Some statistical methods to deal with patient heterogeneity in clinical trial design and causal inference

by

Yi Chen

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Biomedical Data Science)

at the

UNIVERSITY OF WISCONSIN–MADISON

2025

Date of final oral examination: 05/01/2025

The dissertation is approved by the following members of the Final Oral Committee:

Menggang Yu, Professor, Biostatistics, University of Michigan

Guanhua Chen, Associate Professor, Biostatistics and Medical Informatics, UW-Madison

Richard J. Chappell, Professor, Statistics, UW-Madison

Lu Mao, Associate Professor, Biostatistics and Medical Informatics, UW-Madison

ACKNOWLEDGMENTS

The past seven years of work and study at the University of Wisconsin-Madison have been both challenging and rewarding. It would be impossible to reach this milestone without the support, guidance and encouragement of many people who have generously shared their knowledge, time, and kindness along the way.

First and foremost, I would like to express my heartfelt gratitude to my advisor, Dr. Menggang Yu, for his support throughout my professional and academic journey. I was fortunate to start my work under Dr. Yu's supervision, gaining valuable experience through statistical consulting for the Biostatistics Shared Resource and the Health Innovation Program at UW-Madison. These experiences enhanced my ability to provide effective consulting to collaborators, and introduced me to the fields of clinical trials and causal inference, which inspired my decision to pursue a Ph.D. study. Dr. Yu has also been a dedicated mentor for my Ph.D. study, offering valuable guidance and insightful advice, which have a profound and lasting impact. His mentorship was instrumental in expanding my knowledge and equipping me to contribute to meaningful research, ultimately leading to the publication of an award-winning paper. Also, I would like to extend my sincere appreciation to my co-advisor Dr. Guanhua Chen, who closely supervised me on the causal generalization project. I have greatly benefited from our discussions and his thoughtful and detailed

feedback. His mentorship has been essential in my academic journey, and I am truly grateful for the opportunity to learn from him.

Additionally, I am grateful to my committee members, Dr. Richard Chappell and Dr. Lu Mao, for their invaluable feedback and suggestions, which have greatly enhanced the quality of this dissertation. I am deeply grateful to Dr. Chappell for his insightful teaching in the applied biostatistical methods class and his patient explanations during office hours, which greatly enhanced my understanding of many statistical concepts. Our discussions sparked numerous innovative ideas with potential applications in clinical trials. Dr. Chappell also serves on my mentoring committee, and I found our discussion on developing a plan for my professional and academic development to be extremely valuable and insightful. I also want to thank Dr. Mao for the knowledge I gained from his survival analysis class and his supervision for the INVESTED trial. His encouragement of my research, as well as sponsoring my trip to attend conference and chair his organized sessions, has greatly enriched my academic experience.

Moreover, my sincere thanks go to other professors, colleagues and collaborators in BMI and SMPH, for their tremendous support during my professional and academic journey. I am particularly grateful to Dr. Christina Kendziorski and Dr. Yeonhee Park, with whom I completed two rotation research projects, both of which significantly contributed to my growth as a researcher. Special thanks to Dr. Thomas Cook, whose class provided me with a solid foundation in statistical methodologies in clini-

cal trials. I also appreciate Dr. Christie Bartels as I collaborated with her on several projects related to rheumatological and inflammatory disease studies. Working with Dr. Bartels not only strengthened my research skills but also enhanced my understanding of interdisciplinary collaboration. In addition, I would like to express my sincere gratitude to Dr. Roxana Alexandridis and Dr. Kyungmann Kim for their mentorship and supervision during my time as a Biostatistician at BMI. Their mentorship enabled me to further develop my expertise in statistical consulting and clinical trial practice, providing me with essential skills that will continue to shape my career.

Furthermore, I am deeply thankful to my friends in Madison, who have made my years in Madison both joyful and memorable. A special thank you to Zhan Ma for his unconditional love, companionship, and support in every aspect of my life over the past years.

Finally, I am profoundly grateful to my family: my parents and grandparents. Thank you for your sacrifices, your patience, and your love, which have been the foundation of my personal growth. This achievement wouldn't have been possible without you by my side.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Heterogeneous patient data offers both unique opportunities and challenges in biomedical research. Effectively integrating this information while addressing the complexities it introduces is a crucial area of study. My research focuses on two scenarios where heterogeneous patient data are encountered, with the goal to develop methods that improve the reliability and efficiency of the statistical analysis. The first scenario is about incorporating patients' natural ordering information in randomized phase II studies. The exploratory nature of phase II trials makes it quite common to include heterogeneous patient subgroups with different prognoses in the same trial. Incorporating such patient heterogeneity or stratification into statistical calculation can improve efficiency and reduce sample sizes in single-arm phase II trials with binary outcomes. However, such consideration is lacking in randomized phase II trials. In Chapter 1, we propose methods that can utilize some natural order information which may exist in stratified population to gain statistical efficiency for randomized phase II designs. We consider both binary and time-to-event outcomes in our development. Compared with methods that do not use ordering information, our method is shown to improve the probabilities of correct selection and reduce sample size in our simulation and real examples. We also developed its related R package *constrselect* and we discuss its key functions and implementation in Chapter 2.

The second scenario addresses the problem of causal generalization, where differences in the distribution of treatment effect modifiers across populations, known as covariate shift, can result in varying ATEs.Chen et al. [2023] introduced a weighting method to estimate the target ATE using only summary-level information from a target sample while accounting for the possible covariate shifts. However, the asymptotic variance of the estimate was shown to depend on individual-level data from the target sample, hindering statistical inference. In Chapter 3, we propose a resampling-based perturbation method for confidence interval construction for the estimated target ATE, utilizing additional summary-level information. We demonstrate the effectiveness of our approach through simulation and real data settings. We also developed its related R package *EBalGen* and we discuss its key functions and implementation in Chapter 4.

1 RANDOMIZED PHASE II DESIGN WITH ORDER CONSTRAINED STRATA

1.1 Introduction

The rapid change in the therapeutic landscape and medical technology, especially in the field of oncology, makes the randomized phase II trial a popular choice, as it assures better patient comparability, reduces confounding factors, and synchronizes data capture. Indeed, a simple search on the clinicaltrials.gov website leads to about 4,000 registered randomized phase II trials in the last 10 years. Among them, 938 are actively recruiting patients as of today.

In two excellent review articles [Rubinstein et al., 2009, Sharma et al., 2011], many advantages and disadvantages are discussed for randomized phase II trials. One main disadvantage is that the implementation of randomized designs generally requires much more patients than traditional single-arm trials comparing with historical controls under similar settings [Rubinstein et al., 2009]. Our method intends to deal with this issue for randomized phase II trials that include stratification of patients.

Due to disease heterogeneity among patients, patients often have different prognostic factors and thus could be stratified into groups for randomization. For example, it is quite common to have different stage cancer patients (e.g. stages I and II or stages II and III) in the same trial. One

of our motivating examples is a public phase II stratified clinical trial targeting cisplatin-ineligible patients with metastatic urothelial cancer (clinicaltrial.gov Identifier: NCT03451331). The goal is to investigate the effect of adding carboplatin versus oxaliplatin to existing treatment regimen: gemcitabine and nivolumab. The randomization of this study is stratified on the lymph node (LN) only metastasis status since LN only metastasis patients are expected to have higher response rates.

Another motivating example comes from a phase II stratified clinical trial which is still under development. The study targets patients with early-stage triple-negative breast cancer who have completed neoadjuvant therapy and have residual disease. The researchers want to investigate the effect of adding Sacituzumab Govitecan versus Capecitabine to the current single-agent treatment of Pembrolizumab. The stratification of this study is based on nodal status as nodal negative patients are expected to have higher event-free survival (EFS) rates.

Indeed, patient heterogeneity has been long recognized in single-arm phase II studies [Thall et al., 2003, London and Chang, 2005, Wathen et al., 2008, Jung et al., 2012]. Incorporating patient stratification into trial designs has demonstrated improvement in statistical properties including improved efficiency and reduced sample sizes for binary outcomes [Chang et al., 2012, 2011, Sposto and Gaynon, 2009, Xu et al., 2020]. However, such consideration is lacking in randomized phase II trials.

Two main randomized phase II comparative designs are selection and

screening designs. The essential difference between the two is that the selection design does not include a control arm but the screening design does. Therefore the two designs recommend promising investigational agents for further phase III studies based on different logic. The screening design compares investigational agents to the control and screens out those non-promising agents. Traditional hypothesis testing based methods are used to determine trial sample sizes. To ensure feasible sample sizes, quite liberal type I and type II errors are used.

On the other hand, the selection design focuses on “picking a winner" from a pool of testing agents. The design is also sometimes known as the “pick a winner" design. For example, the test agents may have already demonstrated activity in limited scenarios, or they may be combination drugs with new agents added to known active treatments [Liu et al., 1993]. In this type of design, it is not essential that the very best treatment is definitely selected, since we could only make this decision after doing a formal phase III trial. Rather, this design ensures that a substantially inferior treatment will not be selected when a superior treatment exists [Sargent and Goldberg, 2001].

The goal of our method is to utilize some natural order constraints that may exist in stratified population to gain statistical efficiency for randomized phase II designs. Our idea is applicable for both screening and selection designs. However, for thoroughness and simplicity, we focus on the selection design in this chapter, and put our application on screen-

ing design in the appendix. We demonstrate that utilizing such order information is particularly useful in early-phase clinical trials, especially when we want to gain more statistical efficiency and reduce sample sizes. A fundamental reason for the efficiency gain is due to the fact that the constrained estimates for treatment effects have smaller variances and similarly negligible biases compared with unconstrained estimates. Incorporating constraint information will induce positive correlations between estimated treatment effects across ordered strata, therefore increasing the probability of correct selection. We provide such empirical evidence in our appendix.

Besides the popular choice of binary outcomes in phase II trials, we also devote our effort to incorporating time-to-event outcomes as a non-trivial extension in this paper. Randomized phase II selection designs based on time-to-event outcomes with no stratification have been considered by Liu et al. [1993]. Many phase II trials are now designed to assess the promise of a molecularly targeted or an immuno-biological agent, given either alone or in combination with another regimen. In particular, it is not always anticipated that such agents are likely to improve tumor response rates. Rather, they will improve time-to-event outcomes such as EFS, progression-free survival (PFS), or overall survival (OS) through means other than direct cell killing as evidenced by tumor shrinkage. There is an increasing need in oncology to evaluate agents that are anticipated to increase PFS or OS, but not objective tumor response.

For randomized phase II selection design with binary outcome, Simon et al. [1985] first introduced the design for unstratified population and examined its performance and statistical characteristics. In this type of design, patients are randomized to two or more experimental agents and the treatment with the highest observed response rate will be selected for further trial [Simon et al., 1985]. However, there are some additional factors that may influence our decision to select the most appropriate treatment to proceed for a phase III trial, such as toxicity, cost etc. Sargent and Goldberg [2001] recently proposed a flexible randomized phase II selection trial that allows researchers to select the most appropriate treatment based on other factors when the observed response difference is relatively small.

We extend the idea of Sargent and Goldberg [2001] to our setting of randomized selection phase II trials with stratification. In Section 1.2 we present our method for both binary and time-to-event outcomes. In Section 1.3 we evaluate the method under simulated settings by comparing with the method without using the order information. In Section 1.4 we illustrate our method with two motivating examples. Finally in Section 1.5 we conclude this paper with some discussion.

1.2 Method

To better illustrate the specific design that we are proposing, we describe our method based on stratified randomized two-arm trials. Extension to randomized studies with multiple arms are straightforward.

Assume patients are stratified into G strata and randomized to Arms a and b. In total, there are N patients in each arm where the proportion of patients in each stratum is w_{jg} with $j \in \{a, b\}$. Here, $g = 1, ..., G$ and $\sum_{g=1}^{G} w_{jg} = 1$ for $j = a$ or b. Based on this, the number of patients in Arm j Stratum g is defined as $n_{jg} = N \cdot w_{jg}$.

Binary outcome

Assume that, in Arm j and Stratum g, the number of responders r_{jg} are independent binomial random variables with $r_{jg} \sim \text{Bin}(n_{jg}, \pi_{jg})$. Here, we assume the strata in Arm j satisfy the partial stochastic ordering constraints in its strata [Park et al., 2012a] defined by a constraint set $E \subset \{1, ..G\}^2$, i.e.,

$$\forall (u, v) \in E, \pi_{ju} \geqslant \pi_{jv}.$$

If there is total ordering among all strata such that $\pi_{j1} \geqslant \ldots \geqslant \pi_{jG}$, then the constraint set $E = \{(1,2), (2,3), \ldots, (G-1, G)\}$. But partial ordering is also possible. As an example, suppose the strata are formed by stage (1 vs. 2) and nodal involvement (no vs. yes). Then there may be no ordering between the stratum defined by 'stage 1+nodal yes' and the

stratum 'stage 2+nodal no'. As a result, E may take the form $E = \{(1 + no, 1 + yes), (1 + no, 2 + no), (1 + yes, 2 + yes), (2 + no, 2 + yes)\}$ where neither $(1 + yes, 2 + no)$ nor $(2 + no, 1 + yes)$ are included.

Let p_{jg} be the corresponding E-constrained maximum likelihood estimator (MLE) with Arm j Stratum g under the constraint set E. Under the framework described by Sargent and Goldberg [2001], we propose to make the selection of one of the two arms as follows. If for each stratum, the difference of E-constrained MLEs of response rates between two arms is greater than a pre-specified level θ, the arm with the higher response rate will be selected in the phase III trial.

Denote

$$\begin{aligned} \boldsymbol{\pi}_j &= (\pi_{j1}, \ldots, \pi_{jG})^\top, \quad j = a, b; \\ \boldsymbol{p}_j &= (p_{j1}, \ldots, p_{jG})^\top, \quad j = a, b; \\ \boldsymbol{\theta} &= (\theta, \ldots, \theta)^\top_{1\times G}, \quad \theta \geqslant 0\,. \end{aligned}$$

Let $\succeq$ and $\succ$ be the element-wise $\geqslant$ and $>$ functions respectively for a vector. We define $P_{corr} = Pr(\boldsymbol{p}_a \succ \boldsymbol{p}_b + \boldsymbol{\theta} \mid \boldsymbol{\pi}_a \succeq \boldsymbol{\pi}_b)$ as the probability of correctly choosing the better treatment when E-constrained MLE differences are greater than $\boldsymbol{\theta}$. Instead of using the same threshold θ for all strata, stratum-specific threshold θ_g can also be used to define P_{corr}. In actual calculation, we usually specify $\boldsymbol{\pi}_a = \boldsymbol{\pi}_b + \boldsymbol{\theta}^*$ where $\boldsymbol{\theta}^* = (\theta^*, \ldots, \theta^*)^\top_{1\times G}$ with $\boldsymbol{\theta}^* \succeq 0$ and calculate

$$P_{corr} = Pr(\boldsymbol{p}_a \succ \boldsymbol{p}_b + \boldsymbol{\theta} \mid \boldsymbol{\pi}_a = \boldsymbol{\pi}_b + \boldsymbol{\theta}^*)\,.$$

Similar to the specification of the threshold $\boldsymbol{\theta}$, we can also use stratum-specific difference θ_g^* for the gth component of $\boldsymbol{\theta}^*$.

