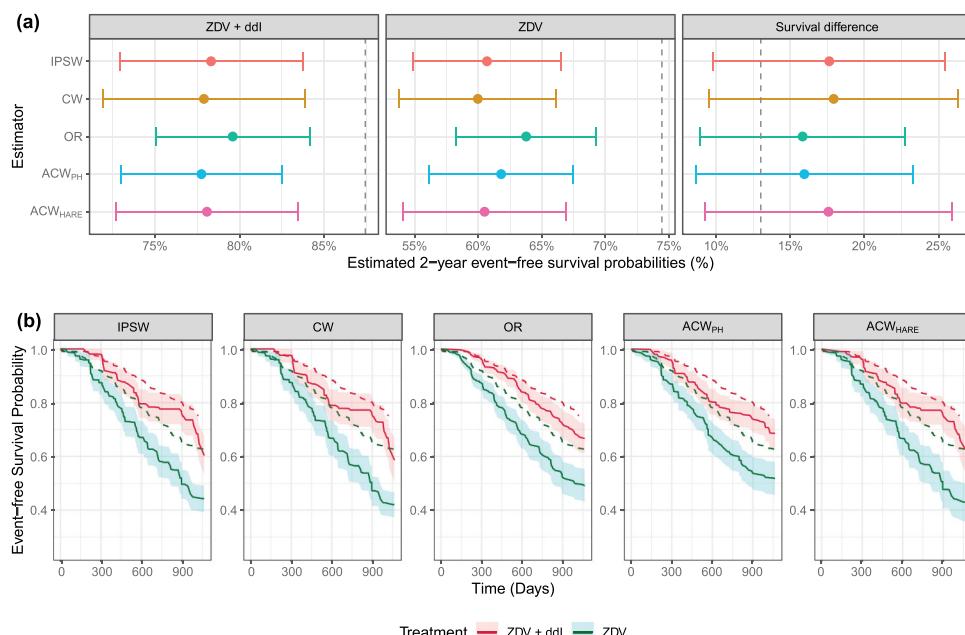
As mentioned in Section 2, the individual-level data for HIV patients in Thailand was not available. However, it is known that even only using summary statistics, the inferential population is well defined (Chu et al. 2022). We emulated the external observational data based on the baseline characteristics in Table 1. Specifically, categorical covariates such as gender and CD4 category were generated from a categorical distribution with the given proportion. The numerical covariate age was generated from



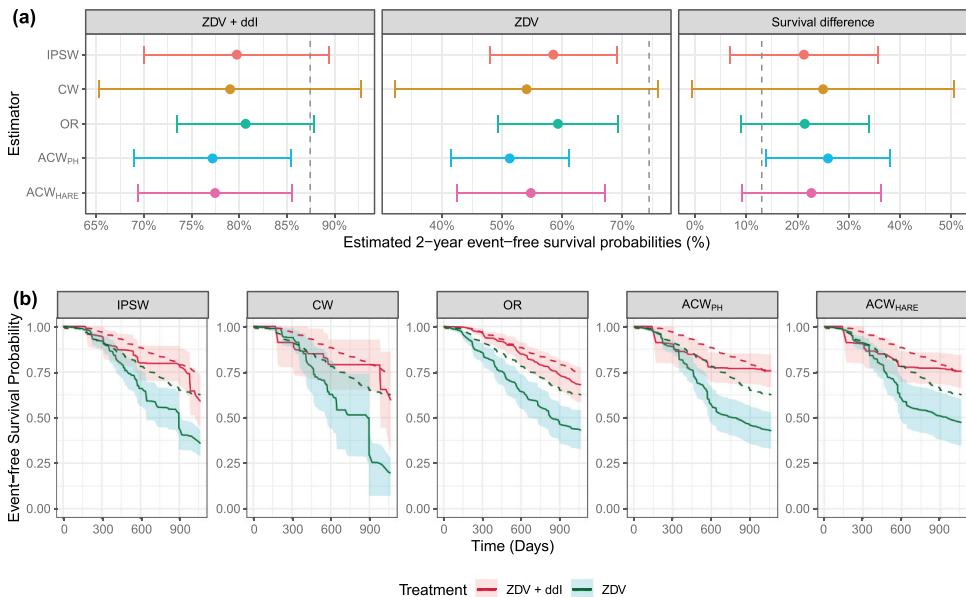
the shifted beta distribution, with mean and variance given from the summary data. The range of the beta distribution was determined following the ACTG 175 trial data unless otherwise specified in the original paper in (Manosuthi et al. 2021). To make the generated data more realistic, we incorporate the correlation structure of the ACTG 175 trial to the emulated Thailand data using the Gaussian copula, assuming that the correlation structure of the trial and the external data are similar. The sensitivity analyses show that the results are robust to the correlation structure as well as the randomness in data generation (see Appendix A).

Figure 3 depicts the estimated treatment-specific event-free survival probabilities in Thailand HIV patients. Figure 3(a) shows that all five estimators that transport the ACTG 175 trial results to the Thailand patients show about 16–17% 2-year survival increase of ZDV + ddI over ZDV monotherapy, which is larger than the 13% survival increase in the ACTG 175 trial patients shown in grey dashed vertical line. According to Figure 3(b), for both treatment groups, the estimated survival probabilities for Thailand HIV patients were found to be lower than that of the trial patients shown in dashed lines. These results make sense as the Thailand study consists of more than half of patients with low CD4 counts ( $\leq 200$ ), considered less healthy patients, than the ACTG 175 trial with only about 9% patients with low CD4 counts. With the PH assumption, the estimated 2-year survival difference is around 15%, whereas RCT<sub>HARE</sub> shows a larger treatment effect, with about a 17.6% survival probability increase. The latter is similar to the estimates from the IPSW and CW estimators that do not involve the PH assumption. In Figure 3(b), estimated survival curves from the RCT<sub>HARE</sub> estimator show different trends from the curves with the PH assumption and are shown to be more efficient than the estimated curves from the IPW-based estimators.

