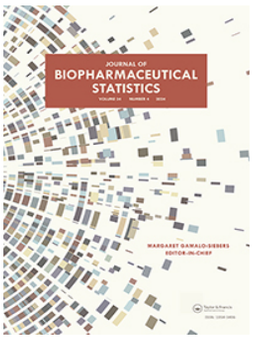


Journal of Biopharmaceutical Statistics

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

Response to comment on "Transporting survival of an HIV clinical trial to the external target populations by Lee et al. (2024)"

Shu Yang & Xiang Zhang

To cite this article: Shu Yang & Xiang Zhang (08 Jul 2024): Response to comment on "Transporting survival of an HIV clinical trial to the external target populations by Lee et al. (2024)", Journal of Biopharmaceutical Statistics, DOI: 10.1080/10543406.2024.2373449

To link to this article: https://doi.org/10.1080/10543406.2024.2373449

Published online: 08 Jul 2024.

JOURNAL OF BIOPHARMACEUTICAL STATISTICS
https://doi.org/10.1080/10543406.2024.2373449

Taylor & Francis
Taylor & Francis Group

Response to comment on "Transporting survival of an HIV clinical trial to the external target populations by Lee et al. (2024)"

1. Responses

Thank the author of the letter and the editor for sharing the ideas and providing insightful comments. We have structured my responses as follows: motivation of treatment effect transportability and impact on clinical practice decision-making (Section 2), statistical considerations (Section 3), feasibility and assumptions required for successful transportability (Section 4), and ethical and fairness considerations (Section 5).

2. Motivation of treatment effects transportability

To evaluate the effects of transportability approaches on patient outcomes and healthcare decision-making, we must consider their application in real clinical settings. We start by asking, "Why transport?" and examining what constitutes the underlying population in a randomized clinical trial (RCT). Typically, the underlying population is not explicitly defined. Can it be delineated by the inclusion/exclusion criteria? Is the RCT population representative of the real-world patient population in clinical practice? The answer is generally no. This discrepancy arises because RCT populations often undergo a consensus process or are enriched to enhance trial efficiency, such as selecting individuals with a high disease risk (Averitt et al. 2020). Furthermore, what if our target population is geographically outside the RCT setting?

It is well known that RCTs exhibit strong internal validity but often lack external validity (i.e., generalizability or transportability; Colnet et al. 2024; Yang and Wang 2022). We will use "generalizability" and "transportability" interchangeably, although there are subtle differences between the two concepts (Colnet et al. 2024). Directly applying RCT results to actual clinical settings may result in the treatment not performing as anticipated. Below, we provide examples where generalizability and transportability are important for patient outcomes and healthcare decision-making.

Example 1 The motivating application in Lee et al. (2022, 2023). Lee et al. (2024) aims to generalize the treatment effect from the CALGB 9633 trial – a randomized phase III trial assessing the efficacy of adjuvant chemotherapy compared to observation in stage 1B non-small cell lung cancer – to a real-world patient population in clinical practice. We considered a representative real-world patient population from the National Cancer Data Base (NCDB), an oncology outcomes database that collects information on 72% of all new invasive cancer diagnoses in the U.S. Despite using the RCT's eligibility criteria to define the target population, there are substantial differences between the patient populations represented by CALGB 9633 and the NCDB sample. Patients in randomized clinical trials tend to be younger, healthier, and have less severe disease status. This discrepancy raises an important question: Can the benefits of adjuvant chemotherapy observed in CALGB 9633 be replicated in the target population represented by NCDB patients?

Example 2 The motivating application in Lee et al. (2024) aims to transport the treatment effect for survival from the AIDS Clinical Trials Group (ACTG) 175 trial, which focused on intermediate-stage disease patients in the US. We considered three external target populations: US early-stage HIV patients, HIV patients in Thailand, and HIV patients in southern Ethiopia. These populations have patient

characteristics that differ significantly from those in the ACTG 175 trial. Consequently, the treatment effect observed in the ACTG 175 trial is likely to differ from the effects in these target populations.

Example 3 Beyond regulatory approval, the concept of health technology assessment (HTA) is widely accepted and implemented by many countries as the cornerstone for decisions regarding market access and reimbursement, as it guides policymakers on the optimal allocation of limited healthcare resources. Because of that, various government agencies (e.g., Gemeinsamer Bundesausschuss, or G-BA in Germany, The National Institute for Health and Care Excellence, or NICE in United Kingdom) and non-profit organizations (e.g., institute for comparative effectiveness research, or ICER in United States) provides scientific guidance on the approaches in generating evidence for HTAs, and generalizability issue of RCT evidence is referred in those guidance documents (Institut fur Qualitat und Wirtschaftlichkeit im Gesundheitswesen 2022; NICE 2022), further demonstrate the importance of generalizing RCT evidence to wider target population in routine clinical practice under each specific country/region's healthcare system.