If the difference for each stratum is in the opposite direction, we define this as the statistically wrong region with a prespecified level θ and we should avoid this situation. Mathematically, this corresponds to $P_{\text{wrong}} = \Pr(\boldsymbol{p}_b \succ \boldsymbol{p}_a + \boldsymbol{\theta} \mid \boldsymbol{\pi}_a \succeq \boldsymbol{\pi}_b)$. Similar to the calculation of P_{corr}, we calculate

$$P_{\text{wrong}} = \Pr(\boldsymbol{p}_b \succ \boldsymbol{p}_a + \boldsymbol{\theta} \mid \boldsymbol{\pi}_a = \boldsymbol{\pi}_b + \boldsymbol{\theta}^*) \, .$$

For the situation other than these two, we define it as statistically ambiguous region. Then the selection of the treatment for the phase III trial will be allowed to include other factors in addition to the response rate. Mathematically, this corresponds to

$$P_{\text{amb}} = 1 - P_{\text{corr}} - P_{\text{wrong}} \, .$$

In order to incorporate the partial stochastic ordering constraint in MLE, we construct the E-constrained MLE of the response rate for each arm and stratum as follows. First, we write the likelihood and log-likelihood functions as

$$\begin{aligned} l(\mathbf{p}_j) &= \frac{N!}{n_{j1}! \cdots n_{jG}!} \Big(\prod_{g=1}^{G} w_{jg}^{n_{jg}} \Big) \left\{ \prod_{g=1}^{G} \binom{n_{jg}}{r_{jg}} p_{jg}^{r_{jg}} (1 - p_{jg})^{n_{jg} - r_{jg}} \right\} \\ &\propto \prod_{g=1}^{G} p_{jg}^{r_{jg}} (1 - p_{jg})^{n_{jg} - r_{jg}}, \\ \log l(\mathbf{p}_j) &\propto \sum_{g=1}^{G} r_{jg} \log(p_{jg}) + (n_{jg} - r_{jg}) \log(1 - p_{jg}) \, . \end{aligned}$$

The optimization problem is to maximize $\log l(\mathbf{p}_j)$ through the following constrained convex optimization problem,

$$\begin{aligned} \min_{\mathbf{p}_j} \quad & -\log l(\mathbf{p}_j) \\ \text{subject to} \quad & E \subset \{1, ..G\}^2, (u, v) \in E, p_{ju} \geqslant p_{jv}; \\ & 0 \leqslant p_{jg} \leqslant 1; \quad j = a, b; \quad g = 1, \ldots, G. \end{aligned} \tag{1.1}$$

Let $c_{jg} = \text{logit}(p_{jg})$. As the logit function is monotonically increasing, we have $p_{ju} \leqslant p_{jv} \iff c_{ju} \leqslant c_{jv}$. With $\eta_{jg} = r_{jg}/n_{jg}$, we transform Eq. (1.1) to be

$$\begin{aligned} \min_{\mathbf{p}_j} \quad & \sum_{g=1}^{G} \left\{ \log(1 + e^{c_{jg}}) - \eta_{jg} c_{jg} \right\} n_{jg} \\ \text{subject to} \quad & E \subset \{1, ..G\}^2, (u, v) \in E, p_{ju} \geqslant p_{jv}; \\ & 0 \leqslant p_{jg} \leqslant 1; \quad j = a, b; \quad g = 1, \ldots, G. \end{aligned} \tag{1.2}$$

This is the generalized isotonic regression problem as $\log(1 + e^{c_{jg}})$ is strictly convex on $(-\infty, \infty)$. By Barlow and Brunk [1972], Eq. (1.2) is equivalent to

$$\begin{aligned} \min_{\mathbf{p}_j} \quad & \frac{1}{2} \sum_{g=1}^{G} \left(p_{jg} - \frac{r_{jg}}{n_{jg}} \right)^2 n_{jg} \\ \text{subject to} \quad & E \subset \{1, ..G\}^2, (u, v) \in E, p_{ju} \geqslant p_{jv}; \\ & 0 \leqslant p_{jg} \leqslant 1; \quad j = a, b; \quad g = 1, \ldots, G. \end{aligned} \tag{1.3}$$

Therefore Eq. (1.3) is a strictly convex and positive definite quadratic programming problem and there are many existing algorithms solving Eq. (1.3). We use an easy-to-implement R package *quadprog* in this paper.

Finally, after obtaining the E-constrained MLEs, we can plug them into Eq. (1.4) to estimate the probabilities that we are interested in,

$$
\begin{aligned}
P_{\text{corr}} = & \sum_{r_{a1}=0}^{n_{a1}} \cdots \sum_{r_{aG}=0}^{n_{aG}} \sum_{r_{b1}=0}^{n_{b1}} \cdots \sum_{r_{bG}=0}^{n_{bG}} \mathbb{1}_{\{\boldsymbol{p}_a > \boldsymbol{p}_b + \boldsymbol{\theta}\}} \prod_{g=1}^{G} \binom{n_{ag}}{r_{ag}} \pi_{ag}^{r_{ag}} (1-\pi_{ag})^{n_{ag}-r_{ag}} \\
& \times \binom{n_{bg}}{r_{bg}} \pi_{bg}^{r_{bg}} (1-\pi_{bg})^{n_{bg}-r_{bg}}, \\
P_{\text{wrong}} = & \sum_{r_{a1}=0}^{n_{a1}} \cdots \sum_{r_{aG}=0}^{n_{aG}} \sum_{r_{b1}=0}^{n_{b1}} \cdots \sum_{r_{bG}=0}^{n_{bG}} \mathbb{1}_{\{\boldsymbol{p}_b > \boldsymbol{p}_a + \boldsymbol{\theta}\}} \prod_{g=1}^{G} \binom{n_{ag}}{r_{ag}} \pi_{ag}^{r_{ag}} (1-\pi_{ag})^{n_{ag}-r_{ag}} \\
& \times \binom{n_{bg}}{r_{bg}} \pi_{bg}^{r_{bg}} (1-\pi_{bg})^{n_{bg}-r_{bg}}, \\
P_{\text{amb}} = & 1 - P_{\text{corr}} - P_{\text{wrong}}.
\end{aligned}
\tag{1.4}
$$

Following Sargent and Goldberg [2001], we use $\lambda \equiv P_{\text{corr}} + \rho P_{\text{amb}}$ with a pre-specified $\rho \in [0,1]$ as the probability for selecting the superior treatment. For determining the desired sample size, the most conservative way is that we assume any statistically ambiguous outcome could result in Arm b being chosen and thus we will try to avoid any ambiguous results. With the prespecified $\phi \in [0,1]$, N should be large enough to ensure that $\lambda > \phi$ when $\rho = 0$. A second approach would be to assume ρ percent of ambiguous cases are indeed correct and being selected. In this case, N would be selected such that $\lambda > \phi$ when $\rho \neq 0$.

Time-to-event outcome

We consider survival probabilities at a pre-fixed time x as the interested endpoint [Rubinstein et al., 2005]. This endpoint is interpretable even in the presence of non-proportional hazard, which is a well-known issue for modern cancer treatments such as immunotherapy [Uno et al., 2015, 2014].

Suppose $S_{jg}(x), j = a, b$ and $g = 1, \ldots, G$ is the true survival probability at time x for Stratum g in Arm j. Further assume that the strata in Arm j satisfy the partial stochastic ordering constraints at a given time x defined by the constraint set $E \subset \{1, ..G\}^2$, i.e., $\forall (u, v) \in E, S_{ju}(x) \geqslant S_{jv}(x)$. Similar to the binary outcome setting, the choice of E is rather flexible to reflect known order constraints among strata. We note that the constraints on the survivor functions are on the given time x only, instead of on the whole functions. Such pointwise constraint has known computational and theoretical advantages as discussed by [Park et al., 2012a,b].

Let $\tilde{S}_{jg}(x)$ be the corresponding E-constrained nonparametric maximum likelihood estimator (NPMLE) for survivor probability of Arm j Stratum g subject to constraint set E applied at a given time x only. Denote

$$\begin{aligned} \boldsymbol{S}_j &= (S_{j1}(x), \ldots, S_{jG}(x))^\top, \quad j = a, b; \\ \tilde{\boldsymbol{S}}_j &= (\tilde{S}_{j1}(x), \ldots, \tilde{S}_{jG}(x))^\top, \quad j = a, b; \\ \boldsymbol{\theta} &= (\theta, \ldots, \theta)^\top_{1\times G}, \quad \theta \geqslant 0\,. \end{aligned}$$

The treatment selection strategy is similar to the above binary outcome

case. That is, we define $P_{corr} = Pr(\tilde{\boldsymbol{S}}_a \succ \tilde{\boldsymbol{S}}_b + \boldsymbol{\theta} \mid \boldsymbol{S}_a \succeq \boldsymbol{S}_b)$ as the probability of correctly choosing the better treatment when the difference of E-constrained NPMLE at time x is greater than θ. Then we can define $P_{wrong} = Pr(\tilde{\boldsymbol{S}}_b \succ \tilde{\boldsymbol{S}}_a + \boldsymbol{\theta} \mid \boldsymbol{S}_a \succeq \boldsymbol{S}_b)$ as the probability of wrongly choosing the worse treatment, and $P_{amb} = 1 - P_{corr} - P_{wrong}$ as the probability of being in the ambiguous region. Similar to the binary outcome setting, we usually specify $\boldsymbol{S}_a = \boldsymbol{S}_b + \boldsymbol{\theta}^*$ where $\boldsymbol{\theta}^* = (\theta^*, \ldots, \theta^*)^\top_{1\times G}$ with, $\boldsymbol{\theta}^* \succeq 0$ and calculate

$$\begin{aligned} P_{corr} &= Pr(\widetilde{\boldsymbol{S}}_a \succ \widetilde{\boldsymbol{S}}_b + \boldsymbol{\theta} \mid \boldsymbol{S}_a = \boldsymbol{S}_b + \boldsymbol{\theta}^*), \\ P_{wrong} &= Pr(\widetilde{\boldsymbol{S}}_b \succ \widetilde{\boldsymbol{S}}_a + \boldsymbol{\theta} \mid \boldsymbol{S}_a = \boldsymbol{S}_b + \boldsymbol{\theta}^*), \\ P_{amb} &= 1 - P_{corr} - P_{wrong}. \end{aligned} \tag{1.5}$$

In order to incorporate the partial stochastic ordering constraints in survival probability estimation, we follow Park et al. [2012a] to construct the pointwise E-constrained NPMLE of survival probability for each arm and stratum. In short, let the observed survival time for each individual i in Arm j and Stratum g be Y_{jgi} and let the event indicator be Δ_{jgi} for $i = 1, ..., n_{jg}$. Then the generalized likelihood function of survival probabilities for each arm j is

$$l(S_{j1}(\cdot), ...S_{jG}(\cdot)) = \prod_{g=1}^{G}\prod_{i=1}^{n_{jg}}\{S_{jg}(Y_{jgi}-) - S_{jg}(Y_{jgi})\}^{\Delta_{jgi}} S_{jg}(Y_{jgi})^{1-\Delta_{jgi}} . \tag{1.6}$$

The estimation of $\tilde{S}_{jg}(x)$ need to maximize Eq. (1.6) subject to the partial-ordering constraint E such that $\forall(u, v) \in E, S_{ju}(x) \geqslant S_{jv}(x)$.

For Arm j and Stratum g, let m_{jg} be the number of distinct events and X_{jgl} be the distinct event times for $l = 1, ...m_{jg}$. Further define $X_{jg0} = 0$ and $X_{jg(m_{jg}+1)} = \infty$. Let $Z_{jg}(x)$ be the number at risk at time x and $M_{jg}(x)$ be the number of distinct events in $(0, x]$. Let d_{jgl} and z_{jgl} be the number of events and number at risk at time X_{jgl}. Let $h_{jg}(t) = \log\{S_{jg}(t)/S_{jg}(t-)\}$ and the corresponding discrete hazard at time t be $1 - \exp\{h_{jg}(t)\}$. Then the loglikelihood function of Eq. (1.6) subject to the partial-ordering constraints E is

$$\begin{aligned}
&\max_{\mathbf{h}_{jg}} \sum_{g=1}^{G} \Big\{ \sum_{l=1}^{m_{jg}} \big(d_{jgl} \log[1 - \exp\{h_{jg}(X_{jgl})\}] + \\
&\qquad (z_{jgl} - d_{jgl}) h_{jg}(X_{jgl}) \big) + Z_{jg}(x) h^{\delta}_{jg}(x) \Big\} \\
&\text{subject to} \sum_{i=1}^{M_{ju}(x)} h_{ju}(X_{jui}) + h^{\delta}_{ju}(x) \geqslant \sum_{i=1}^{M_{jv}(x)} h_{jv}(X_{jvi}) + h^{\delta}_{jv}(x), \text{for } (u, v) \in E, \\
&\qquad h^{\delta}_{jg}(x) \leqslant 0 \,,
\end{aligned} \tag{1.7}$$

where $\mathbf{h}_{jg} = \{h_{jg}(X_{jg1}), ..., h_{jg}(X_{jgm_{jg}}), h^{\delta}_{jg}(x)\}$, $g = 1, \ldots, G$. Here, we define $h^{\delta}_{jg}(x) = \mathbb{1}_{(x \neq X_{jgM_{jg}(x)})} h_{jg}(x)$ to account for the fact that we do not need to add this extra term $Z_{jg}(x)h_{jg}(x)$ if $x = X_{jgM_{jg}(x)}$.

Now, this becomes a linearly constrained concave maximization problem. One challenge here is that our data contain many more observed event times than strata. Thus, we need to transform this problem into a simple concave maximization problem subject to linear constraints by

using the profile likelihood.

Let q_{jg} satisfies the relationship $S_{jg}(x) = \exp(q_{jg})$ at time x. Suppose $\hat{k}_{jg}$ is the unique solution of the equation $\sum_{i=1}^{M_{jg}(x)}\{\log(1 - d_{jgi}/(z_{jgi} + k_{jg})\} = q_{jg}$. If $q_{jg} = 0$, $\hat{k}_{jg} = \infty$ and if $q_{jg} = -\infty$, $\hat{k}_{jg} = d_{jgM_{jg}(x)} - z_{jgM_{jg}(x)}$. If $M_{jg}(x) = 0$, let $K_{jg}(q_{jg}; x) = -Z_{jg}(x)$, and otherwise let $K_{jg}(q_{jg}; x) = \max(-Z_{jg}(x), \hat{k}_{jg})$. Based on Park et al. [2012a], we can then transform Eq. (1.7) into the profile loglikelihood function as

$$\begin{aligned}
\max_{q_j} \sum_{g=1}^{G} \ell_{jg}(q_{jg}; x) = \sum_{g=1}^{G} \Bigg(& \sum_{i=1}^{M_{jg}(x)} [(z_{jgi} - d_{jgi}) \log\{z_{jgi} + K_{jg}(q_{jg}; x) - d_{jgi}\} \\
& - z_{jgi} \log\{z_{jgi} + K_{jg}(q_{jg}; x)\}] + \mathbb{1}_{\{K_{jg}(q_{jg}; x) = -Z_{jg}(x)\}} Z_{jg}(x) \\
& \times \Big[q_{jg} - \sum_{l=1}^{M_{jg}(x)} \log\{1 - \frac{d_{jgl}}{z_{jgl} + K_{jg}(q_{jg}; x)}\} \Big] \Bigg) \\
\text{subject to} \quad & q_{ju} \geqslant q_{jv}, \text{ for all } (u, v) \in E; \\
& q_{jg} \leqslant 0.
\end{aligned} \tag{1.8}$$

The corresponding derivative of Eq. (1.8) is $d\ell_j(q_j; x)/dq_j^\top = \{-K_{j1}(q_{j1}; x), ..., -K_{jG}(q_{jG}; x)\}^\top$. In order to maximize Eq. (1.8), only G parameters $q_j = (q_{j1}, ..., q_{jG})$ need to be estimated for each arm to get $\tilde{S}_{jg}(x) = \exp(\hat{q}_{jg})$. This is a strictly concave maximization problem subject to linear constraints. There are many existing algorithms solving this problem. We use an easy-to-implement R function *contrOptim* to solve this optimization problem.

Finally, after getting the E-constrained NPMLE of survival probability at time x, we could plug in Eq. (1.5) to estimate P_{corr} and P_{amb} and calculate λ using Monte Carlo simulations.

1.3 Evaluation with simulated settings

In this section, we would like to compare the performance of the proposed method with the simple randomized stratified selection design, which does not incorporate order information. We consider the setting that each of the Arms a and b has N patients. There are $G = 2$ strata in each arm. The patient proportions in different strata are $w_{a1} = w_{b1} = 0.4$ (therefore $w_{a2} = w_{b2} = 0.6$).

Binary outcome

Assume $\pi_{ag} \geqslant \pi_{bg}$ and $\pi_{j2} \geqslant \pi_{j1}, j = a, b, g = 1, 2$ without loss of generality. Accordingly, the E-constraint is $\pi_{j2} \geqslant \pi_{j1}$. The binomial response rate MLE without considering the order information would be the observed response rate $\hat{\pi}_{jg} = {n_{jg}}^{-1} r_{jg}$. We will compare the calculated λ between the two methods.