We are also interested in assessing the treatment effect in HIV patients in southern Ethiopia. A retrospective study was conducted including the HIV-infected patients enrolled in Dilla University Hospital located in southern Ethiopia (Hailemariam et al. 2016). The study consists of more females and patients having lower CD4 count compared to the ACTG 175 patients. As the individual-level data for southern Ethiopia patient data is not available, we emulate the external observational data based on



**Figure 3.** Estimated treatment-specific event-free survival probabilities in Thailand HIV patients; (a) 2-year event-free survival probabilities; (b) treatment-specific survival curves. Dashed lines indicate the RCT<sub>PH</sub> model.



**Figure 4.** Estimated treatment-specific event-free survival probabilities in southern Ethiopia HIV patients; (a) 2-year event-free survival probabilities; (b) treatment-specific survival curves. Dashed lines indicate the estimates from the RCT<sub>PH</sub> model.

the summary statistics in Table 1 and the correlation structure of the ACTG 175 data, similar to the Thailand data.

Figure 4 plots the estimated treatment-specific event-free survival probabilities in southern Ethiopia HIV patients. Figure 4(a) shows the estimated 2-year survival difference after transporting the ACTG 175 trial to the southern Ethiopia patients. The estimated ATE in southern Ethiopia patients is larger than that in the ACTG 175 trial in a grey dashed vertical line. Figure 4(b) shows estimated treatment-specific survival curves. All five curves show a trend of no significant difference at the initiation of the treatment and then significance after a year or more on treatment. The RCT<sub>PH</sub> and RCT<sub>HARE</sub> show similar estimated survival curves. The IPSW and CW estimator show relatively large variability, which could be due to the misspecification of the sampling score model.

## 7. Conclusions

In this paper, we have explored various models for transporting the treatment effect from the ACTG 175 trial to the three target populations including both the PH and non-PH models. Due to the heterogeneity in the patient characteristics between the ACTG 175 trial and the external data that represent the target populations, the TATE is different from the treatment effect estimated in the trial. Specifically, as patients in the US early-stage HIV patients are relatively healthier than the patients in ACTG 175 trial who are in intermediate-stage disease, the estimated treatment-specific survival probabilities are higher and their difference is smaller in the target population compared to those in the trial. On the other hand, as HIV patients in Thailand and southern Ethiopia are relatively sicker with lower baseline CD4 counts than the ACTG 175 trial patients, the estimated treatment-specific survival probabilities were found to be lower and differences were larger in general. These results correspond to that of (Kennedy et al. 2021) in that combination therapy could benefit more the high-risk or sicker patients. Moreover, the estimated TATEs from the ACW with the HARE model and the ACW with the Cox PH model are different. As the PH assumption may be violated in the ACTG 175 trial, the latter might be biased whereas the former corrects the bias by accounting for the non-PH model. The ACW with the HARE model also shows more robustness and efficiency in general than the



IPSW and CW estimators which do not depend on the PH assumption but are singly robust to the sampling score model.

The presented analysis is based on the assumption that the trial participation is ignorable, i.e., all covariates related to the trial participation and survival time are captured. Even though some covariates that are known to be highly predictive of the disease progression were captured in the external observational data, some important covariates may not be available as the observational samples were not originally collected for the research purpose. Sensitivity analyses can be conducted to assess the robustness of the presented results in the presence of unmeasured covariates in external data, e.g., VanderWeele and Ding (2017) and Yang and Lok (2017). Moreover, some external data, such as Thailand or southern Ethiopia studies, do not provide individual-level data but only the summary statistics of the covariate distribution from the target populations. Several recent works suggest that the analysis based on reliable summary statistics for covariates could lead to the valid inference of the target population, e.g., Chu et al. (2022). Also, the proposed external data generating process was found to be robust to the correlation structure and the randomness in data generation. Nonetheless, the results from the emulated data could be different from the results analyzing the actual individual-level data.

## Acknowledgements

We are grateful to the patients who volunteered for AIEDRP, to the AIEDRP study team, and to Susan Little and Davey Smith for their help and advice in interpreting the AIEDRP database.

## Disclosure statement

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

## Funding

Yang is partly supported by SES 2242776, NIH 1R01AG066883 and 1R01ES031651.

## ORCID

Shu Yang <http://orcid.org/0000-0001-7703-707X>

## References

- Austin, P. C. 2012. Generating survival times to simulate cox proportional hazards models with time-varying covariates. *Statistics in Medicine* 31 (29):3946–3958. doi:[10.1002/sim.5452](https://doi.org/10.1002/sim.5452).
- Chen, P. Y., and A. A. Tsiatis. 2001. Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics Bulletin* 57 (4):1030–1038. doi:[10.1111/j.0006-341X.2001.01030.x](https://doi.org/10.1111/j.0006-341X.2001.01030.x).
- Chu, J., W. Lu, and S. Yang. 2022. Targeted optimal treatment regime learning using summary statistics. *Biometrika* 110 (4):913–931.
- 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](https://doi.org/10.1093/aje/kwq084).
- Colnet, B., I. Mayer, G. Chen, Dieng, A., Li, R., Varoquaux, G. & Yang, S. 2020. Causal inference methods for combining randomized trials and observational studies: A review. *Statistical science*, 39 (1): 165–191. *arXiv Preprint arXiv: 201108047*.
- 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](https://doi.org/10.1007/s10654-019-00533-2).
- Dahabreh, I. J., L. C. Petito, S. E. Robertson, M. A. Hernán, and J. A. Steingrimsson. 2020. Toward causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a new target population. *Epidemiology* 31 (3):334–344. doi:[10.1097/EDE.0000000000001177](https://doi.org/10.1097/EDE.0000000000001177).
- Dahabreh, I. J., S. E. Robertson, J. A. Steingrimsson, E. A. Stuart, and M. A. Hernán. 2020. Extending inferences from a randomized trial to a new target population. *Statistics in Medicine* 39 (14):1999–2014. doi:[10.1002/sim.8426](https://doi.org/10.1002/sim.8426).