Evaluating the treatment effect in various populations is crucial, as they offer insights into different aspects of treatment effects, thereby assisting regulatory agencies in evaluating a drug's efficacy and safety during the approval process.

3. Statistical framework relaxing the proportional hazard assumption

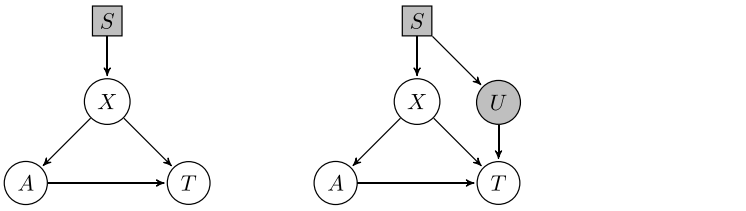
How do we address the problem of external validity? The proposed "genRCT" framework (Lee et al. 2022, 2023, 2024) is an analytical approach designed to address the issue of external validity in RCTs and to generalize or transport trial results to the target population of interest.

We develop a doubly robust estimator, called the ACW (Augmented Calibration Weighting) estimator, leveraging outcome regression and propensity score modeling. The ACW estimator is designed to be consistent for the population treatment effect even if only the outcome regression or propensity score is well approximated, though not necessarily both. The original ACW estimator, proposed by Lee et al. (2023) for continuous outcomes, was extended by Lee et al. (2022) for survival outcomes. In this latter work, Cox proportional hazard regression was used to estimate the survival outcome mean function in ACW, with an acknowledgment of its potential limitations in certain scenarios. To address these limitations, Lee et al. (2024) introduced the HARE (HAzard Regression Estimator) in ACW as an alternative to Cox proportional hazard regression. HARE, based on linear splines and their tensor products, does not rely on the proportional hazard assumption (Kooperberg et al. 1995). Specifically, HARE considers splines of the covariates and time, along with the pairwise interactions between covariates and time, and performs variable selection simultaneously.

While HARE relaxes the Cox proportional hazard assumption, it inherits practical challenges associated with spline approximation, such as the placement and number of knots, and the sufficiency of pairwise interactions between covariates and time. Further investigation is necessary to assess its finite-sample performance. Other potential alternatives to HARE include the additive hazards model, survival random forest and many other machine learning estimators. Integrating flexible models within the ACW framework shows promise, given that ACW is rate doubly robust. The cross-fitting strategy of Chernozhukov et al. (2018) can be shown to be valid for ACW, enabling weaker conditions on the estimators for the nuisance parameters, seamless incorporation of flexible machine learning estimators, and a straightforward variance estimator. Although Lee et al. (2024) did not implement it in this manner, future work to demonstrate its potential is worthwhile.

Table 1. Causal assumptions for successful generalizability and transportability.

Assumptions	Implications in practice
1. Stable unit treatment value assumption	The versions of the treatment in both the trial and target populations are identical. Additionally, interference structures, if present, must be the same between the trial and target populations.
2. Ignorability and positivity of trial treatment assignment	The trial treatment is randomized, and all patients in the trial have positive probabilities of receiving both the control and active treatments. This is typically achieved through a well-controlled trial design.
3. Censoring ignorability and positivity of uncensored probability	The censoring mechanism can be fully explained by the observed covariates.
4. Conditional survival transportability	The covariate set captures all confounding variables that exhibit differing distributions between the trial and target populations and are associated with the outcomes.
5. Positivity of trial participation	All patients in the target population have positive probabilities of participating in the trial study.



notation: S: study participation indicator; A : treatment assignment; X: observed covariates; U: unobserved covariates; and T : survival outcome of interest

Figure 1. Selection diagrams illustrating Assumption 4 (conditional survival transportability). White nodes represent observed variables, while the dark node represents an unmeasured variable. In both diagrams, the two populations differ by covariate distributions, as indicated by *S* pointing to *X*. Assumption 4 is satisfied in the left diagram, but not in the right diagram, where *S* points to *U* and *U* points to *T*.