The probabilities of selecting the superior treatment (λ) across different N and θ are shown in Figure 1.1. When calculating λ, we consider $\rho = 0$ or $\rho = 0.5$. The former means that we do not want any ambiguity (i.e. P_{amb}) in selecting the superior treatment whereas the latter means that

we use a coin flip to recommend the superior treatment when the result falls in the ambiguous region. Also, we consider different values of $\boldsymbol{\theta}$ in the design as a trade-off between minimizing the sample size N and the clinical consideration of other factors. Overall, we see our method gives uniformly larger λ than the original method without using order information.

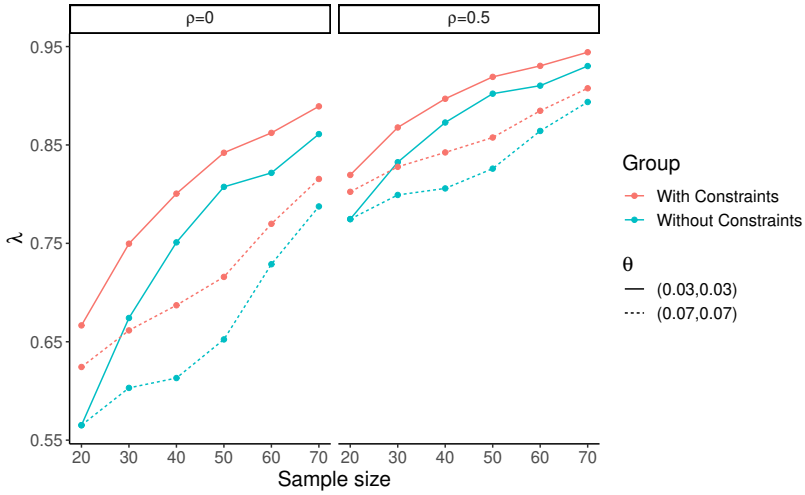

Figure 1.1: Probability of selecting the superior treatment (λ) for various N and $\boldsymbol{\theta}$, between $\rho = 0$ and $\rho = 0.5$, fixing $\boldsymbol{\pi}_a = (0.55, 0.65)$, $\boldsymbol{\theta}^* = (0.2, 0.2)$.

Table 1.1 lists required sample sizes per arm for common response probabilities to get $\lambda = 0.8$. The table also shows that λ is not a monotone function of response probabilities.

Table 1.1: Sample size per arm for various response probabilities to get $\lambda = 0.8$, assuming $\boldsymbol{\theta} = (0.05, 0.05)$, $\pi_{j2} - \pi_{j1} = 0.1$ and $\boldsymbol{\theta}^* = (0.2, 0.2)$.

	With constraints		Without constraints	
π_{a1}	$\rho = 0.5$	$\rho = 0$	$\rho = 0.5$	$\rho = 0$
0.25	14	29	18	32
0.4	17	45	23	63
0.55	18	58	24	73
0.7	17	45	23	63
0.85	13	29	17	33

Survival outcome

Assume $S_{ag}(x) \geqslant S_{bg}(x)$, $S_{j2}(x) \geqslant S_{j1}(x), j = a, b, g = 1, 2$ without loss of generality. Accordingly, $S_{jg}(x)$ satisfy the E-constraints at time x that $S_{j2}(x) \geqslant S_{j1}(x), j = a, b$. The survival probability NPMLE without considering the order information would be the Kaplan-Meier estimator $\hat{S}_{jg}(x) = \prod_{i:x_i \leqslant x}(1 - \frac{d_{jgi}}{n_{jgi}})$. We will then compare the calculated λ between the two methods.

Suppose patients enroll according to a Poisson process with an accrual rate of 4 patients per month for each of the treatment arm stratum. We will continue to follow up for an additional 6 months after the last patient is enrolled for each stratum. Suppose the survival time follows exponential distribution and we are constraining and comparing survival probabilities at 6 months. The estimation of P_{corr} and P_{amb} is based on 8,000 simulations.

The probability of selecting the superior treatment (λ) across different N and $\boldsymbol{\theta}$ is shown in Figure 1.2. Note that from the figure, λ is not a mono-

tonically increasing function of sample size. Thus, we should calculate the required sample size needed for each scenario without assuming λ should increase with the increase of sample size. Under different selection criteria ρ and trade-off values $\boldsymbol{\theta}$, we see our method gives uniformly larger λ than the original method without using order information.

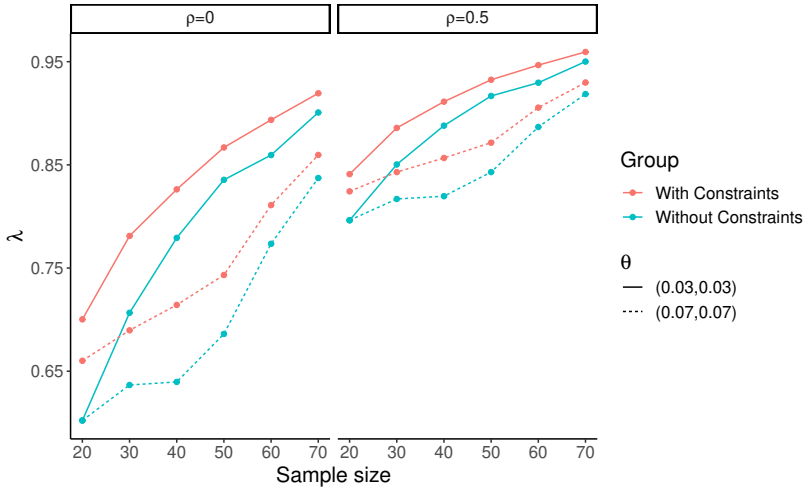

Figure 1.2: Probability of selecting the superior treatment (λ) for various N and $\boldsymbol{\theta}$, between $\rho = 0$ and $\rho = 0.5$, fixing $\boldsymbol{S}_a = (0.75, 0.85)$, $\boldsymbol{\theta}^* = (0.2, 0.2)$.

Table 1.2 gives sample sizes per arm for common survival probabilities for different additional follow-up (FUP) months after the last patient is enrolled for each stratum. Similar to the binary outcome setting, we see that the required sample sizes are smaller when constraints are used, except for the setting with $S_{a1}(6) = 0.85, \text{FUP} = 4, \rho = 0.5$. This might be due to the high censoring probability and the fact that P_{amb} is included in the definition of λ.

Table 1.2: Sample size per arm for various survival probabilities at 6 months to get $\lambda = 0.8$, assuming $\boldsymbol{\theta} = (0.05, 0.05)$, $S(6)_{j2} - S(6)_{j1} = 0.1$ and $\boldsymbol{\theta}^* = (0.2, 0.2)$.

		With Constraints		Without Constraints	
FUP (months)	$S_{a1}(6)$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.5$	$\rho = 0$
6	0.25	14	29	18	32
	0.4	17	45	22	64
	0.55	19	58	24	73
	0.7	17	47	23	64
	0.85	12	28	17	39
5	0.25	15	34	19	42
	0.4	19	55	25	67
	0.55	22	58	27	74
	0.7	19	52	24	67
	0.85	15	33	19	43
4	0.25	21	42	25	52
	0.4	27	58	32	72
	0.55	28	63	32	77
	0.7	26	55	29	69
	0.85	22	40	21	50

1.4 Evaluation with real settings

Here we would like to use two real clinical trial examples to demonstrate the advantage of our method in binary and survival settings. First, let's consider our motivating clinical trial example for patients with metastatic urothelial cancer. This trial has two treatment arms: treatment nivolumab, gemcitabine, oxaliplatin versus treatment nivolumab, gemcitabine, carboplatin. The primary outcome of the study is response rate and the stratification is based on lymph node only metastasis versus metastasis of other sites. Previous studies reported that there are around 30% of patients

with lymph node only metastasis have better response rates. Here, a randomized stratified phase II selection trial would be the most appropriate.

We hypothesize that the lymph node only group will have a higher response rate than the other group. We assume two strata of the inferior treatment arm have response rates $\boldsymbol{\pi}_b = (0.4, 0.5)$, $\boldsymbol{\theta}^* = (0.2, 0.2)$ and 30% of the total study patients have lymph node only metastasis. Figure 1.3 shows the calculated λ under different sample sizes using our proposed method and the method not considering order information. We see that the sample size N per arm derived from our method is around 20 versus 30 using the method without considering order information to achieve $\lambda = 0.8$ with $\rho = 0.5$ and $\boldsymbol{\theta} = (0.05, 0.05)$.

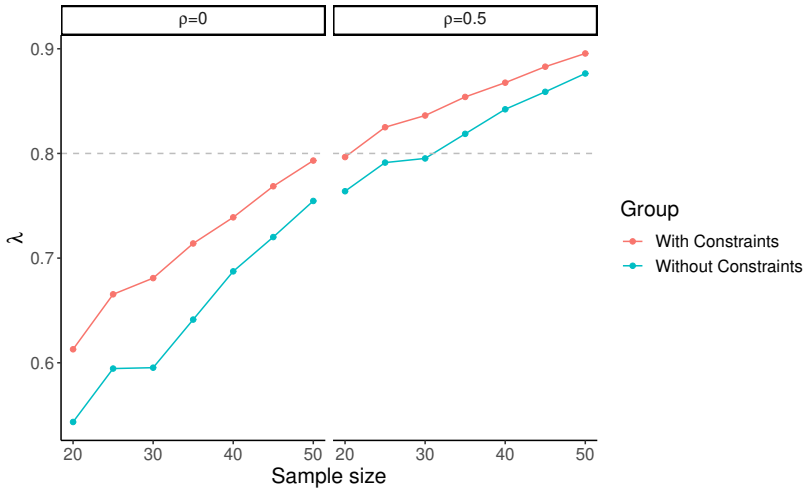

Figure 1.3: Probability of selecting the superior treatment (λ) for various N, between $\rho = 0$ and $\rho = 0.5$ with $\boldsymbol{\theta} = (0.05, 0.05)$, $\boldsymbol{\pi}_b = (0.4, 0.5)$, $\boldsymbol{\theta}^* = (0.2, 0.2)$.

For the other motivating example, it has two treatment arms for ex-

amination with the goal of comparing the effect of Sacituzumab Govitecan in combination with Pembrolizumab versus Capecitabine and Pembrolizumab in patients with triple negative breast cancer and residual disease. The primary outcome of the study is EFS and the stratification of this study is based on nodal status. Previous studies reported that the node positive group had lower 2-year EFS than node negative group. The prevalence of node positive is around 30%. Again, a randomized stratified phase II selection trial would be the most appropriate.

Suppose the two strata of the inferior treatment arm have 2-year EFS $\boldsymbol{S}_{\mathfrak{b}} = (0.6, 0.7)$, and sample size is determined based on an improvement of $\boldsymbol{\theta}^* = (0.15, 0.15)$ for the better treatment arm. Suppose patients enroll according to a Poisson process with an accrual rate of 8 patients per year for each of the treatment arm stratum. We will continue follow-up for an additional 2 years after the last patient is enrolled for each stratum. Suppose the survival time follows exponential distribution and we are constraining and comparing survival probabilities at 2 years. Based on 8,000 Monte Carlo simulations, the estimated λ under different sample sizes using our proposed method and the method not considering order information are shown in Figure 1.4. We see that the sample size derived from our method is around 25 versus 35 using the method without considering order information to achieve $\lambda = 0.8$ with $\rho = 0.5$ and $\boldsymbol{\theta} = (0.02, 0.02)$.

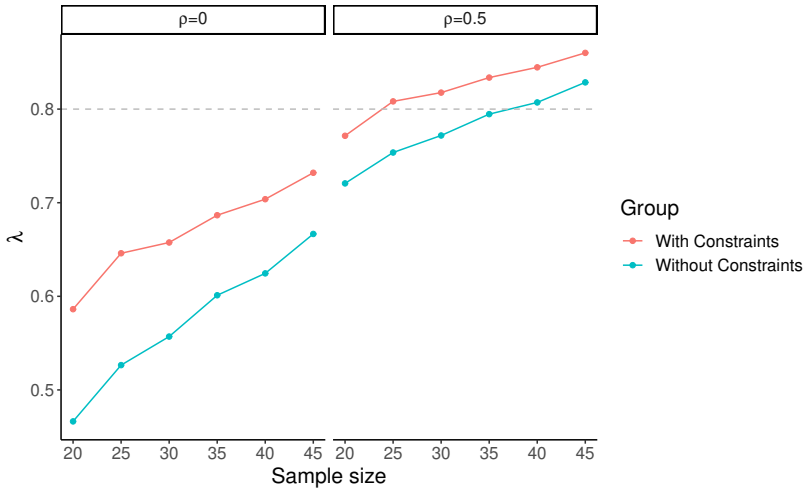

Figure 1.4: Probability of selecting the superior treatment (λ) for various N, between $\rho = 0$ and $\rho = 0.5$ with $\boldsymbol{\theta} = (0.02, 0.02)$, $\boldsymbol{S}_b = (0.6, 0.7)$, $\boldsymbol{\theta}^* = (0.15, 0.15)$.

1.5 Discussion and conclusion

We have considered designing stratified randomized phase II cancer trials using order constraints. Given the exploratory nature of phase II trials, it is important to incorporate known constraints into the sample size calculation procedure to improve statistical efficiency. Using a two-arm randomized selection design as our setting, we demonstrated improvement of selection probabilities or reduction of sample sizes for both binary and time-to-event outcomes.

Our results are easily generalizable to randomized phase II screening designs and we put the details in the appendix. In addition, we can simply use the E-constrained MLEs in most of the calculations laid out in [Jung and George, 2009]. Such a generalization can be an interesting future

work. In addition, our approach shows promise for Phase III clinical trial applications, particularly when individual strata contain limited numbers of patients. For example, this adaptation would be especially valuable in rare disease studies, where recruitment challenges naturally result in small stratified subgroups. By incorporating the natural ordering relationships between strata, our method can enhance statistical power in these challenging late-phase trial contexts.

A fundamental assumption for our method is the E-constraint assumption. Because the constraint set E is very flexible, we recommend including only well-established ordering. In other words, when there is uncertainty or insufficient data to support a particular order relationship, it may be better not to include such a relationship in the set E.

Statistically, one can also try to empirically evaluate the ordering of an assumption when there is existing data. A recent method known as nearly isotonic regression [Tibshirani et al., 2011, Matsuda and Miyatake, 2022] may be used to visually evaluate such an assumption. We leave the details to the Appendix. In there, we also evaluate the performance of our method when there is a violation of the ordering assumption.

In our calculation, we defined P_{corr} as the probability of differences in the response rates or survival probabilities surpassing a common threshold for all strata. Alternative definitions can be used. For example, it can be defined as 'winning' in at least one stratum, instead of in all strata. We are also extending our design to recommend a subgroup-specific winner

where the subgroup consists of certain strata.

Acknowledgments

Research reported in this work was partially supported by the Specialized Program of Research Excellence (SPORE) program, through the National Cancer Institute (NCI), grant P50CA278595. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Data Availability

The data that support the findings in this paper were derived from simulation and from published summary information. An R package for the method in this paper is available on GitHub: https://github.com/yc702/constrselect. The package includes help files, unit testing, the simulation code for the paper and a readme file with instructions about installing the package and using the package functions.

2.1 Overview

This chapter introduces an R package *contrselect*, which is used to incorporate patient heterogeneity and stratification into randomized phase II selection design. This package focuses on estimating the probability of correct selection (λ) when comparing two treatments, accounting for additional factors when the observed response difference is relatively small. The package offers flexibility in handling both survival and binary outcomes, as well as with and without constraints. Regarding patient heterogeneity, it can accommodate both total and partial ordering information, making it particularly useful when we have multiple strata. We illustrate the implementation of our package under different scenarios. The key functions for our package are *pickwin_bin_multiple*() for binary outcome and *pickwin_surv_fun*() for survival outcome. Our package is available on github https://github.com/yc702/constrselect and it passes R-CMD-check.

2.2 Package dependencies

contrselect was developed with dependence on 7 packages:

- *parallel*, *doParallel*, *foreach* and *doRNG* are essential tools for implementing parallel computing, as most of our methods are based on

Monte Carlo simulations. These packages can significantly increase the efficiency of our method by distributing computational tasks across multiple CPU cores. Especially, *doRNG* is used to ensure reproducibility in parallel computing given the same seed [R Core Team, 2023, Microsoft and Weston, 2022a,b, Gaujoux, 2023].