- 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](https://doi.org/10.1111/biom.13009).
- Grambsch, P. M., and T. M. Therneau. 1994. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 81 (3):515–526. doi:[10.1093/biomet/81.3.515](https://doi.org/10.1093/biomet/81.3.515).
- Hailemariam, S., G. Tenkolu, H. Tadese, Vata, P. K. 2016. Determinants of survival in hiv patients: A retrospective study of dilla university hospital hiv cohort. *International Journal of Virology and AIDS* 3(2):23. doi:[10.23937/2469-567X/1510023](https://doi.org/10.23937/2469-567X/1510023).
- 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](https://doi.org/10.1093/pan/mpr025).
- Hammer, S. M., D. A. Katzenstein, M. D. Hughes, H. Gundacker, R. T. Schooley, R. H. Haubrich, W. K. Henry, M. M. Lederman, J. P. Phair, M. Niu, et al. 1996. A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine* 335 (15):1081–1090. doi:[10.1056/NEJM199610103351501](https://doi.org/10.1056/NEJM199610103351501).
- Hecht, F. M., L. Wang, A. Collier, S. Little, M. Markowitz, J. Margolick, J. & Kilby, E. Daar, B. Conway, S. Holte, et al. 2006. A multicenter observational study of the potential benefits of initiating combination antiretroviral therapy during acute hiv infection. *The Journal of Infectious Diseases* 194(6):725–733. doi:[10.1086/506616](https://doi.org/10.1086/506616).
- Josey, K. P., S. A. Berkowitz, D. Ghosh, and S. Raghavan. 2021. Transporting experimental results with entropy balancing. *Statistics in Medicine* 40 (19):4310–4326. doi:[10.1002/sim.9031](https://doi.org/10.1002/sim.9031).
- Kennedy, E. H., S. Balakrishnan, and L. Wasserman. 2021. Semiparametric counterfactual density estimation. *Biometrika* 110 (4):875–896.
- Kooperberg, C., C. J. Stone, and Y. K. Truong. 1995. Hazard regression. *Journal of the American Statistical Association* 90 (429):78–94. doi:[10.1080/01621459.1995.10476491](https://doi.org/10.1080/01621459.1995.10476491).
- Lee, D., S. Yang, L. Dong, X. Wang, D. Zeng, and J. Cai. 2021. Improving trial generalizability using observational studies. *Biometrics Bulletin* 79 (2):1213–1225. doi:[10.1111/biom.13609](https://doi.org/10.1111/biom.13609).
- Lee, D., S. Yang, and X. Wang. 2022. Generalizable survival analysis of randomized controlled trials with observational studies. *arXiv Preprint arXiv:* 220106595.
- Manosuthi, W., L. Charoenpong, and C. Santiwarangkana. 2021. A retrospective study of survival and risk factors for mortality among people living with hiv who received antiretroviral treatment in a resource-limited setting. *AIDS Research and Therapy* 18 (1):1–10. doi:[10.1186/s12981-021-00397-1](https://doi.org/10.1186/s12981-021-00397-1).
- Pearl, J., E. Bareinboim. 2011. Transportability of causal and statistical relations: A formal approach. In 2011 IEEE 11th International Conference on Data Mining Workshops, 540–547. Vancouver, BC: IEEE Computer Society.
- 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](https://doi.org/10.1016/S0140-6736(04)17670-8).
- Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66 (5):688. doi:[10.1037/h0037350](https://doi.org/10.1037/h0037350).
- Rubin, D. B. 1986. Comment: Which ifs have causal answers. *Journal of the American Statistical Association* 81 (396):961–962. doi:[10.1080/01621459.1986.10478355](https://doi.org/10.1080/01621459.1986.10478355).
- Rudolph, K. E., and M. J. van der Laan. 2017. Robust estimation of encouragement design intervention effects transported across sites. *Journal of the Royal Statistical Society Series B, Statistical Methodology* 79 (5):1509–1525. doi:[10.1111/rssb.12213](https://doi.org/10.1111/rssb.12213).
- Sheng, A., and S. K. Ghosh. 2020. Effects of proportional hazard assumption on variable selection methods for censored data. *Statistics in Biopharmaceutical Research* 12 (2):199–209. doi:[10.1080/19466315.2019.1694578](https://doi.org/10.1080/19466315.2019.1694578).
- Stuart, E. A., S. R. Cole, C. P. Bradshaw, Leaf, P. J. 2011. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174(2):369–386. doi:[10.1111/j.1467-985X.2010.00673.x](https://doi.org/10.1111/j.1467-985X.2010.00673.x).
- Tipton, E. 2013. Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational Statistics* 38 (3):239–266. doi:[10.3102/1076998612441947](https://doi.org/10.3102/1076998612441947).
- Tsiatis, A. A. 2006. *Semiparametric theory and missing data*. New York: Springer.
- VanderWeele, T. J., and P. Ding. 2017. Sensitivity analysis in observational research: Introducing the e-value. *Annals of Internal Medicine* 167 (4):268–274. doi:[10.7326/M16-2607](https://doi.org/10.7326/M16-2607).
- 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](https://doi.org/10.1093/aje/kwx164).
- Williamson, E. J., A. Forbes, and I. R. White. 2014. Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in Medicine* 33 (5):721–737. doi:[10.1002/sim.5991](https://doi.org/10.1002/sim.5991).
- Yang, S., and J. J. Lok. 2017. Sensitivity analysis for unmeasured confounding in coarse structural nested mean models. *Statistica Sinica* 28:1703–1723. doi:[10.5705/ss.202016.0133](https://doi.org/10.5705/ss.202016.0133).
- Yang, S., Y. Zhang, G. F. Liu, and Q. Guan. 2021. Smim: A unified framework of survival sensitivity analysis using multiple imputation and martingale. *Biometrics Bulletin* 79 (1):230–240. doi:[10.1111/biom.13555](https://doi.org/10.1111/biom.13555).
- Zhang, M., and D. E. Schaubel. 2012. Contrasting treatment-specific survival using double-robust estimators. *Statistics in Medicine* 31 (30):4255–4268. doi:[10.1002/sim.5511](https://doi.org/10.1002/sim.5511).