4. Causal assumptions for successful transportability

The analytic framework we provided is not a panacea. It is important to understand the feasibility and the assumptions necessary for successful transportability. Table 1 outlines key causal assumptions, and any violation of these assumptions can hinder the success of treatment generalizability and transportability. Notably, some assumptions, such as Assumption 4, are not testable with observed data. For Assumption 4 to be valid, the covariate set must encompass all confounding variables that have differing distributions between the trial and target populations and are associated with outcomes. The causal diagrams in Figure 1 help assess the plausibility of this assumption based on domain knowledge. For example, in Lee et al. (2024), patient characteristics are identified and adjusted for as potential sources of variation between the RCT and external populations. However, as the author of the letter indicated, other factors like access to resources, cultural differences, and health systems may also influence treatment success. If this is the case and these critical factors are unmeasured, Assumption 4 is violated, leading to biases in treatment transportability.

Given that Assumption 4 is not testable, sensitivity analysis should be conducted to determine whether the study's conclusions are sensitive to the violation of Assumption 4 and the amount of unmeasured confounding affects the estimated population treatment effect. For further examples, refer to our recent work (Jiang et al. 2024) and the referenced works (Nguyen et al. 2017, 2018).

5. Ethical and fairness considerations

It is imperative to address moral and fairness issues with ethical considerations in healthcare research. Using real-world evidence alongside RCTs can enhance the external validity of study findings. For

example, combining data from electronic health records with RCT data can provide insights into how treatments perform in diverse, real-world populations. However, researchers must be cautious when extrapolating RCT results to broader populations to avoid exacerbating existing inequities in healthcare. As the author of the letter indicated, the specific needs and circumstances of various patient groups, including differences in health conditions, socioeconomic statuses, cultural backgrounds, and levels of healthcare access, must be considered. For example, elderly patients often have weaker immune systems and are more vulnerable. Transporting RCT results from a young population to an elderly population without considering these differences can lead to inappropriate treatment and adverse effects.

To mitigate inequities, it is crucial to design fairness-aware studies and transportability techniques. Implementing stratified clinical trials ensures that various demographic groups (e.g., gender, ethnicity, socioeconomic status) are adequately represented and analyzed, thereby promoting fairness. Transportability techniques must account for *individual variation* and *fairness*, which will be discussed separately below.

Most studies on treatment effect transportability have focused on transporting the average treatment effect (ATE) from RCTs. However, ATE may not be sufficiently insightful for individual patient treatment, as different patients may respond differently to the same treatment. Precision medicine aims to estimate the heterogeneous treatment effect or determine the optimal individualized treatment rule (ITR), tailoring treatment recommendations to patients based on their individual characteristics, such as age, gender, and clinical history. Wu and Yang (2023) and Zhao et al. (2023) have proposed statistical methods for transferring ITR learned from RCTs to a target population for continuous outcomes and right-censored survival outcomes, respectively. Chu et al. (2023) extended this approach to scenarios where only summary statistics from the target population are available, addressing privacy and confidentiality concerns.

Ongoing research is needed to refine ITR learning techniques, ensuring they are both reliable and suitable for diverse clinical settings. Zhao et al. (2024) considered fair policy learning tasks as a constrained optimization problem under the Demographic Parity (DP) (Calders et al. 2009) or Equal Opportunity (EO) (Hardt et al. 2016) metrics. DP requires that the predicted positive rate be the same across different sensitive groups, while EO focuses on promoting equal predicted positive rates for true positives in different sensitive groups. Since the optimal policy is typically defined as the maximizer of the expected potential outcome over the entire population, it may be suboptimal or even detrimental to certain disadvantaged subgroups. Fang et al. (2023) proposed the fairness-oriented optimal policy learning framework to estimate an optimal ITR that maximizes the average value while ensuring its tail performance exceeds a prespecified threshold (Protect the Vulnerable, PV). Other examples in the literature include individual fairness, which demands that similar individuals be treated similarly (Dwork et al. 2012), principal fairness, which incorporates causality into fairness by ensuring non-discrimination among individuals similarly affected by the decision (Imai and Jiang 2023), and the counterfactual no-harm criterion by the principal stratification method (Li et al. 2023).

Further studies on transporting optimal ITRs under various fairness constraints are needed to improve the reliability and suitability of transportability techniques for individual treatment decision-making in clinical settings.

Disclosure statement

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

Funding

The work was supported by the National Science Foundation [SES 2242776]; National Institutes of Health [1R01AG066883].

References

Averitt, A. J., C. Weng, P. Ryan, and A. Perotte. 2020. Translating evidence into practice: Eligibility criteria fail to eliminate clinically significant differences between real-world and study populations. *NPJ Digital Medicine* 3 (1):67. 2. doi: 10.1038/s41746-020-0277-8.