- *quadprog* provides functions for solving quadratic programming (QP) problems, which minimizes a quadratic function subject to linear constraints. This package is essential for the binary outcome method given total and partial ordering of strata [Turlach and Weihs, 2019].

- *survival* offers comprehensive tools for analyzing time-to-event (survival) data. Here, we primarily use its functions for getting Kaplan-Meier estimators at a given time [Therneau, 2023].

- *dplyr* is a powerful and efficient tool for tidy data manipulation [Wickham et al., 2023].

2.3 Key functions

Binary outcome

pickwin_bin_exact() and *pickwin_bin_multiple*() are two crucial functions to implement in our package for binary outcome. While *pickwin_bin_exact*() only works for the two-strata case, *pickwin_bin_multiple*() is more flexible, allowing us to incorporate more than two strata. Another important

difference between these two functions is that *pickwin_bin_exact*() employs the exact binomial method to do the statistical calculation, while *pickwin_bin_multiple*() relies on Monte Carlo simulations, resulting a significantly longer computation time. Since both functions share similar input arguments, here I will mainly focus on *pickwin_bin_multiple*() for demonstration purpose.

- *n* Total sample size for each treatment arm.
- *p_inf* A vector of response probabilities for the inferior treatment arm for each stratum.
- *D* A vector of two treatment arms differences for each stratum, Default: c(0.15, 0.15, 0.15).
- *d* A vector of ambiguous region for each stratum, Default: c(0.05, 0.05, 0.05) for three strata.
- *prop.strat* The sample size proportion for each stratum, Default: c(0.2, 0.3, 0.5) for three strata.
- *study* Could be either "Constrained" or "Origin" for the two type of study design with or without using constraints, Default: 'Constrained'.
- *S* Number of simulations for calculating the probabilities.
- *cluster* Number of parallel running CPU cores, Default: 6.

- *order_list* A list of strata order allowing for total and partial ordering, grouped in a vector within a list. Eg. list(1,2,3) for total ordering and list(1,c(2,3)) for partial ordering.

- *with_seed* Random seed for simulation, Default: NULL.

The output of the function returns a data frame of whether each simulated scenario would result in a correct and wrong decision with a total of *S* number of simulations.

Survival outcome

pickwin_surv_fun () is the main function for survival outcome. It is flexible as it allows us to incorporate more than two strata. The statistical calculations in this function is based on Monte Carlo simulation.

- *n, prop.strat, S, d, study, cluster, order_list, with_seed* are the same as the input as function *pickwin_bin_multiple*().

- *surv_inf* The survival probability at time *x* for patients in the inferior treatment arm.

- *surv_sup* The survival probability at time *x* for patients in the superior treatment arm.

- *arrival_rate* The Poisson arrival rate for patients, number of patients accrued each month/year.

- *FUP* Additional follow up time after the last patient is accrued.
- x Time we are interested in comparing the survival probabilities.

The output of the function similarly returns a data frame of whether each simulated scenario would result in a correct and wrong decision with a total of S number of simulations.

2.4 Example implementation

For a two-strata example with a binary outcome, we will look at a scenario similar to the real-world setting described in Chapter 1. Suppose a clinical trial has two treatment arms to study and we would like to pick a winner. The primary outcome of the study is the response rate, and the patients' stratification is based on lymph node only metastasis versus metastasis of other sites. Historical literature mentions that around 30% of patients with lymph node only metastasis have better response rates.

Suppose the lymph node only group has a higher response rate than the other group. We assume two strata of the inferior treatment arm having response rates $(0.4, 0.5)$ while the better treatment arm having $(0.6, 0.7)$ constraining on strata 2 has a better response rate than strata 1. We use exact binomial function, *pickwin_bin_exact*() to calculate $\mathrm{P}_{\mathtt{corr}}$ and $\mathrm{P}_{\mathtt{amb}}$. We see that the sample size N per arm derived from our method is around

20 to achieve $\lambda = \rho \times P_{amb} + P_{corr} = 0.8$ with $\rho = 0.5$ and ambiguous regions (0.05,0.05).

```
result = pickwin_bin_exact(n = 20, p_inf = c(0.4,0.5),
                                    D=c(0.2,0.2),d=c(0.05,0.05),
                                    prop.strat=0.7,study="Constrained",
                                    order_list=list(1,2))

result
#>       pcorr      pamb
#> 0.6128794 0.3673020
```

With a slight modification of this example, suppose the patients now are stratified into three strata, based on cancer stage 1,2,3. Previous research indicated that the larger the cancer stage, the worse the prognosis, with the sample proportion of 4:3:3.

We assume three strata of the inferior treatment arm having response rates (0.5,0.4,0.3) while the better treatment arm having (0.65,0.55,0.45) for cancer stage 1,2,3. Using 5000 Monte Carlo simulations, the *pickwin_bin_multiple*() function calculates P_{corr} and P_{amb}. In order to incorporate the ordering constraints, we specify *order_list* to be 'list(3,2,1)' which indicates the total ordering constraints of response rates strata 3 < strata 2 < strata 1. If we want to specify partial ordering constraints, eg. strata 3 < strata 2 and strata 3 < strata 1 without specifying the order between strata 1 and 2, we could set *order_list=list(3,c(1,2))*. We see

that the sample size N per arm derived from our method is around 58 to achieve $\lambda = \rho \times P_{amb} + P_{corr} = 0.8$ with $\rho = 0.5$ and ambiguous regions (0.02,0.02,0.02).

```
result <- pickwin_bin_multiple(n = 58, p_inf = c(0.5,0.4,0.3),
                                D=c(0.15,0.15,0.15),
                                d=c(0.02,0.02,0.02),
                                prop.strat=c(0.4,0.3,0.3),
                                study="Constrained", S = 5000,
                                cluster=6,order_list=list(3,2,1))
```

```
Pcorr = sum(result$Corr)
Pwrong = sum(result$Wrong)
(Pcorr+0.5*(5000-Pcorr-Pwrong))/5000
#> 0.8052
```

For a two-strata example with survival outcome, we will also look at a scenario similar to the real-world setting described in Chapter 1. Suppose two treatment arms are evaluated, with event-free survival (EFS) as the primary outcome and patients' stratification based on nodal status. Previous studies showed that the node positive group had a lower 2-year EFS with the prevalence to be around 30%.

Suppose two strata of the inferior treatment arm have 2-year EFS (0.6, 0.7), and the sample size is determined based on an improvement of 0.15 for the better treatment arm. Suppose patients enroll according to a

Poisson process with an accrual rate of 8 patients per year for each of the treatment arm stratum. We will continue to follow up with patients for an additional two years after the last patient is enrolled in each stratum. Assuming that survival time follows an exponential distribution, we will compare and constrain survival probabilities at two years. Based on 8000 Monte Carlo simulations, we need a sample size of 30 to achieve $\lambda = \rho \times P_{amb} + P_{corr} = 0.8$ with $\rho = 0.5$ and ambiguous region (0.02, 0.02). We could also generalize it to include more than two strata.

```
result <- pickwin_surv_fun(n=25,prop.strat=c(0.3,0.7),
                           surv_inf=c(0.6,0.7),
                           surv_sup=c(0.75,0.85),
                           d=c(0.02,0.02), arrival_rate=8,
                           FUP=2,x=2,
                           S=8000,study ="Constrained",
                           cluster=2,order_list=list(1,2),
                           with_seed = 111)

## Pamb
pamb=8000-sum(result$Corr)-sum(result$Wrong)

## lambda calculation with rho = 0.5
(sum(result$Corr)+(pamb)/2)/8000
#> 0.80825
```

3 CONFIDENCE INTERVAL CONSTRUCTION FOR CAUSALLY GENERALIZED ESTIMATES WITH TARGET SAMPLE SUMMARY INFORMATION

3.1 Introduction

Causal inference plays a pivotal role in population health research, providing essential tools for understanding and shaping effective health interventions. One of its popular research questions is how to generalize causal findings from a study population to a target population (Degtiar and Rose, 2023, Colnet et al., 2023, Chen et al., 2023). For example, we may want to generalize findings about the effectiveness of a treatment from a properly conducted randomized clinical trial (RCT) to its target population. We usually refer to this type of problem as generalizability [Cole and Stuart, 2010, O'Muircheartaigh and Hedges, 2013], transportability [Rudolph and van der Laan, 2017, Pearl and Bareinboim, 2011], or data fusion [Bareinboim and Pearl, 2016, Graham et al., 2025, Li and Luedtke, 2023]. There are some differences between these terminologies, and more detailed explanations can be found in Colnet et al. [2023].

For much of this article, for demonstration purposes, we focus on generalizing the average treatment effect (ATE), although similar considerations can be given to other causal estimands such as the Average Treatment effect on the Treated (ATT) or Average Treatment effect on the Overlap

population (ATO) [Colnet et al., 2023]. Our method mainly deals with causal generalization from a source to a target population when individual treatment effects are heterogenenous. Specifically, the individual treatment effects may depend on certain covariates, known as effect modifiers. In addition, the distributions of the effect modifiers can differ between the two populations [Sugiyama et al., 2007].

Much of the existing literature take a data fusion or integrated data analysis approach to this problem [Colnet et al., 2023, Bareinboim and Pearl, 2016, Graham et al., 2025, Li and Luedtke, 2023, Dahabreh et al., 2023]. Such approaches typically require individual data from both populations. However, there can be settings when comprehensive data at the individual level may not be consistently accessible within a target sample, owing to various practical considerations such as restricted data sharing, storage constraints, and privacy apprehensions [Degtiar and Rose, 2023]. On the contrary, obtaining summary-level information from the target sample is comparatively more feasible. This type of information can be readily gathered from diverse sources such as healthcare databases, census data, and published literature.

To deal with the challenges posed by lack of individual data from the target population, Dong et al. [2020] adapted the entropy balancing weights approach [Hainmueller, 2012, Zhao and Percival, 2016] for generalizing ATE estimation from an RCT to a given target population. Josey et al. [2020] then extended the approach to the setting when the source

sample is from observational studies. In particular, they proposed a two-step procedure to adjust for covariate shift and confounding separately. By showing that the weights produced by the two-step procedure of Josey et al. [2020] can be consolidated into a one-step procedure, Chen et al. [2023] developed a more intuitive strategy that may further mitigate bias under mild conditions, which rely solely on summary-level information from the target sample and individual-level covariates from the source sample. Recently, Chattopadhyay et al. [2024] proposed a very similar strategy.

The purpose of this article is to provide a practical solution to a key limitation with these methods: how to construct confidence intervals (CIs) for the resulting causally generalizable estimates. Chen et al. [2023] showed that the asymptotic variance of their estimator depends on individual-level data in the target sample. Similarly the asymptotic variance of the estimator from Chattopadhyay et al. [2024] also depends on the individual-level data in the target sample. This article addresses this limitation by proposing a method to construct CIs for the proposed estimator from Chen et al. [2023] using resampling-based perturbation, without requiring individual-level data from the target sample.

This paper is organized as follows: In Section 3.2, we present general notations and assumptions for our method. In Section 3.3 we present two methods to do the resampling-based perturbation for CI construction. In Sections 3.4 and 3.5, we evaluate the proposed methods using simulation

studies and a real data application using cross-validation. In Section 3.6, we conclude the paper with a discussion.

3.2 Notation and framework

Suppose we have individual-level data in a representative sample of our source population $\mathcal{S}$, denoted as $\{(X_i; A_i; Y_i) : i \in \mathcal{S}\}$ with n_s subjects. We denote $X_i \in \mathcal{X} \subset \mathbb{R}^p$ as the pre-treatment covariates which include confounders and treatment effect modifiers. The treatment indicator is denoted as $A_i \in \{0, 1\}$, and Y_i is the outcome we are interested in. For a representative sample of our target population $\mathcal{T}$, the sample size is n_t but we do not observe the individual-level data. Instead, we only have the information for the first moments based on a set of linearly independent covariate functions $h_k : \mathcal{X} \to \mathbb{R}; k = 1, \ldots, K_h$ from the target sample as follows.

$$\bar{h}_{k,\mathcal{T}} \equiv \frac{1}{n_t} \sum_{i \in \mathcal{T}} h_k(X_i), k = 1, \ldots, K_h \, .$$

Each h_k is usually defined on one or two covariates, instead of on the full covariate vector. For continuous covariates, if h_k is defined as an identity function, then $\bar{h}_{k,\mathcal{T}}$ represents the mean of this component. If h_k is defined as a polynomial function of degree 2, $\bar{h}_{k,\mathcal{T}}$ corresponds to the second moment, or variance, of this component. For discrete covariates, h_k could be defined as an indicator function to count the number of subjects in a particular category.

Here, we formulate the causal problem using the potential outcome framework [Rubin, 1974, Rosenbaum and Rubin, 1983]. For each subject we define a "full" random vector $(X_i, S_i, A_i, Y_i(0), Y_i(1))$, where S_i is a population indicator in source or target such that $S_i = 1$ for $i \in \mathcal{S}$ and $S_i = 0$ for $i \in \mathcal{T}$. The total sample size is $n = n_s + n_t$ and each subject assumed to be i.i.d. from a joint distribution of $(X, S, A, Y(0), Y(1))$. Moreover, $\mathcal{S}_0$ is used to denote the subjects in the source control group, and mathematically, $\mathcal{S}_0 = \{i : S_i = 1; A_i = 0\}$; $\mathcal{S}_1$ is defined for the source treated group similarly. According to Rosenbaum and Rubin [1983], we use the propensity score $\pi(x) = \mathbb{P}(A = 1|X = x, S = 1)$ to determine the treatment assignment mechanism. The main estimand in this paper, ATE of the target population, is

$$\tau^* = \mathbb{E}\{Y(1) - Y(0)|S = 0\}, \tag{3.1}$$

The following 3 standard assumptions are used which enable identification of causal effects within the source population.

Assumption 1. (*Stable Unit Treatment Value Assumption or SUTVA*) *There is no interference between different subjects and no hidden variation of treatments.*

Assumption 2. (*No unmeasured confounders of treatment assignment*) *In the source population,* $(Y(0), Y(1))$ *are conditionally independent of* A *given* X*:* $(Y(0), Y(1)) \perp\!\!\!\perp A|X, S = 1$.

Assumption 3. (*Positivity of treatment assignment*) *The propensity score of*

the source population is bounded away from 0 and 1: for some $c > 0$, $c \leqslant \pi(X) \leqslant 1 - c$ *almost surely.*

To extend the generalizability of the causal estimates to the target population, a key quantity is the participation probability between source and target defined as $\rho(x) = \mathbb{P}(S = 1|X = x)$. We further adopt two additional assumptions from Rudolph and van der Laan [2017] and Dahabreh et al. [2020].

Assumption 4. (*Mean exchangeability across populations*) *The conditional mean of the potential outcomes given the covariates are equal between the two populations:* $\mathbb{E}\{Y(a)|X, S = 1\} = \mathbb{E}\{Y(a)|X, S = 0\}$ *almost surely for* $a \in \{0, 1\}$.

Assumption 5. (*Positivity of participation probability*) *The participation probability is bounded away from 0:* $\rho(X) > c$ *almost surely for some* $c > 0$.

We further denote the conditional mean and variance of the potential outcomes in the source population as $\mu_a(x) = \mathbb{E}\{Y(a)|X = x, S = 1\}$ and $\sigma_a(x) = \text{Var}\{Y(a)|X = x, S = 1\}$. Under Assumption 4, we have $\mu_a(x) = \mathbb{E}\{Y(a)|X = x, S = 0\} = \mathbb{E}\{Y(a)|X = x\}$. The conditional average treatment effect (CATE) function is denoted as $\tau(x) \equiv \mu_1(x) - \mu_0(x)$.