## Appendix A. Robustness of the individual-level data generation process in Thailand and southern Ethiopia

In this section, we explore the robustness of the individual-level data generation in Thailand and southern Ethiopia based on the summary statistics in Table 1.

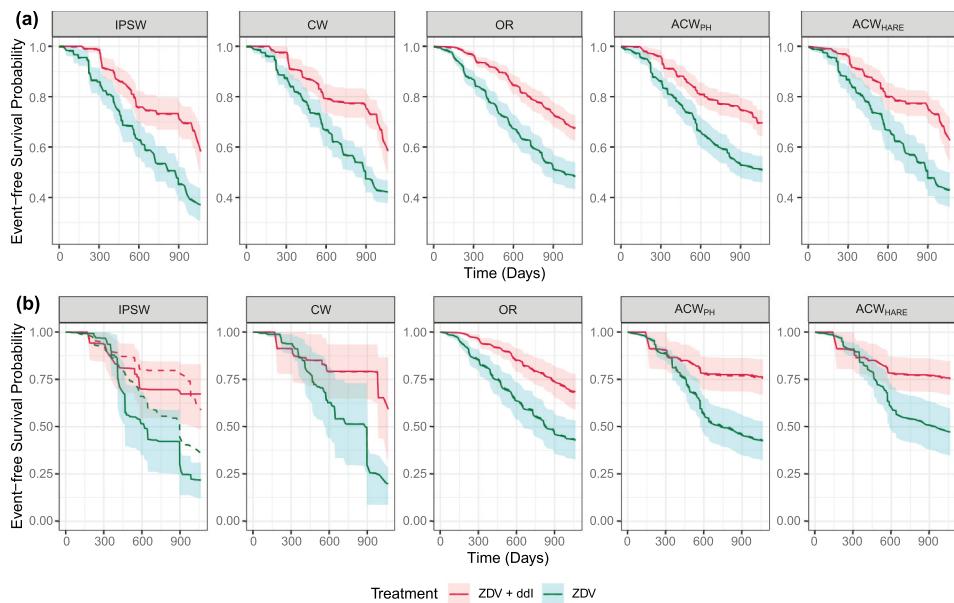
### A.1. Under the increased correlation between age and CD4 count

For the ACTG 175 trial, the Pearson's correlation between age and CD4 count is  $-0.043$  and between the age and the CD4 category is  $-0.04$ . Assuming that the trial data and the external data have the same correlation structure, we use these values to emulate the external data in Section 6.2. However, the correlation in the trial and the external data could be different; it is reasonable to assume that age and CD4 count are highly correlated as both the elderly and lower CD4 counts imply sicker patients in general. Therefore, we investigate the effect of increasing the correlation between age and CD4 count. Specifically, we emulated the external observational data with an increased correlation of  $-0.8$  instead of  $-0.04$  between the age and CD4 category for Thailand HIV patients. Similarly, for southern Ethiopia HIV patients, we emulated the individual-level data with a correlation of  $-0.8$  between the age and CD4 count.