Calders, T., F. Kamiran, and M. Pechenizkiy. 2009. Building classifi s with independency constraints. In *2009 IEEE international conference on data mining workshops*, 13–18, IEEE.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21 (1):C1–C68. doi: 10.1111/ectj.12097.

Chu, J., W. Lu, and S. Yang. 2023. Targeted optimal treatment regime learning using summary statistics. *Biometrika* 110 (4):913–931. doi: 10.1093/biomet/asad020.

Colnet, B., I. Mayer, G. Chen, A. Dieng, R. Li, G. Varoquaux, J.-P. Vert, J. Josse, and S. Yang. 2024. Causal inference methods for combining randomized trials and observational studies: A review. *Statistical Science* 39:165–191. doi: 10.1214/23-STS889.

Dwork, C., M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226. ITCS 2012 Cambridge, MA USA.

Fang, E. X., Z. Wang, and L. Wang. 2023. Fairness-oriented learning for optimal individualized treatment rules. *Journal of the American Statistical Association* 118 (543):1733–1746. 5. doi: 10.1080/01621459.2021.2008402.

Hardt, M., E. Price, and N. Srebro. 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems* 29:3323–3331.

Imai, K., and Z. Jiang. 2023. Principal fairness for human and algorithmic decision-making. *Statistical Science* 38 (2):317–328. 5. doi: 10.1214/22-STS872.

Institut fur Qualitat und Wirtschaftlichkeit im Gesundheitswesen. 2022. Allgemeine methoden: Version 6.0. https://www.iqwig.de/methoden/general-methodsversion−6−0.pdf. *Online*.

Jiang, K., X.-X. Lai, S. Yang, Y. Gao, and X.-H. Zhou. 2024. A practical analysis procedure on generalizing comparative eff eness in the randomized clinical trial to the real-world trial eligible population. *arXiv preprint arXiv:2406.04107*.

Kooperberg, C., C. J. Stone, and Y. K. Truong. 1995. Hazard regression. *Journal of the American Statistical Association* 90 (429):78–94. 3. doi: 10.1080/01621459.1995.10476491.

Lee, D., C. Gao, S. Ghosh, and S. Yang. 2024. Transporting survival of an hiv clinical trial to the external target populations. *Journal of Biopharmaceutical Statistics* 1–22. doi: 10.1080/10543406.2024.2330216.

Lee, D., S. Yang, M. Berry, T. Stinchcombe, H. J. Cohen, and X. Wang. 2024. genrct: A statistical analysis framework for generalizing rct findings to real-world population. *Journal of Biopharmaceutical Statistics*. doi: 10.1080/10543406.2024.2333136.

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.

Li, H., C. Zheng, Y. Cao, Z. Geng, Y. Liu, and P. Wu. 2023. Trustworthy policy learning under the counterfactual no-harm criterion. *International Conference on Machine Learning* 202:20575–20598.

Nguyen, T. Q., B. Ackerman, I. Schmid, S. R. Cole, E. A. Stuart, and N. Mitra. 2018. Sensitivity analyses for effect modifiers not observed in the target population when generalizing treatment effects from a randomized controlled trial: Assumptions, models, effect scales, data scenarios, and implementation details. *PLOS ONE* 13 (12):e0208795. doi: 10.1371/journal.pone.0208795.

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 eff. *The Annals of Applied Statistics* 11 (1):225–247. doi: 10.1214/16-AOAS1001.

NICE. 2022. Nice Health Technology Evaluations: The Manual. *Process and methods [PMG36]*. 3.

Wu, L., and S. Yang. 2023. Transfer learning of individualized treatment rules from experimental to real-world data. *Journal of Computation and Graphical Statistics* 32 (3):1036–1045. doi: 10.1080/10618600.2022.2141752.

Yang, S., and X. Wang. 2022. Rwd-integrated randomized clinical trial analysis. *Biopharmaceutical Report* 29 (2):15–21. doi: 10.1111/biom.13609.

Zhao, P., A. Chambaz, J. Josse, and S. Yang. 2024. Positivity-free policy learning with observational data. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024*, Valencia, Spain. 238:1918–1926. PMLR.

Zhao, P., J. Josse, and S. Yang. 2023. Efficient and robust transfer learning of optimal individualized treatment regimes with right-censored survival data. *arXiv preprint arXiv:2301.05491*.

Shu Yang
Department of Statistics, North Carolina State University, Raleigh, USA
syang24@ncsu.edu http://orcid.org/0000-0001-7703-707X

Xiang Zhang
CSL Behring