3.3 Method

Gap in the existing work

Given Assumptions 1-5, we can estimate τ^* in terms of the observable from the source sample data by a difference of weighted outcomes as follows:

$$\hat{\tau}_w = \frac{1}{n_s}\sum_{i\in\mathcal{S}_1} w_i Y_i - \frac{1}{n_s}\sum_{i\in\mathcal{S}_0} w_i Y_i \ . \tag{3.2}$$

The weights $\{w_i : i \in \mathcal{S}\}$ take the following form [Chen et al., 2023]:

$$w_i = \left\{\frac{A_i}{\pi(X_i)} + \frac{1-A_i}{1-\pi(X_i)}\right\}\frac{E(S_i)(1-\rho(X_i))}{(1-E(S_i))\rho(X_i)}.$$

Directly estimation of w_i is usually computationally unstable. Without the individual data it is also infeasible. Therefore Chen et al. [2023] proposed a method for estimation of the weights as follows, based on entropy balancing weighting framework

$$\begin{aligned}
&\min_{w\succeq 0}\sum_{i\in\mathcal{S}} w_i \log w_i \\
\text{subject to}\quad &\frac{1}{n_s}\sum_{i\in\mathcal{S}_1} w_i h_k(X_i) = \bar{h}_{k,\mathcal{T}},\ k=1,\dots,K_h; \\
&\frac{1}{n_s}\sum_{i\in\mathcal{S}_0} w_i h_k(X_i) = \bar{h}_{k,\mathcal{T}},\ k=1,\dots,K_h; \\
&\frac{1}{n_s}\sum_{i\in\mathcal{S}_1} w_i g_k(X_i) = \frac{1}{n_s}\sum_{i\in\mathcal{S}_0} w_i g_k(X_i),\ k=1,\dots,K_g; \\
&\frac{1}{n_s}\sum_{i\in\mathcal{S}_1} w_i = \frac{1}{n_s}\sum_{i\in\mathcal{S}_0} w_i = 1.
\end{aligned} \tag{3.3}$$

In particular, functions $\{h_k : X \to \mathbb{R}; k = 1, \ldots, K_h\}$ are used to address covariate shift between source and target samples, while functions $\{g_k : X \to \mathbb{R}; k = 1, \ldots, K_g\}$ are employed to further correct for imbalances between the treatment and control groups within the source sample. From the theorem below, we can see that ideally the h_k functions should be chosen so that the linear span formed by them can at least cover treatment modifiers, even all outcome related variables if possible. The g_k functions should be chosen to complement h_k to determine the treatment assignment mechanism.

The weight normalization constraint at the last line of Equation (3.3) can be absorbed to the first two constraints by introducing $h_0(x) \equiv 1$. Denote $H = (h_0, h_1, ..., h_{K_h})$ and $G = (g_1, ..., g_{K_g})$. The following theorem is adopted directly from Chen et al. [2023] which originally listed 3 conditions under any of which could lead to consistency of the resulting weighting estimator for τ^*. Here we only list two of them as the other one was not as intuitive.

Theorem 3.1. *Suppose $\hat{w}$ is the solution of (3.3). If either of Conditions (a) or (b) below holds, $\hat{\tau}_{\hat{w}}$ is a consistent estimator of τ^*:*

Condition (a). $\mu_a(x) \in Span\{H(x)\}$, $a = 0, 1$.

Condition (b). $\log\{\pi(x)/(1 - \pi(x))\} \in Span\{H(x), G(x)\}$ *and* $\tau(x) \in Span(\{H(x)\})$.

Chen et al. [2023] further derived the asymptotic variance for $\hat{\tau}_w$.

However, estimation of the asymptotic variance directly from their formula requires individual covariate values in the target sample. We intend to overcome this limitation by introducing a resampling-based perturbation method for CI construction that do not require such information from the target sample.

Resampling-based perturbation for confidence interval construction

Parzen et al. [1994] introduced a straightforward resampling method for inference based on pivotal estimating functions within a semiparametric model framework. The authors demonstrated that for a broad class of estimating functions meeting two mild convergence conditions, a valid asymptotic CI could be constructed using the resampling method on the pivotal estimating functions. Hu and Kalbfleisch [2000] further broadened the idea by using bootstrapped general estimating functions for statistical inference. In particular, when the estimating functions are sums of independent terms, we can resample or bootstrap these terms to obtain an empirical distribution of the estimating functions. Solving the corresponding bootstrapped estimation equations then leads to valid statistical inference for the resulting estimators. Here, we extend this idea to our setting.

Since Equation (3.3) has constraints, we work with its dual problem

which is unconstrained for our purpose. In particular, we have the following characterization of the weights from Equation (3.3).

$$\hat{w}_i = \begin{cases} \exp\{\hat{\lambda}_1^\top H(X_i) + \hat{\gamma}^\top G(X_i)\}, & i \in \mathcal{S}_1 \\ \exp\{\hat{\lambda}_0^\top H(X_i) - \hat{\gamma}^\top G(X_i)\}, & i \in \mathcal{S}_0 \end{cases}$$

where $(\hat{\lambda}_1, \hat{\lambda}_0, \hat{\gamma}) \in \mathbb{R}^{K_h+1} \times \mathbb{R}^{K_h+1} \times \mathbb{R}^{K_g}$ is the solution to the dual problem:

$$\begin{aligned} \min_{\lambda_1, \lambda_0, \gamma} \ & \frac{1}{n_s} \sum_{i \in \mathcal{S}_1} \exp\{\lambda_1^\top H(X_i) + \gamma^\top G(X_i)\} + \frac{1}{n_s} \sum_{i \in \mathcal{S}_0} \exp\{\lambda_0^\top H(X_i) - \gamma^\top G(X_i)\} \\ & - (\lambda_1^\top + \lambda_0^\top) \bar{H}_{\mathcal{T}} \end{aligned} \tag{3.4}$$

Here $\bar{H}_{\mathcal{T}} = (\bar{h}_{0,\mathcal{T}}, ..., \bar{h}_{K_h,\mathcal{T}})$ with $\bar{h}_{0,\mathcal{T}} = 1$.

Equation (3.4) leads to the following first order condition to solve for $(\hat{\lambda}_1, \hat{\lambda}_0, \hat{\gamma})$:

$$\begin{aligned} {n_s}^{-1} \textstyle\sum_{i \in \mathcal{S}_1} H(X_i) \exp\{\lambda_1^\top H(X_i) + \gamma^\top G(X_i)\} - \bar{H}_{\mathcal{T}} = 0 \\ {n_s}^{-1} \textstyle\sum_{i \in \mathcal{S}_0} H(X_i) \exp\{\lambda_0^\top H(X_i) - \gamma^\top G(X_i)\} - \bar{H}_{\mathcal{T}} = 0 \\ \textstyle\sum_{i \in \mathcal{S}_1} G(X_i) \exp\{\lambda_1^\top H(X_i) + \gamma^\top G(X_i)\} - \\ \textstyle\sum_{i \in \mathcal{S}_0} G(X_i) \exp\{\lambda_0^\top H(X_i) - \gamma^\top G(X_i)\} = 0 \end{aligned} \tag{3.5}$$

Therefore, if we can use bootstrap to capture the variance of the estimating equations in (3.5), we can back-propagate the estimation error to the estimated weights, enabling us to construct a CI for the estimator $\hat{\tau}_w$ in (3.2). The classic bootstrap [Efron, 1979] can be applied to the elements in the source population $\{H(X_i), G(X_i)\}$ to generate bootstrapped versions $\{H(X_i)^{(b)}, G(X_i)^{(b)}\}$ for $b = 1, \ldots, B$. However for the summary

level information $\bar{H}_{\mathcal{T}}$, we resort to parametric bootstrap [Efron, 2012]. Because $\bar{H}_{\mathcal{T}}$ are sample averages, we assume that $\bar{H}_{\mathcal{T}} \sim N(\boldsymbol{\mu}_{\bar{H}}, \boldsymbol{\Sigma}_{\bar{H}})$ asymptotically. Therefore when $\boldsymbol{\Sigma}_{\bar{H}}$ is available from the target sample, we can draw $\bar{H}_{\mathcal{T}}^{(b)}$ from the multivariate normal distribution with mean $\bar{H}_{\mathcal{T}}$ and variance-covariance matrix $\boldsymbol{\Sigma}_{\bar{H}}$. It is more common that only the diagonal elements of $\boldsymbol{\Sigma}_{\bar{H}}$ is available, especially if the summary information is from published literature. Then we propose to estimate the correlation matrix corresponding to $\boldsymbol{\Sigma}_{\bar{H}}$ using the individual data from the source population.

Algorithm 1 Resampling-based perturbation method for CI construction (RPM-CI)

procedure RPM-CI($\{(X_i; A_i; Y_i) : i \in \mathcal{S}\}$, $\bar{H}_{\mathcal{T}}$, $\text{var}(\bar{H}_{\mathcal{T}})$)
Estimate correlation of target moments $\hat{R}_{\bar{H}_{\mathcal{T}}} = \text{corr}(H(X_i)), i \in \mathcal{S}$.
Estimate $\hat{\tau}_w$ using $\{(X_i; A_i; Y_i) : i \in \mathcal{S}\}$ and $\bar{H}_{\mathcal{T}}$ by Equations (3.2) and (3.3).
for each $b = 1, \ldots, B$ **do**
Draw $\{(X_i^{(b)}; A_i^{(b)}; Y_i^{(b)}) : i \in \mathcal{S}\}$ from the source population.
Generate perturbed means $\bar{H}_{\mathcal{T}}^{(b)} \sim \mathcal{N}(\bar{H}_{\mathcal{T}}, \text{var}(\bar{H}_{\mathcal{T}})^{1/2} \hat{R}_{\bar{H}_{\mathcal{T}}} \text{var}(\bar{H}_{\mathcal{T}})^{1/2})$.
Estimate $\{\hat{w}_i^{(b)} : i \in \mathcal{S}\}$ using $\{(X_i^{(b)}; A_i^{(b)}; Y_i^{(b)}) : i \in \mathcal{S}\}$ and $\bar{H}_{\mathcal{T}}^{(b)}$ by Equation (3.3).
Estimate $\hat{\tau}_w^{(b)}$ using $\{\hat{w}_i^{(b)} : i \in \mathcal{S}\}$ and $\{(X_i^{(b)}; A_i^{(b)}; Y_i^{(b)}) : i \in \mathcal{S}\}$ by Equation (3.2).
end for
Construct a 95% CI based on 2.5 and 97.5 percentiles of $\hat{\tau}_w^{(b)}, b = 1, \ldots, B$.
end procedure

We formalize the above proposed resampling-based perturbation method

to construct the confidence interval (RPM-CI) in Algorithm 1. For a particular data set, in step 1, we estimate the correlation of target moments $\hat{R}_{\bar{H}_{\mathcal{T}}} = \text{corr}(H(X_i)), i \in \mathcal{S}$ using the source data and estimate the corresponding target ATE $\hat{\tau}_w$. In step 2, for each b^{th} over B iteration, we first sample the source population with replacement as $\{(X_i^{(b)}; A_i^{(b)}; Y_i^{(b)}) : i \in \mathcal{S}\}$. Then we use multivariate normal distribution to perturb the target sample mean $\bar{H}_{\mathcal{T}}$ and generate target data perturbed means $\bar{H}_{\mathcal{T}}^{(b)} \sim \mathcal{N}(\bar{H}_{\mathcal{T}}, \text{var}(\bar{H}_{\mathcal{T}})^{1/2}\hat{R}_{\bar{H}_{\mathcal{T}}}\text{var}(\bar{H}_{\mathcal{T}})^{1/2})$ assuming $\text{var}(\bar{H}_{\mathcal{T}})$ is available from the target sample. Next, we use the simulated $\{(X_i^{(b)}; A_i^{(b)}; Y_i^{(b)}) : i \in \mathcal{S}\}$ and $\bar{H}_{\mathcal{T}}^{(b)}$ to estimate weights $\{\hat{w}_i^{(b)} : i \in S\}$ and its corresponding $\hat{\tau}_w^{(b)}$ for $b = 1, \ldots, B$ using Equation (3.3) and (3.2). Finally, we construct the CI for our estimator $\hat{\tau}_w$ based on 2.5 and 97.5 percentiles of $\hat{\tau}_w^{(b)}, b = 1, \ldots, B$.

Resampling-based perturbation method with approximate balancing

In practice, the exact balancing approach may not always produce a feasible solution due to finite sample. Therefore, Wang and Zubizarreta [2019] advocated a more flexible approach for covariate balancing weight construction for causal inference. The approach can be extended in a

straight-forward fashion to our causal generalization setting as follows.

$$
\begin{aligned}
\min_{w_i \geq 0} \quad & \sum_{i \in S} w_i \log w_i \\
\text{subject to} \quad & \left| \frac{1}{n_s} \sum_{i \in \mathcal{S}_1} w_i h_k(X_i) - \bar{h}_{k,\mathcal{T}} \right| \leqslant \delta_1, k = 1, ..., K_h; \\
& \left| \frac{1}{n_s} \sum_{i \in \mathcal{S}_0} w_i h_k(X_i) - \bar{h}_{k,\mathcal{T}} \right| \leqslant \delta_1', k = 1, ..., K_h; \\
& \left| \frac{1}{n_s} \sum_{i \in \mathcal{S}_1} w_i g_k(X_i) - \frac{1}{n_s} \sum_{i \in \mathcal{S}_0} w_i g_k(X_i) \right| \leqslant \delta_2, k = 1, ..., K_g; \\
& \frac{1}{n_s} \sum_{i \in \mathcal{S}_1} w_i = \frac{1}{n_s} \sum_{i \in \mathcal{S}_0} w_i = 1.
\end{aligned}
\tag{3.6}
$$

Therefore the exact balancing constraints is relaxed in Equation (3.6) by introducing $\delta_1, \delta_1' \in \mathbb{R}^{K_h}$ and $\delta_2 \in \mathbb{R}^{K_g}$. The weight normalization constraint at the last line of Equation (3.6) again can be absorbed to the first two constraints by introducing an extra element $\bar{h}_{0,\mathcal{T}} = 1$ and setting the corresponding relaxation $\delta_{1,0}, \delta_{1,0}' \equiv 0$. This flexibility trades bias for variance and offers two key advantages: it enables us to incorporate a broader set of covariate functions, and it helps overcome computational challenges when exact balancing is infeasible during the construction of CIs with the resampling-based perturbation.

Naturally, a practical consideration when using approximate balancing is how to determine the appropriate degree of approximate balance. Chattopadhyay et al. [2024] advocated using a constant factor (i.e., 0.1 times)

of each covariate's standard deviation. Instead of the standard deviation, we advocate the following strategy based on the dual problem of Equation (3.6).

In particular, the dual of Equation (3.6) takes the following form:

$$\tilde{w}_i = \begin{cases} \exp\{\tilde{\lambda}_1^\top H(X_i) + \tilde{\gamma}^\top G(X_i)\}, & i \in \mathcal{S}_1 \\ \exp\{\tilde{\lambda}_0^\top H(X_i) - \tilde{\gamma}^\top G(X_i)\}, & i \in \mathcal{S}_0 \end{cases}$$

where $\tilde{\lambda}_0, \tilde{\lambda}_1, \tilde{\gamma}$ minimize

$$\min_{\lambda_1,\lambda_0,\gamma} \frac{1}{n_s}\sum_{i\in\mathcal{S}_1} \exp\{\lambda_1^\top H(X_i) + \gamma^\top G(X_i)\} + \frac{1}{n_s}\sum_{i\in\mathcal{S}_0} \exp\{\lambda_0^\top H(X_i) - \gamma^\top G(X_i)\} - \lambda_1^\top \bar{H}_{\mathcal{T}} - \lambda_0^\top \bar{H}_{\mathcal{T}} + |\lambda_1|^\top \delta_1 + |\lambda_0|^\top \delta_1' + |\gamma|^\top \delta_2. \tag{3.7}$$

Compared with the dual form (3.4) for the exact balancing, (3.7) contains three additional L_1 regularization terms for the dual parameters: $|\lambda_1|^\top \delta_1 + |\lambda_0|^\top \delta_1' + |\gamma|^\top \delta_2$. Thus, we propose to use the Adaptive LASSO [Zou, 2006] to determine the degree of approximate balancing. In particular assume that we have estimates $(\hat{\lambda}_1, \hat{\lambda}_0, \hat{\gamma})$ from (3.4) based on the exact balancing problem. Then for resampling-based perturbations that have no exact balancing solution or are infeasible for (3.3), we use fractions of $(\hat{\lambda}_1, \hat{\lambda}_0, \hat{\gamma})$ for $(\delta_1, \delta_1', \delta_2)$. The details are listed in Algorithm 2.