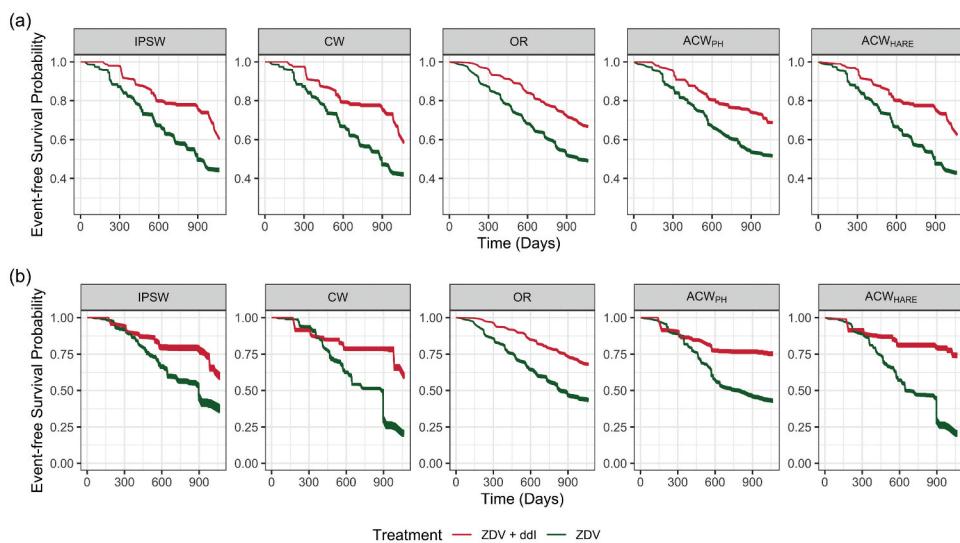
Figure A1 depicts the estimated treatment-specific survival curves with the increased correlation of  $-0.8$  between age and CD4 count/category. Dashed lines indicate the estimated treatment-specific survival curves based on the correlation structure in the ACTG 175 trial. Figure A1(a) shows that all five transport methods give an almost identical estimation of survival curves under the trial correlation of  $-0.05$  and increased correlation of  $-0.8$  for Thailand HIV patients. Similarly, Figure A1(b) shows that, except the IPSW estimator, the estimated survival curves for southern Ethiopia HIV patients under the highly negative age-CD4 count correlation are almost the same as that under the correlation in the trial data. For the IPSW estimator, the discrepancy between the different correlation structures may be due to its intrinsic weakness that it is unstable under the extreme sampling score.

### A.2. Robust to randomness in data generation

We now show that the external data generation process we used is robust to the randomness in data generation. Figure A2 plots the estimated treatment-specific event-free survival curves from 1,000 sets of emulated data based on the process described in Section 6.2 in the main text. Specifically, categorical covariates were generated from a categorical distribution based on the summary proportion, and the numerical covariates were generated from the shifted beta distribution with mean and standard deviation from the summary statistics. Even though the 1,000 emulated datasets were different from each other due to the randomness in the probability distribution, it can be seen that for all five transport methods, the estimated curves were almost identical for both Thailand and southern Ethiopia.



**Figure A1.** Estimated treatment-specific event-free survival curves with increased correlation between age and CD4 count/category; (a) estimated survival curves in Thailand HIV patients; (b) estimated survival curves in southern Ethiopia HIV patients.



**Figure A2.** Estimated treatment-specific event-free survival curves from 1,000 sets of emulated data; (a) Thailand; (b) southern Ethiopia.



## Appendix B. Effect of two other treatments - ZDV + ZAL and ddI - over ZDV

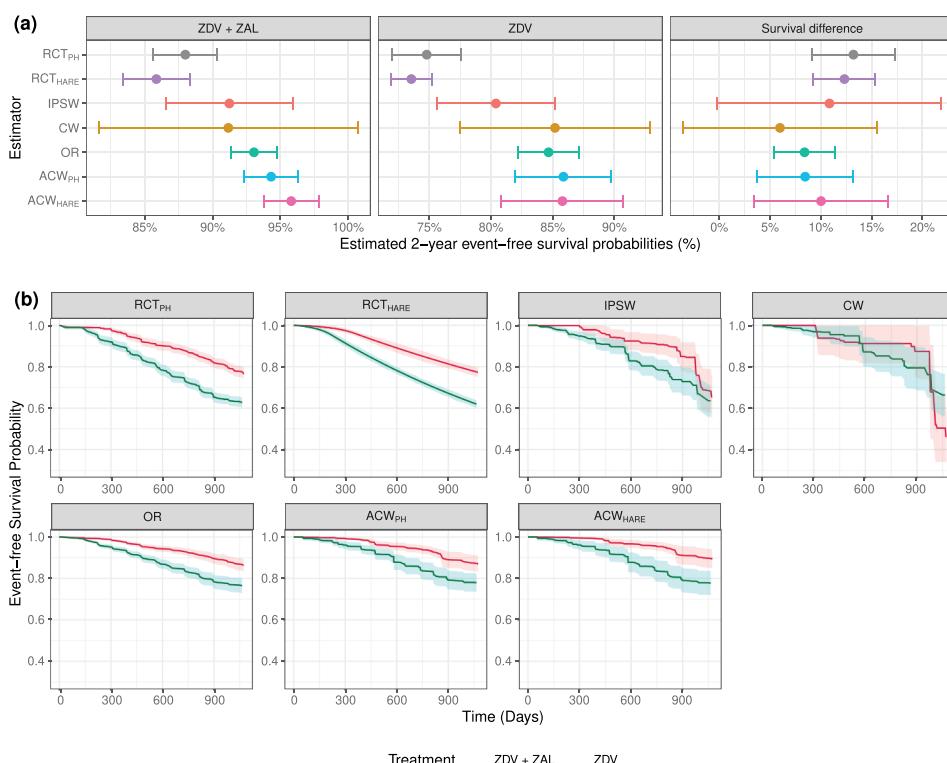
In the ACTG 175 trial, enrolled patients were randomly assigned to the four treatments, ZDV + ddI, ZDV + ZAL, ddI, and ZDV. In the main text, we focus on the effect of ZDV + ddI combination therapy over ZDV for illustrative purpose. This section provides the results of transporting the effect of two other treatments, i.e., ZDV + ZAL and ddI, over ZDV to the external target populations.

### B.1 Transporting the effect of ZDV + ZAL over ZDV

In this section, we consider ZDV + ZAL and ZDV as binary treatment, which consists of 524 ZDV + ZAL patients and 532 ZDV monotherapy patients. Similar to the analysis in the main text, the primary endpoint is the progression of HIV disease, defined as a more than 50% decline in the CD4 cell count or development of the acquired immunodeficiency syndrome, or death. The causal estimand of interest is a 2-year event-free survival difference between ZDV + ZAL and ZDV monotherapy, and about 73% of the survival times were right-censored.