When there is no exact balancing solution for the original exact balancing problem $\hat{\tau}_w$, we advocate using Chattopadhyay et al. [2024] idea of allowing imbalances to be up to a constant factor (i.e., 0.1 times) each

covariate's standard deviation until we get a solution and then follow Algorithm 2 for CI construction.

Algorithm 2 Resampling-based perturbation with approximate balancing (RPM-AB)

procedure RPM-AB($\{(X_i; A_i; Y_i) : i \in \mathcal{S}\}$, $\bar{H}_{\mathcal{T}}$, $\text{var}(\bar{H}_{\mathcal{T}})$)
  Estimate correlation of target moments $\hat{R}_{\bar{H}_{\mathcal{T}}} = \text{corr}(H(X_i)), i \in \mathcal{S}$.
  Estimate $\{\hat{w}_i : i \in \mathcal{S}\}$, $(\hat{\lambda}_1, \hat{\lambda}_0, \hat{\gamma})$ and $\hat{\tau}_w$ using Equation (3.2) and (3.3).
  for each $b = 1, \ldots, B$ **do**
    Draw $\{(X_i^{(b)}; A_i^{(b)}; Y_i^{(b)}) : i \in \mathcal{S}\}$ from the source population.
    Generate perturbed means $\bar{H}_{\mathcal{T}}^{(b)} \sim \mathcal{N}(\bar{H}_{\mathcal{T}}, \text{var}(\bar{H}_{\mathcal{T}})^{1/2} \hat{R}_{\bar{H}_{\mathcal{T}}} \text{var}(\bar{H}_{\mathcal{T}})^{1/2})$.
    Estimate $\{\hat{w}_i^{(b)} : i \in \mathcal{S}\}$ using $\{(X_i^{(b)}; A_i^{(b)}; Y_i^{(b)}) : i \in \mathcal{S}\}$ and $\bar{H}_{\mathcal{T}}^{(b)}$ by Equation (3.3).
    if $\{\hat{w}_i^{(b)} : i \in \mathcal{S}\}$ could not admit a solution **then**
      $c \leftarrow 0$
      while $\{\hat{w}_i^{(b)} : i \in S\}$ not admitting solutions **do**
        $\delta_1 \leftarrow (c + 0.1)|\frac{1}{\hat{\lambda}_1}|$, $\delta_1' \leftarrow (c + 0.1)|\frac{1}{\hat{\lambda}_0}|$, $\delta_2 \leftarrow (c + 0.1)|\frac{1}{\hat{\gamma}}|$
        Estimate $\{\hat{w}_i^{(b)} : i \in \mathcal{S}\}$ using $\delta_1, \delta_1', \delta_2$ and Equation (3.6).
      end while
    end if
    Estimate $\hat{\tau}_w^{(b)}$ using $\{\hat{w}_i^{(b)} : i \in \mathcal{S}\}$ and $\{(X_i^{(b)}; A_i^{(b)}; Y_i^{(b)}) : i \in \mathcal{S}\}$ by Equation (3.2).
  end for
  Construct the 95% CI based on 2.5 and 97.5 percentiles of $\hat{\tau}_w^{(b)}$, $b = 1, \ldots, B$.
end procedure

3.4 Evaluation with simulated settings

In this section, we conduct simulation studies to evaluate the performance of the proposed methods in finite sample settings. For each simulation set-up, we could estimate τ^* using Equation (3.1). Then, for each m over a total of M simulations, we estimate $\hat{\tau}_w^{(m)}$ and its corresponding 95% CI using methods we proposed in Algorithm 1, or Algorithm 2 if exact balancing is infeasible. The performance is measured in terms of bias of $E(\hat{\tau}_w^{(m)})$ and the empirical coverage of τ^* within the 95% CI constructed by the proposed methods.

Exact balancing and Algorithm 1 evaluation

We first examine exact balancing CI construction method in Algorithm (1) with simulation settings that always have feasible solutions for (3.3) and the corresponding resampling-based perturbations. We set the total sample size $n = n_s + n_t = 800$ with the bootstrap iteration $B = 1000$ and $M = 500$ simulated data sets. In our simulations, due to random sampling, the source sample size n_s varies between 350 and 450 observations. We generate 5 covariates $X = (X_1, \ldots, X_5)$ from a uniform distribution $U(-2, 2)$. We consider the case when the covariates are independent of each other and the case when correlation among them are 0.1 and 0.3.

In the target sample we only have summary-level information of X_1, X_2, and X_3. We set $H(x) = (1, x_1, x_2, x_3)$ and $G(x) = (x_4, x_5)$. We consider

balancing on the first moments of all covariates.

In light of Theorem 1, we consider scenarios when the conditions for consistency hold and also when none of them holds. We consider the settings when either Condition (a) or Condition (b) holds or neither of them holds.

Therefore for the propensity score model, we first assume a scenario when the treatment assignment is related to H linearly with $\text{logit}\{\pi(x)\} = 0.7x_2 + 0.5x_3$. In this case, all the confounders are included in H, and it is enough that we only balance on H to account for confounding. We also assume a scenario when the propensity score is related to H and G nonlinearly with $\text{logit}\{\pi(x)\} = 0.35x_2 - 0.4\max(x_3, x_4) - 0.7x_5$.

For the outcome model, we assume it has the form of $Y_i = m(X_i) + (A_i - 0.5)\tau(X_i) + \epsilon_i$ with $\epsilon_i \overset{\text{i.i.d}}{\sim} N(0,1)$. We assume the CATE function comes from the following settings:

(T1) $\tau(x) = x_1 - 0.6x_2 - 0.4x_3$.

(T2) $\tau(x) = x_1 - 0.6x_2 - 0.4x_3 + 0.8x_4 - 0.3x_5$.

(T3) $\tau(x) = x_1 - 0.5\exp(x_2 - 0.8x_3)$

We assume the main effect $m(x)$ comes from the following settings:

(M1) $m(x) = 0.5x_1 + 0.3x_2 + 0.3x_3$.

(M2) $m(x) = 0.5x_1 + 0.3x_2 + 0.3x_3 - 0.4x_4 - 0.7x_5$.

(M3) $m(x) = 0.5x_1 + 0.8x_2^2 + 0.2\exp(0.5x_3 - x_4 - 1) - 0.7x_5$.

When the propensity score lies within the linear span of H and the CATE function follows (T1), $\tau(x)$ is linearly related to H and satisfies the

consistency Condition (b) in Theorem 1, regardless of the main effect settings. However, this condition does not hold under (T2) or (T3) because (T2) depends on both H and G, while (T3) is nonlinearly related to H. When the propensity score is not within the linear span of H, but the outcome satisfies (T1) and (M1), $\mu_a(x)$ remains linear in H, thereby meeting Condition (a). Outside these scenarios, neither Condition (a) nor Condition (b) holds, which may introduce bias in the estimator $\hat{\tau}_w$.

For covariate shift, similar to the propensity score model, we also consider a linear setting when the participation probability is $\text{logit}\{\rho(x)\} = 0.4x_1 + 0.3x_2 - 0.2x_4$. That is, there is shift in the distribution of (X_1, X_2, X_4). We also consider a nonlinear setup when the participation probability is $\text{logit}\{\rho(x)\} = 0.3x_1 + 0.5x_2 \cdot x_4 - 0.2x_4$.

The performance of our method is summarized in Table 3.1 for independent covariates and Table 3.2 for correlated covariates. For bias evaluation, we see that the bias of average $\hat{\tau}_w$ for our method is ignorable under linear and nonlinear settings when the consistency conditions are met. In terms of the CI construction, when the consistency conditions are met, we find that under both linear and nonlinear settings, the constructed CI by Algorithm 1 can cover around 95% of the time. Even when the consistency conditions are not fully met, our method maintains approximately 95% coverage as long as the estimator is not severely biased. However, when the estimator exhibits significant bias, the constructed confidence interval (CI) results in lower coverage of τ^*.

Table 3.1: Empirical evaluation for Target ATE estimation and CI coverage using the RPM-CI method (independent covariates)

Settings		Consistency Condition	τ^*	Empirical coverage of τ^*	Average $\hat{\tau}_w$ (95% CI)
Linear	T1+M2	b	-0.140	95.2%	-0.131 (-0.422, -0.159)
	T1+M3	b	-0.138	94.4%	-0.160 (-0.545, 0.213)
	T2+M1	No	-0.039	76%	-0.259 (-0.579, -0.059)
Nonlinear	T1+M1	a	-0.179	94.4%	-0.182 (-0.451, 0.083)
	T3+M1	No	-1.525	94.2%	-1.496 (-1.843, -1.164)
	T1+M3	No	-0.179	93.2%	-0.188 (-0.538, 0.157)

Table 3.2: Empirical evaluation for Target ATE estimation and CI coverage using the RPM-CI method (correlated covariates)

Settings		Consistency Condition	τ^*	Covariate correlation	Empirical coverage of τ^*	Average $\hat{\tau}_w$ (95% CI)
Linear	T1+M2	b	-0.126	0.1	94.8%	-0.122 (-0.410, 0.164)
			-0.097	0.3	94.5%	-0.096 (-0.382, 0.190)
	T2+M1	No	-0.054	0.1	76.8%	-0.249 (-0.566, 0.066)
			-0.079	0.3	79.8%	-0.251 (-0.560, 0.055)
Nonlinear	T1+M1	a	-0.166	0.1	94%	-0.175 (-0.437, -0.083)
			-0.140	0.3	94.6%	-0.144 (-0.391, 0.100)
	T1+M3	No	-0.169	0.1	94%	-0.176 (-0.518, 0.164)
			-0.143	0.3	94.6%	-0.139 (-0.467, 0.186)

Approximate balancing and Algorithm 2 evaluation

Now we consider settings that Algorithm 2 needs to be invoked due to infeasibility of perturbed Equation (3.4), in particular in smaller sample size settings with noisy covariates. We set the total sample size $n = n_s + n_t = 400$ with bootstrap iteration $B = 800$, and generate covariates $X = (X_1, \ldots, X_5)$ from uniform distribution $U(-2, 6)$. The rest of the settings are the same as in the previous subsection.

The performance of our methods is summarized in Table 3.3. We also report the percent of non-feasible solutions over $M \times B$ iterations for exact

balancing. In terms of bias evaluation, for exact balancing, even though some simulations may not admit solutions, the bias of $\hat{\tau}_w$ is small when the consistency conditions are met. When we use approximate balancing for target ATE estimation simulation cases with no exact balancing solution, the bias is larger as it trades bias for variance. In terms of the CI construction, when most of the perturbations admit an exact balancing solution, our findings about the CI coverage are consistent with what we observe in Table 3.1 and Table 3.2. Especially, when the consistency conditions are satisfied, we find that both exact and approximate balancing CI could cover τ^* around 95% times. However, regardless of the consistency conditions, if we could not admit enough feasible solutions during the CI construction process, we would get poor CI coverage for exact balancing. The CI constructed by the approximate balancing method is wider and thus could help in this situation with a better coverage.

Table 3.3: Empirical evaluation for Target ATE estimation and CI coverage using RPM-CI and RPM-AB methods

Settings	Consistency Condition	τ^*	Methods	% Infeasible	Empirical coverage τ^*	Average $\hat{\tau}_w$ (95% CI)
Linear	(b):T1+M2	-0.652	RPM-CI	26.0%	87.3%	-0.660 (-1.335, 0.125)
			RPM-AB	0%	92.6%	-0.624(-1.414, 1.174)
	No: T2+M1	0.800	RPM-CI	26.9%	60.9%	-0.017(-1.029, 1.075)
			RPM-AB	0%	90.2%	0.005(-0.958, 1.890)
Nonlinear	(a):T1+M1	0.038	RPM-CI	5.6%	95.2%	0.034(-0.702, 0.767)
			RPM-AB	0%	94.6%	0.030(-0.806, 0.808)
	No: T2+M2	0.539	RPM-CI	4.9%	70.4%	1.083(0.256, 2.036)
			RPM-AB	0%	74.2%	1.095(0.223, 2.212)

3.5 Cross-validation based evaluation with a real setting

Here, we use the same example as in Chen et al. [2023], derived from the MIMIC-III database [Johnson et al., 2016], but with CI construction and evaluation. However, due to the fact that we don't know the true target ATE, we employ a cross-validation based strategy as we explain below.

This observational dataset comprises 6,361 ICU patients, with 51.3% having undergone transthoracic echocardiography (TTEC) either during or within 24 hours before ICU admission. The primary outcome of interest is 28-day survival. Our goal is to assess the effect of TTEC on the survival of ICU patients with sepsis.

The dataset encompasses demographic details, such as age, gender, and weight, along with severity at admission measured by the Simplified Acute Physiology Score (SAPS), Sequential Organ Failure Assessment (SOFA) score, and Elixhauser comorbidity score. Additionally, it includes comorbidity indicators (denoted as cmb_i), including congestive heart failure, atrial fibrillation, respiratory failure, and malignant tumor. Vital signs like mean arterial pressure, heart rate, and temperature, as well as laboratory results, are also part of the dataset. To address right-skewed distributions of lab results, a log transformation is applied, and standardization is employed for continuous variables. Missing values are addressed through imputation using the missForest method, which

is a flexible non-parametric missing value imputation approach with no assumptions needed [Stekhoven and Bühlmann, 2011].

As the real data is observed only once and we do not know the true treatment effect, we create a cross-validation (CV) based sampling procedure for evaluation of bias and coverage. In particular, we partition the data set into subsets, train the model on some of these subsets, and then evaluate its performance on the remaining subset.

Figure 3.1 is a diagram showing the whole workflow of our CV-based evaluation. For each round of CV, we first partition p_S proportion of the total study population into the source population, with the remainder as the target population. The partition or sampling probability is in proportion to a function Ψ, which includes some important effect modifiers. Then, within the source population, we generate source samples using a function g, which includes some key confounding factors. In particular, p_{S_1} proportion of the treated and p_{S_0} of the control populations become treated and control samples. Next, we randomly split the target population with p_T proportion into target and the rest into test samples. Suppose in the target sample, we only know the summary level information while in the test sample we know all the individual level information. The above steps will be repeated many times and the source treated, control, and target samples will be used as the training data to estimate target population treatment effect and CI. To obtain an oracle estimate of the target population ATE, we repeat the splitting procedures and use the test sample

to construct a CV-based treatment effect τ^*_{cv} as our oracle estimator. We will use this τ^*_{cv} as the benchmark to evaluate the bias and CI of the target population treatment effect. The rest of the section illustrates the details of this CV-based procedure used in our real-setting.

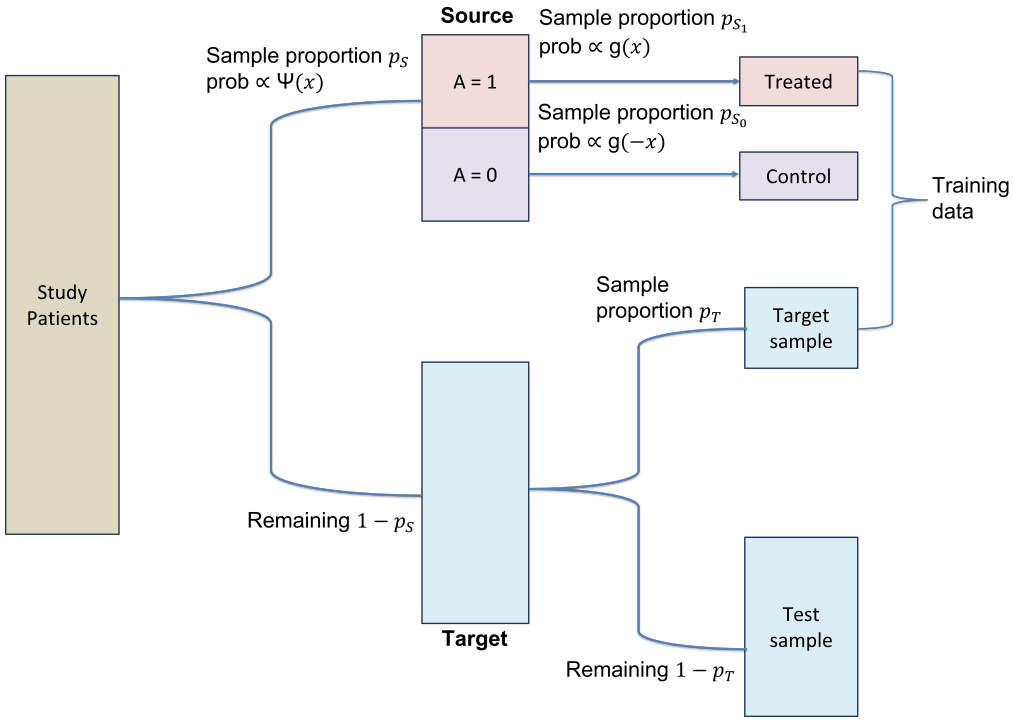

Figure 3.1: CV-based evaluation workflow in a real setting

To comprehensively assess across diverse scenarios, we manipulate various degrees of confounding and covariate shift while keeping the covariate-outcome relationship unchanged in the actual data. Initially, we select $p_S = 40\%$ of the entire dataset with probability proportional to $\Psi(x)$ to form the source population. The remaining data is then randomly divided into a target sample, with a $p_T = 1/3$ probability, and a test sample.

The probability of being selected as a source sample is proportional to

$$\Psi\{\kappa_S(-0.3 \times \text{age} + 0.3 \times \text{cmb}_1 + 0.4 \times \text{cmb}_2 + 0.3 \times \text{cmb}_3 + 0.4 \times \text{cmb}_4 - 0.5)\}\,.$$

where $\Psi(x) = 0.8\Phi(x) + 0.1$ with $\Phi(x)$ being the standard normal CDF. Here, κ_S is a parameter to reflect different levels of covariate shift, where κ_S equals 1 for small covariate shift and 5 for large covariate shift. Under this sampling design, the source population is younger and more likely to have comorbidities than the target population. Among the source samples, we randomly select $p_{S_1} = 1/2$ of the TTEC patients and $p_{S_0} = 1/2$ of the non-TTEC patients to form the source sample. The TTEC patients are selected with probability proportional to $g(x)$ while the non-TTEC patients are selected with probability proportional to $g(-x)$ with

$$g(x) = \Psi\{\kappa_{\mathcal{A}}(0.3 \times \text{SAPS} + 0.4 \times \text{SOFA} - 0.5 \times \text{Elixhauser})\}\,.$$

We consider two choices of $\kappa_{\mathcal{A}}$ in g: (a) $\kappa_{\mathcal{A}} = 0$, so all the patients in this step are sampled with equal probability; (b) $\kappa_{\mathcal{A}} = 1$, which induces additional confounding determined by a linear combination of the severity scores.

In total, we have four settings. Under each one, we run $M = 500$ times replications of the above CV procedures and $B = 1000$ perturbations using the exact balancing method to construct the estimator $\hat{\tau}_{cv}$ and its confidence intervals. We assume that the target sample only includes information about the average and variance of the demographic covariates

and comorbidity indicators. The oracle estimate of the target population ATE τ^*_{cv} is estimated using the entropy balancing weighting method incorporating full information in the test data. To be more specific, we repeat the above CV procedures 8000 times and get the estimated target ATE. The average of it would be our oracle estimator of the target population ATE τ^*_{cv}. The CV splitting set-up here makes a good overlap between source and target samples. Thus, we did not encounter the infeasibility issue for exact balancing.

For evaluation, we examine the bias and percentage of times the constructed CI covers the oracle estimator τ^*_{cv}. Table 3.4 summarizes τ^*_{cv} empirical coverage percentage and the estimator $\hat{\tau}_{cv}$ under each scenario. As we can see, the proposed method could cover the oracle estimator around 95% of the time. In evaluating bias, we find that the bias of empirical $\hat{\tau}_{cv}$ is ignorable under different settings.

Table 3.4: Target ATE estimation and CI coverage in a real setting

Setting		τ^*_{cv}	Empirical	Average
Confounding	Covariate shift		coverage of τ^*_{cv}	$\hat{\tau}_{cv}$ (95% CI)
Extra ($\kappa_A = 1$)	Small ($\kappa_s = 1$)	0.051	96.4%	0.048 (-0.006, 0.102)
No ($\kappa_A = 0$)	Small ($\kappa_s = 1$)	0.050	96%	0.052 (0.001, 0.104)
Extra ($\kappa_A = 1$)	Big ($\kappa_s = 5$)	0.049	95.6%	0.052 (-0.005, 0.107)
No ($\kappa_A = 0$)	Big ($\kappa_s = 5$)	0.049	95.8%	0.054 (0.001, 0.107)

3.6 Discussion and conclusion

We have developed a resampling-based perturbation method for CI construction to make inference about generalizing ATE estimation to a target population. It is an important step to complement the work of Chen et al. [2023] to quantify the uncertainty associated with the estimated treatment effect for the target population. Although we require slightly more information from a target sample than Chen et al. [2023] did, our requirement is minimum as we only need the variance of the summary statistics $\bar{H}_{\mathcal{T}}$. Note that for binary and discrete variables, such variance is not needed as we can directly use $\bar{H}_{\mathcal{T}}$ to estimate its variance. When the target sample's individual data is available but can not be shared due to privacy reasons, then requesting this further information is relatively straightforward.

To achieve an unbiased causal generalization, exact balancing is essential, as it ensures that covariates are equally balanced between populations. For the CI construction using the resampling-based perturbation method, exact balancing should be prioritized because it directly aligns with the goal of unbiased causal generalization by precisely matching covariate distributions. However, when a feasible solution for exact balancing is unattainable due to sample size limitations or high-dimensional covariates, approximate balancing can be a practical alternative. Although it may introduce a small bias, approximate balancing provides a close solution that maintains the integrity of the analysis by minimizing discrepancies in

covariate distributions. Therefore, we recommend approximate balancing only as a secondary option, to be used when exact balancing solutions are not feasible.

Data Availability Statement

The data that support the findings in this paper were derived from the following resources available in the public domain: MIMIC-III Clinical Database Version1.4 (https://physionet.org
/content/mimiciii/1.4/)

4 AN R PACKAGE: EBALGEN

4.1 Overview

This chapter introduces an R package *EBalGen*, which is used to implement the exact and approximate balancing methods in Chapter 3 for causal generalization in the presence of covariate shift using target sample summary-level information. This package is designed to estimate causally generalized balancing weights, the target Average Treatment Effect (ATE), and its corresponding confidence interval (CI). It provides flexibility in achieving both exact and approximate balance generalizing causal findings from source to target population when we only have summary level information of the target. We illustrate the implementation of this package across various scenarios. The key functions of this package is `ebal_wts()` for estimating weights and *ebal_ATE*() for ATE estimation. *RPM_CI*() and *RPM_AB*() are important functions for exact and approximate balancing CI estimation. Our package is available on github https://github.com/yc702/EBalGen and it passes R-CMD-check.

4.2 Package dependencies

EBalGen was developed with dependence on 9 packages:

- *parallel*, *doParallel*, *foreach* and *doRNG* are essential tools for imple-

menting parallel computing, as most of our methods are based on simulations [R Core Team, 2023, Microsoft and Weston, 2022a,b, Gaujoux, 2023]. *dplyr* is also used for efficient data manipulation [Wickham et al., 2023].

- *CVXR* provides functions to solve convex optimization problems, and is compatible with different solvers. In our problem, we use *MOSEK*, which is a commercial high-performance solver for large-scale convex optimization problems. It is numerically stable and could efficiently solve exponential cone problem, which can be challenging for some other solvers [Fu et al., 2020, MOSEK-ApS, 2024].

- *resample* provides essential functions for resampling-based inference, enabling our method to perform resampling-based perturbation of target sample moments and estimate CI. [Hesterberg, 2022].

- *stats* offers powerful functions to estimate correlation structures and extract quantiles from sampling distributions [R Core Team, 2023].

- *rockchalk* provides functions to perturb the target sample moments and generate multivariate normal distributed random variables [Johnson, 2022].

4.3 Key functions

The first important function is *ebal_wts*(), which is used to compute the exact and approximate entropy balancing weights. Another similar function *ebal_wts_simple* () is used to compute the weights calibrating the whole source sample to the target moments without distinguishing source treated and control groups. Since these two functions contain similar input arguments, we mainly discuss *ebal_wts*() for demonstration purpose.

- *x* A data matrix for the source sample. Each column represents source sample covariate and each row represents an observation.
- *trt* A vector of 0, 1 or FALSE/TRUE of treatment assignment for the source sample.
- *H_vars* A vector of numbers indexing which covariate in *x* need to be balanced between source and target samples.
- *target_moments* A vector of first moments of the target sample covariates that needs to be balanced between source and target.
- *H_add_intercept* A logical value determines whether to include 1 as intercept in H covariates, default as TRUE.
- *delta* A vector specifying the approximate balancing tolerance margin. The vector has a total length of $H + H + G$, where H represents the number of covariates balanced between the source (treatment and

control) and the target moments, and G represents the covariates balanced solely between the source treatment and control groups. If we are doing exact balancing, delta are all zeros.

The output of the function returns a list containing w, which is a vector of entropy balancing weights and *theta* which is the dual parameters estimated from the optimization process.

The second key function is *ebal_ATE*(), which is used to compute the exact and approximate balancing ATE. If the tolerance margin argument *delta* is all 0, it computes the exact balancing ATE. Otherwise, it computes the approximate balancing ATE.

As we want this function to return a feasible solution, if exact balancing does not yield a feasible solution, the standard deviation of x is used as the input argument *delta*, which convert exact into approximate balancing. If the specified *delta* does not yield a feasible solution, for approximate balancing, the constant c is increased (starting from 1) by 1 times *delta* until a solution is found. For exact balancing that later uses the standard deviation for *delta*, the constant is increased (starting from 0) by 0.1 times *delta* until a solution is achieved.

- *x, trt, H_vars, H_add_intercept, target_moments, delta* are the same as the input for function *ebal_wts*().

- *y* A vector of the source sample response values.

The output of the function returns a list containing *ate_est*, which is the target ATE for causal generalization. and *constant* which is the final constant c used for the approximate balancing tolerance margin if no feasible solution is achieved with the specified *delta*. If the specified *delta* results in a feasible solution, the constant remains 0. Otherwise, the constant will increase until a feasible solution is found.

The third major function is *RPM_CI*(), which is used to compute the exact balancing CI according to Chapter 3 Algorithm 1.

- *x, y, trt, H_vars, H_add_intercept, delta* are the same as the input for functions *ebal_ATE*().
- *target_mean* A vector of means of the target sample covariates that needs to be balanced between source and target samples.
- *target_sd* A vector of standard deviations of the target sample covariates that needs to be balanced between source and target samples.
- *num_sim* A numeric value shows the number of simulations used in resampling-based perturbation.
- *cluster* Number of parallel running CPU cores, Default: 1.
- *with_seed* Random seed for simulation, Default: 111.

The output of the function returns a list containing *mean_ATE*, which is the mean ATE over *num_sim* perturbations; *lb_ATE* and *ub_ATE*, which

are the lower and upper bounds of CI and *n_success* which is the number of feasible solutions in *num_sim* perturbations if using exact balancing.

The last key function is *RPM_AB*(), which is used to compute the approximate balancing CI according to Chapter 3 Algorithm 2. The input arguments are the same as function *RPM_CI*(). The output of the function also returns a list containing *mean_ATE*, *lb_ATE* and *ub_ATE*. In addition, *n_success* is the number of feasible solutions in *num_sim* perturbations if using approximate balancing. *use_exact* is the number of times that exact balancing could be achieved in the perturbations.

4.4 Example implementation

For the example implementation, we will use scenarios similar to the simulation setting described in Chapter 3. Suppose we set the total sample size $n = n_s + n_t = 800$, which is split into source $n_s = 401$ and target $n_t = 399$ samples. We generate 5 covariates $X = (X_1, \ldots, X_5)$ from a uniform distribution $U(-2, 2)$. The source/target participation probability $\rho(x)$ follows $\text{logit}\{\rho(x)\} = 0.4x_1 + 0.3x_2 - 0.2x_4$. That is, there is shift in the distribution of (X_1, X_2, X_4). For the propensity score $\pi(x)$ model, we assume the treatment assignment is related to H linearly with $\text{logit}\{\pi(x)\} = 0.7x_2 + 0.5x_3$. In this case, all the confounders are included in H, and it is enough that we only balance on H to account for confounding. For the outcome model, we assume $Y_i = m(X_i) + (A_i - 0.5)\tau(X_i) + \epsilon_i$ with

$\epsilon_i \overset{i.i.d}{\sim} N(0,1)$. For the CATE function, we assume $\tau(x) = x_1 - 0.6x_2 - 0.4x_3$. For the main effect $m(x)$, it has the form of $m(x) = 0.5x_1 + 0.3x_2 + 0.3x_3 - 0.4x_4 - 0.7x_5$. In this setting, the target ATE is -0.138.

Figure 4.1 here visually checks the propensity scores of source and target samples fitted using simple logistic regression including all 5 covariates. The distribution of propensity scores in both samples shows a substantial degree of overlap, indicating that the covariate distributions between the two samples are sufficiently similar. This overlap suggests that the generalization of treatment effects from the source population to the target population is reliable and exact balancing could be achieved.

Here is the summary statistics of the exact balancing weights.

```
## Source sample
wts_gen <- ebal_wts(xs, trts,H_vars, target_moments,
                    H_add_intercept = TRUE, delta)$w
summary(wts_gen)
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>  0.2415  1.0163  1.5952  2.0000  2.5774 14.0762
```

Here is the generalized target ATE using the weights above.

```
ebal_ATE(xs,ys,trts,H_vars, target_moments,
         H_add_intercept=TRUE, delta)$ATE
#>       value
#> 0.02294481
```

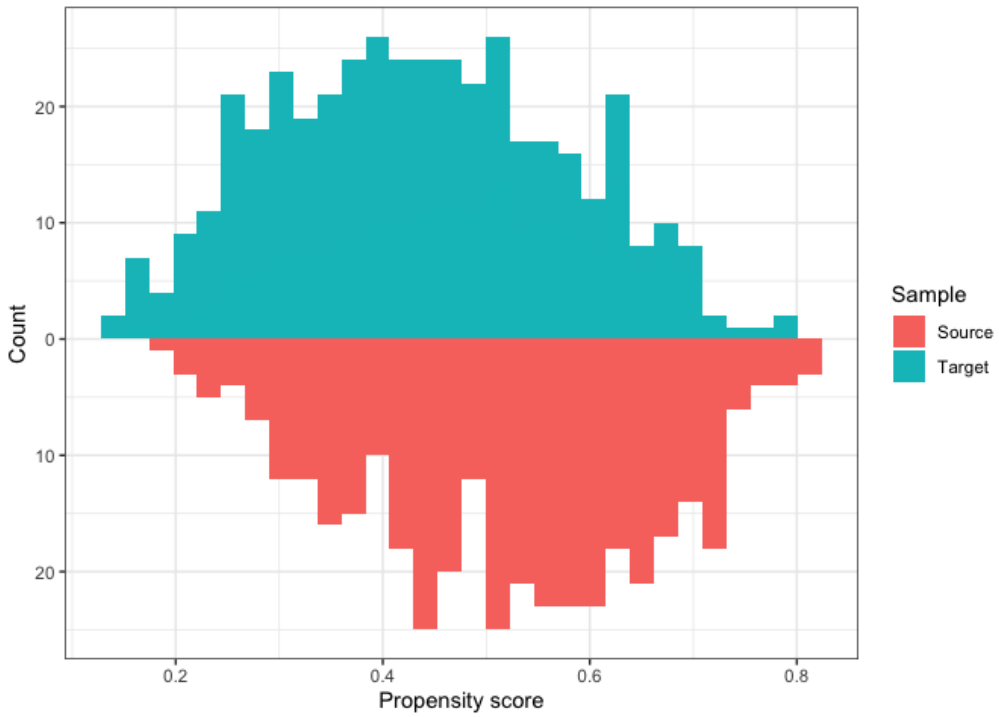

Figure 4.1: Propensity scores distribution between source and target samples (good overlap)

For CI estimation, we use resampling-based perturbation *RPM_CI()* with input of *target_sd* and the number of bootstrap iteration of 300.

```
## CI construction
target_sd = colStdevs(xt)[H_vars]
ATE_CI = RPM_CI(xs, ys,trts,
                H_vars=H_vars,target_mean=target_moments,
                target_sd=target_sd,num_sim=300,
                H_add_intercept=TRUE,
                cluster=5, set_seed=100)
```

```
## Lower bound of 95% CI
ATE_CI$lb_ATE
#>       2.5%
#> -2.130664

## Upper bound of 95% CI
ATE_CI$ub_ATE
#>     97.5%
#> 2.310287
```

For the approximate balancing example, we set the total sample size $n = n_s + n_t = 400$ which is split into source $n_s = 281$ and target $n_t = 119$ samples. We generate 5 covariates $X = (X_1, \ldots, X_5)$ from a uniform distribution $U(-2, 6)$. The remaining settings are identical to those in the previous example. In this setting, the target ATE is -0.641.