**Figure B1** plots estimated treatment-specific event-free survival probabilities in the US early-stage HIV patients. **Figure B1(a)** shows that the transported 2-year survival probabilities in the US early-stage HIV patients population are higher for both treatment groups, and their differences are smaller in general than those estimated from the ACTG 175 trial. For instance, the estimators based only on the trial data, i.e.,  $\text{RCT}_{\text{PH}}$  and  $\text{RCT}_{\text{HARE}}$ , give an estimate of a 13% higher survival probability for ZDV + ZAL over ZDV monotherapy. On the other hand, the transport methods estimate that the 2-year survival differences in the US early-stage HIV patients are 8%-10%. **Figure B1(b)** illustrates that after transporting to the US early-stage HIV patients, survival curves for both treatment groups are higher than those in the ACTG 175 trial participants. These results suggest that as the US early-stage patients are healthier than the ACTG-175 trial participants, the effect of the combined therapy over the ZDV monotherapy is less significant in the target population, possibly due to the toxicity and the low compliance. There is no notable difference between the estimated survival probabilities from the  $\text{ACW}_{\text{HARE}}$  estimator and the estimators with the PH assumption, i.e.,  $\text{OR}_{\text{PH}}$ ,  $\text{ACW}_{\text{PH}}$ .

For Thailand and southern Ethiopia HIV patients, we emulated the external data using the same data generation process described in the main text. **Figures B2** and **Figure B3** depict the estimated survival probabilities for the ZDV + ZAL and ZDV treatment groups in Thailand and southern Ethiopia HIV patients, respectively. As patients in both target



**Figure B1.** Estimated treatment-specific event-free survival probabilities in the US early-stage HIV patients for ZDV + ZAL vs. ZDV; (a) 2-year event-free survival probabilities; (b) treatment-specific survival curves.

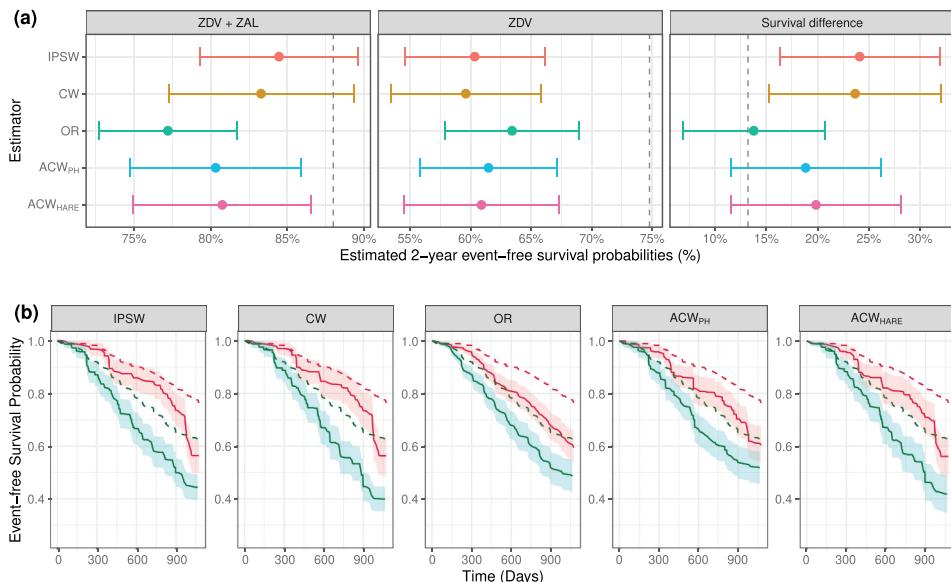
populations are considered to be sicker than the patients in the ACTG 175 trial, the transported treatment-specific survival curves are lower than the survival curves estimated from the trial data, depicted in dashed lines. According to Figure B2(b), it can be seen that the estimated survival curves by  $\text{ACW}_{\text{HARE}}$  are steeper than those by the  $\text{OR}_{\text{PH}}$  and  $\text{ACW}_{\text{PH}}$  estimators which depend on the PH assumption. The former estimator is robust to the violation of the PH assumption whereas the latter estimators are not, which may result in the overestimation of the survival probabilities. Figure B3(b) shows that there is no significant effect of the ZDV + ZAL combined therapy over ZDV in southern Ethiopia HIV patients. However, this result may not be meaningful due to the large variability of the transported probabilities.

## B.2. Transporting the effect of ddI over ZDV

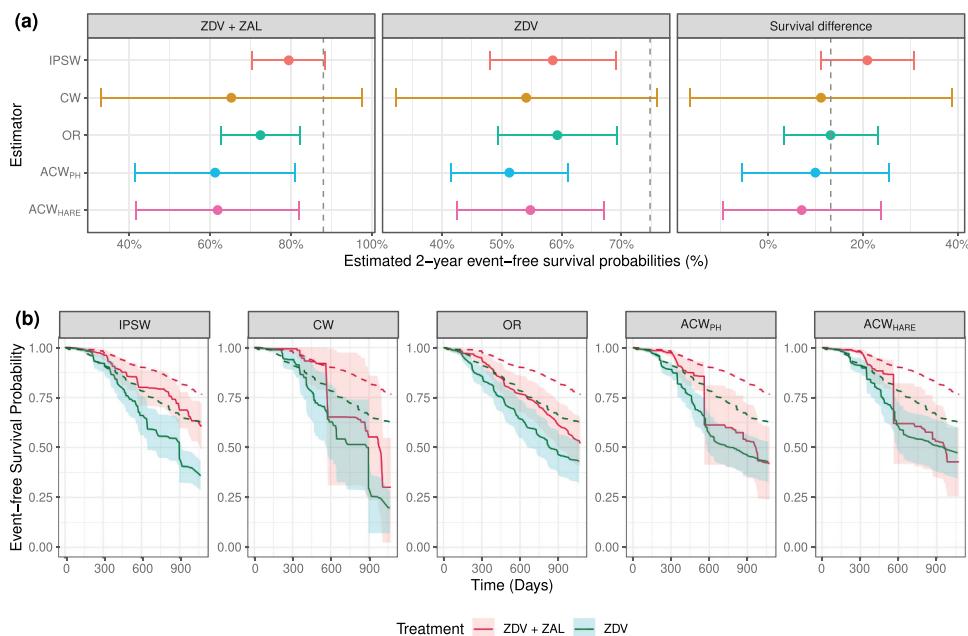
This section considers ddI and ZDV as binary treatment, consisting of 561 ddI patients and 532 ZDV monotherapy patients. The analyses are similar to those of ZDV + ddI vs. ZDV in the main text and ZDV + ZAL vs. ZDV in Appendix B1. About 72% of the survival times were right-censored.

Figure B4 shows that the estimated survival probabilities for both the ddI and ZDV treatment groups in the US early-stage HIV patients population are higher than those in the ACTG 175 trial as the former is healthier. The OR,  $\text{RCT}_{\text{PH}}$ , and  $\text{RCT}_{\text{HARE}}$  estimators suggest that the effect of ddI over ZDV is smaller in the target population. This result contradicts the estimated survival probabilities from the IPSW and CW estimators which is larger than the one in the trial data. The estimated survival probabilities from the  $\text{ACW}_{\text{HARE}}$  estimator and that from the estimators with the PH assumption, i.e.,  $\text{OR}_{\text{PH}}$ ,  $\text{ACW}_{\text{PH}}$ , were found to be similar.

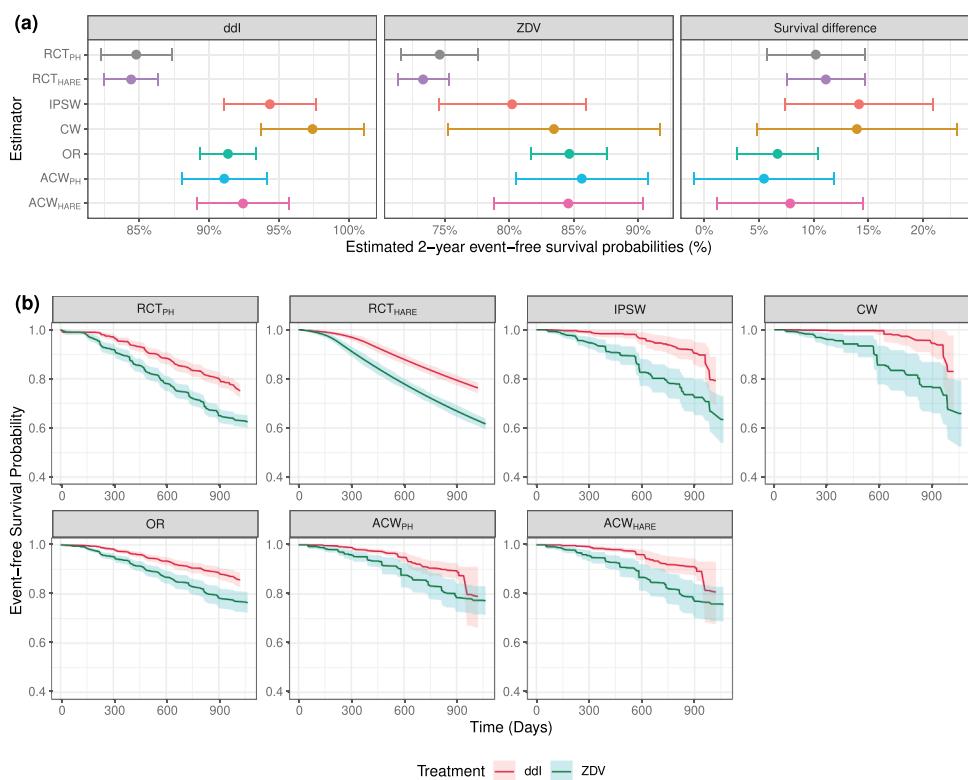
The transported results of ddI vs. ZDV for Thailand and southern Ethiopia HIV patients are parallel to those of other treatment sets; the estimated treatment-specific survival curves in the target populations are lower than the estimated survival curves in the trial data as the target populations include less healthy patients. Under the same external data generation process described in the main text, Figures B5 and B6 show the estimated survival probabilities for the ddI and ZDV treatment groups in Thailand and southern Ethiopia HIV patients, respectively, which are lower than the survival curves estimated from the trial data, depicted in dashed lines. According to Figure B5(b), the estimated survival curves by  $\text{ACW}_{\text{HARE}}$  were found to be steeper than those by the  $\text{OR}_{\text{PH}}$  and  $\text{ACW}_{\text{PH}}$  estimators which depend on the PH assumption. As the  $\text{OR}_{\text{PH}}$  and  $\text{ACW}_{\text{PH}}$  estimators can not adjust for the violation of the PH assumption, the survival probabilities might be overestimated. In Figure B6(a), all estimators except the OR estimator estimate the transported 2-year survival difference in the southern Ethiopia HIV patients to be higher than that in the trial data. According to Figure B6(b), all transport methods besides the OR estimator show a delayed treatment effect and crossing of the transported survival curves for ddI and ZDV.



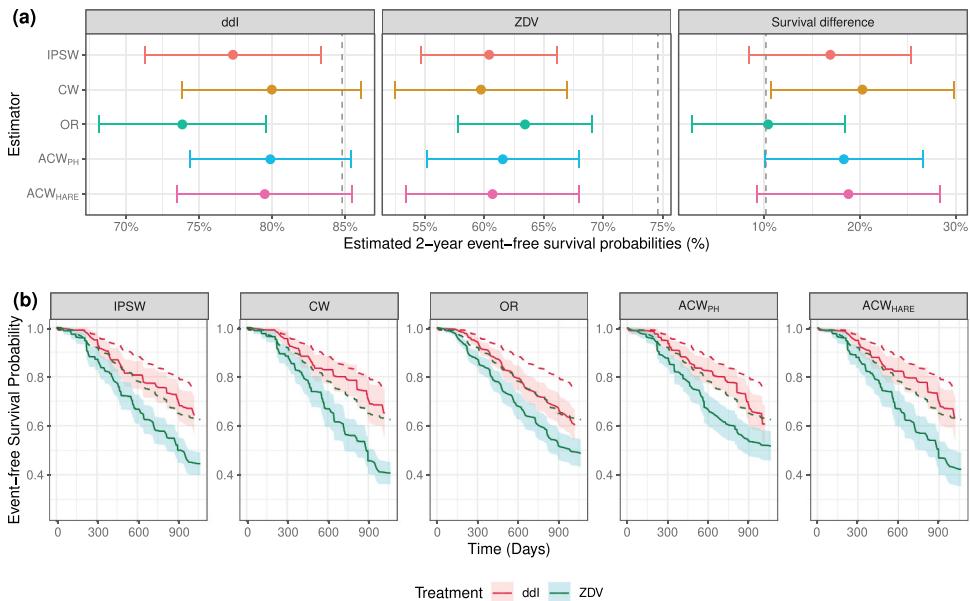
**Figure B2.** Estimated treatment-specific event-free survival probabilities in Thailand HIV patients for ZDV + ZAL vs. ZDV; (a) 2-year event-free survival probabilities; (b) treatment-specific survival curves. Dashed lines indicate the estimates from the  $\text{RCT}_{\text{PH}}$  model.



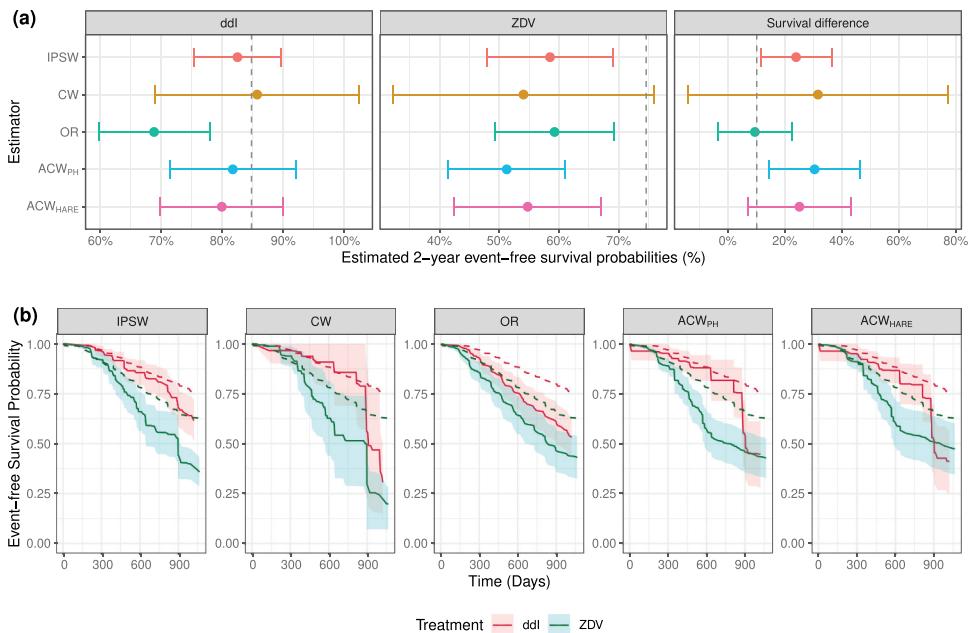
**Figure B3.** Estimated treatment-specific event-free survival probabilities in southern Ethiopia HIV patients for ZDV + ZAL vs. ZDV; (a) 2-year event-free survival probabilities; (b) treatment-specific survival curves. Dashed lines indicate the estimates from the RCT<sub>PH</sub> model.



**Figure B4.** Estimated treatment-specific event-free survival probabilities in the US early-stage HIV patients for ddl vs. ZDV; (a) 2-year event-free survival probabilities; (b) treatment-specific survival curves.



**Figure B5.** Estimated treatment-specific event-free survival probabilities in Thailand HIV patients for ddl vs. ZDV; (a) 2-year event-free survival probabilities; (b) treatment-specific survival curves. Dashed lines indicate the estimates from the  $\text{RCT}_{\text{PH}}$  model.



**Figure B6.** Estimated treatment-specific event-free survival probabilities in southern Ethiopia HIV patients for ddl vs. ZDV; (a) 2-year event-free survival probabilities; (b) treatment-specific survival curves. Dashed lines indicate the estimates from the  $\text{RCT}_{\text{PH}}$  model.