Figure 4.2 here visually checks the propensity scores of source and target samples fitted using simple logistic regression including all 5 covariates. The distribution of propensity scores in both samples shows a limited degree of overlap, indicating that the covariate distributions between the two samples are quite different. This overlap suggests that approximate balancing should be used.

Here is the summary statistics of the approximate balancing weights

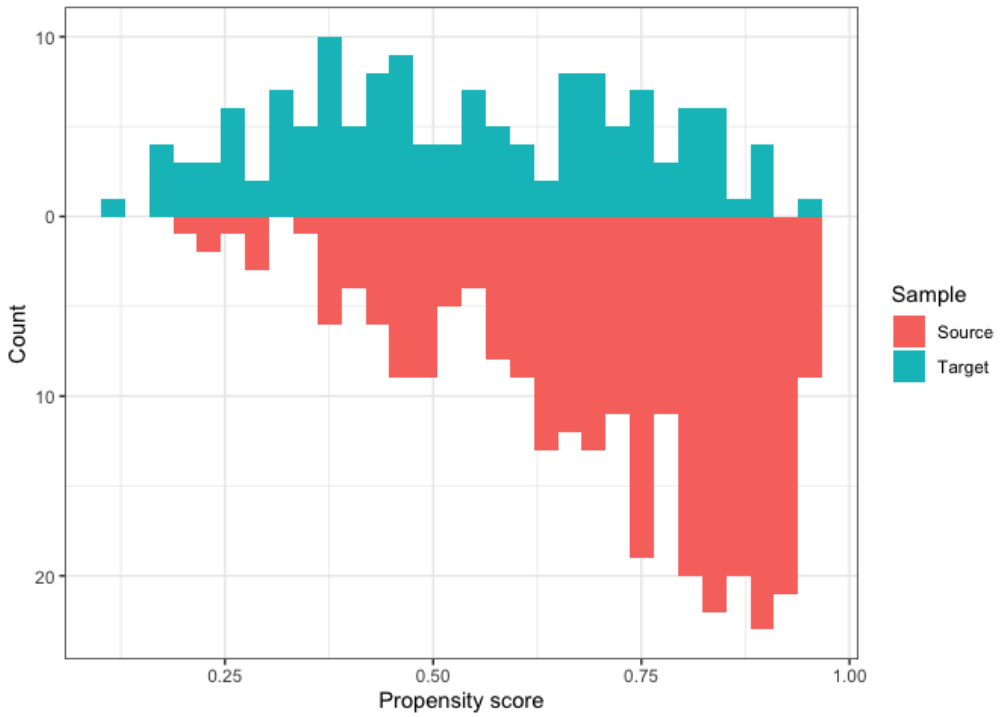

Figure 4.2: Propensity scores distribution between source and target samples (bad overlap)

for the source sample if we set the *delta* to be of 0.1 for all covariates to be balanced.

```
wts_gen <- ebal_wts(xs, trts ,H_vars,
                    target_moments,H_add_intercept = TRUE,
                    delta=numeric(8)+0.1)$w
summary(wts_gen)
#>      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
#>   0.04555   0.33305   0.75514   2.00001   1.85307  93.90129
```

Here is the generalized target ATE using the weights above.

```
ebal_ATE(xs,ys,trts,H_vars, target_moments,
         H_add_intercept=TRUE,
         delta=numeric(8)+0.1)$ATE
#>       value
#> -1.046427
```

For CI estimation, we use resampling-based perturbation *RPM_AB*() with additional input of *target_sd* and the number of bootstrap as 300.

```
target_sd = colStdevs(xt)[H_vars]
ATE_CI = RPM_AB(xs, ys,trts, H_vars=H_vars,
                target_mean=target_moments,
                target_sd=target_sd,num_sim=300,
                H_add_intercept=TRUE,
                cluster=5, set_seed=100)

## Lower bound of 95% CI
ATE_CI$lb_ATE
#>       2.5%
#> -2.807896

## Upper bound of 95% CI
ATE_CI$ub_ATE
#>    97.5%
```

```
#> 4.54158

## Number of simulations that uses exact balancing over 300
ATE_CI$use_exact
#> [1] 67
```

A SUPPLEMENTARY MATERIALS FOR CHAPTER 1 "RANDOMIZED PHASE II DESIGN WITH ORDER CONSTRAINED STRATA"

A.1 Randomized phase II screening design with order constrained strata

Introduction

In the main text, we focused on incorporating ordering information in the context of randomized selection design. In this section of the Appendix, we discuss incorporating ordering information in randomized screening design.

Randomized phase II screening design was introduced by Rubinstein et al. [2005], extending previous research by Simon et al. [2001] and Korn et al. [2001]. The goal of this type of design is to design a randomized study that could yield sample sizes and statistical properties suitable for phase II studies. The design provides preliminary comparisons between experimental and standard treatments by carefully adjusting and balancing the type I (α) and II errors (β), ensuring that the targeted treatment benefit is appropriate while the sample size remains restricted. Here we want to show that, with minor modifications, the advantage of incorporating ordering information can still be observed in randomized selection design.

Method

Here we use a similar setting as in the main text. Assume patients are stratified into G strata and randomized to treatment arm T and control arm C. In total, there are N patients in each arm where the proportion of patients in each stratum is w_{jg} with $j \in \{T, C\}$. Here, $g = 1, ..., G$ and $\sum_{g=1}^{G} w_{jg} = 1$. So the number of patients in Arm j Stratum g is defined as $n_{jg} = N \cdot w_{jg}$.

Binary outcome

Assume that, in Arm j and Stratum g, the number of responders r_{jg} are independent binomial random variables with $r_{jg} \sim \text{Bin}(n_{jg}, \pi_{jg})$. Again, we assume the strata in Arm j satisfy the partial stochastic ordering constraints in its strata defined by a constraint set $E \subset \{1, ..G\}^2$, i.e., $\forall (u, v) \in E, \pi_{ju} \geqslant \pi_{jv}$.

Denote

$$\begin{aligned} \boldsymbol{\pi}_j &= (\pi_{j1}, \ldots, \pi_{jG})^\top, \quad j = T, C; \\ \boldsymbol{p}_j &= (p_{j1}, \ldots, p_{jG})^\top, \quad j = T, C\,. \end{aligned}$$

Let p_{jg} be the corresponding E-constrained maximum likelihood estimator (MLE) with Arm j Stratum g under the constraint set E. Under the framework described by Rubinstein et al. [2005], we design a randomized phase II screening trial that will allow us to assess whether treatment arm T is more promising than standard control arm C. The hypotheses

associated with this type of comparison are

$$H_0 : \boldsymbol{\pi}_T = \boldsymbol{\pi}_C \qquad \text{vs} \qquad H_1 : \boldsymbol{\pi}_T \succ \boldsymbol{\pi}_C$$

Given a specific sample size N, we control the type I error by choosing the critical value $\boldsymbol{\gamma} = (\gamma, \ldots, \gamma)$ to maximize the probability in (A.1) so that the probability of accepting treatment arm is no larger than α under H_0.

$$\max_{\boldsymbol{\gamma}} \left[\Pr(\boldsymbol{p}_T - \boldsymbol{p}_C \succ \boldsymbol{\gamma} \mid \boldsymbol{\pi}_T = \boldsymbol{\pi}_C) \right] \quad < \quad \alpha \tag{A.1}$$

Denote θ^* as the clinically significant difference in response rate with $\boldsymbol{\theta}^* = (\theta^*, \ldots, \theta^*)^\top_{1\times G}$, $\quad \theta^* > 0$. The probability of correct screening, which is the power of the test given the critical value $\boldsymbol{\gamma}$ is determined by (A.2):

$$\Pr(\boldsymbol{p}_T - \boldsymbol{p}_C \succ \boldsymbol{\gamma} \mid \boldsymbol{\pi}_T = \boldsymbol{\pi}_C + \boldsymbol{\theta}^*) = 1 - \beta \tag{A.2}$$

Given the constrained resources for phase II trials, we aim to limit the two-arm trial's sample size while still ensuring the ability to effectively screen the effective treatment. We thus want to keep the type I error rate α to be either 10% or 20% while allowing the power to also be either 90% or 80% evaluated at treatment effect of θ^*.

To determine the critical values $\boldsymbol{\gamma}$ and power, a Monte Carlo simulation based algorithm is proposed given a specific sample size. Under the null, we repeat 10,000 times to obtain a simulated distribution of the estimated

treatment effect for each stratum g. The critical value γ is determined by finding the maximum γ obtaining (A.1). Under the alternative, we repeat 10,000 times to compute the power by the proportion of simulated estimated treatment effect for each stratum g being correctly rejected H_0 based on the critical value γ.

Time-to-event outcome

Suppose $S_{jg}(x), j = T, C$ and $g = 1, \ldots, G$ is the true survival probability at time x for Stratum g in Arm j. Further assume that the strata in Arm j satisfy the partial stochastic ordering constraints at a given time x defined by the constraint set $E \subset \{1, ..G\}^2$, i.e., $\forall (u, v) \in E, S_{ju}(x) \geqslant S_{jv}(x)$.

Let $\tilde{S}_{jg}(x)$ be the corresponding E-constrained nonparametric maximum likelihood estimator (NPMLE) for survival probability of Arm j Stratum g subject to constraint set E applied at a given time x only. Denote

$$\begin{aligned} \boldsymbol{S}_j &= (S_{j1}(x), \ldots, S_{jG}(x))^\top, \quad j = T, C; \\ \tilde{\boldsymbol{S}}_j &= (\tilde{S}_{j1}(x), \ldots, \tilde{S}_{jG}(x))^\top, \quad j = T, C\,. \end{aligned}$$

The hypothesis testing construction is similar as the above binary outcome case. That is, the hypotheses associated with this type of comparison are

$$H_0 : \boldsymbol{S}_T = \boldsymbol{S}_C \qquad \text{vs} \qquad H_1 : \boldsymbol{S}_T \succ \boldsymbol{S}_C$$

Given a specific sample size, we control the type I error α by choosing a critical value γ to maximize the probability in (A.3) so that the probability

of accepting treatment arm is no larger than α under H_0.

$$\max_{\gamma} [\Pr(\tilde{\boldsymbol{S}}_T - \tilde{\boldsymbol{S}}_C \succ \boldsymbol{\gamma} \mid \boldsymbol{S}_T = \boldsymbol{S}_C)] \quad < \quad \alpha \tag{A.3}$$

Denote $\boldsymbol{\theta}^* = (\theta^*, \ldots, \theta^*)^\top_{1\times G}, \quad \theta^* > 0$ as the clinically significant difference in survival probability at time x. The power of the screening design given the critical value $\boldsymbol{\gamma}$ can be determined by (A.4):

$$\Pr(\tilde{\boldsymbol{S}}_T - \tilde{\boldsymbol{S}}_C \succ \boldsymbol{\gamma} \mid \boldsymbol{S}_T = \boldsymbol{S}_C + \boldsymbol{\theta}^*) = 1 - \beta \tag{A.4}$$

Similar as the binary outcome case, a Monte Carlo simulation based algorithm is proposed to determine the critical value $\boldsymbol{\gamma}$ and power given a specific sample size.

Evaluation with simulated setting

We seek to evaluate the performance of our proposed method in comparison to a simple randomized stratified screening design, which does not account for order information. We use simulation studies conducted under similar settings as the selection design in Chapter 1. We consider the setting that each of the Arms T and C has N patients. There are $G = 2$ strata in each arm. The patient proportions in different strata are $w_{T1} = w_{C1} = 0.4$ (therefore $w_{T2} = w_{C2} = 0.6$).

Binary outcome

Assume $\pi_{Tg} \geqslant \pi_{Cg}$ and $\pi_{j2} \geqslant \pi_{j1}, j = T, C, g = 1, 2$ without loss of generality. Accordingly, the E-constraint is $\pi_{j2} \geqslant \pi_{j1}$. The binomial response rate MLE without considering the order information would be the observed response rate $\hat{\pi}_{jg} = n_{jg}{}^{-1} r_{jg}$. Given different sample sizes, we will compare the estimated power between two methods, based on 10,000 simulations.

Figure A.1 presents the power of the screening trial across different values of N controlled at $\alpha = 0.1, 0.2$. Overall, we see our method gives a slightly larger power than the method without using order information. Table A.1 numerically shows power, across different values of N controlled at $\alpha = 0.1$.

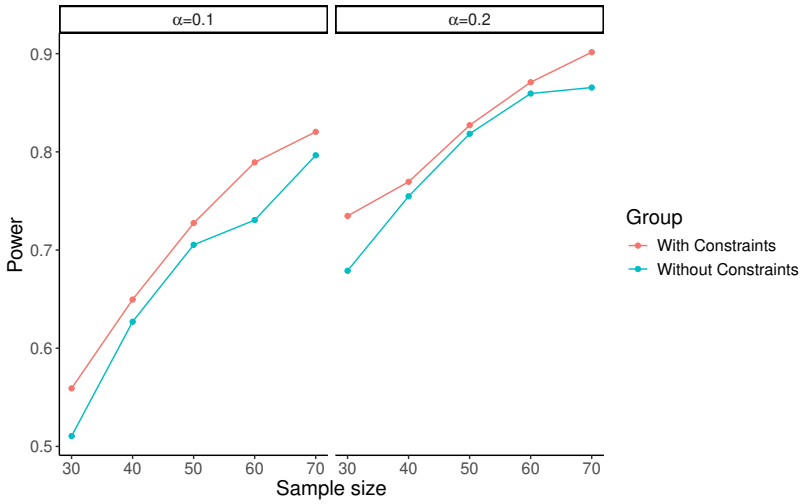

Figure A.1: Power of the screening trials for various N controlled at $\alpha = 0.1, 0.2$, fixing $\boldsymbol{\pi}_C = (0.25, 0.35)$, $\boldsymbol{\theta}^* = (0.2, 0.2)$.

Table A.1: Power of the screening trials for various N controlled at $\alpha = 0.1$, fixing $\boldsymbol{\pi}_C = (0.25, 0.35)$, $\boldsymbol{\theta}^* = (0.2, 0.2)$.

	With constraints		Without constraints	
N	α	$1-\beta$	α	$1-\beta$
30	0.089	0.559	0.074	0.511
40	0.082	0.650	0.094	0.627
50	0.097	0.727	0.099	0.705
60	0.099	0.789	0.080	0.731
70	0.099	0.820	0.096	0.797

Survival outcome

Assume $S_{Tg}(x) \geqslant S_{Cg}(x)$, $S_{j2}(x) \geqslant S_{j1}(x), j = T, C, g = 1, 2$ without loss of generality. Accordingly, $S_{jg}(x)$ satisfy the E-constraints at time x that $S_{j2}(x) \geqslant S_{j1}(x), j = T, C$. The survival probability NPMLE without considering the ordering would be the Kaplan-Meier estimator $\hat{S}_{jg}(x) = \prod_{i:x_i \leqslant x}(1 - \frac{d_{jgi}}{n_{jgi}})$. Given different sample sizes, we will compare the calculated power between the two methods, based on 10,000 simulations.

Suppose patients enroll according to a Poisson process with an accrual rate of 4 patients per month for each of the treatment arm stratum. We continue to follow up for an additional 6 months after the last patient is enrolled. Suppose the survival time follows exponential distribution and we are constraining and comparing survival probabilities at 6 months.

The power of the screening trial across different N controlled at $\alpha = 0.1, 0.2$ are shown in Figure A.2. Overall, we observe similar results as the binary case that our method gives a slightly larger power than the method without using order information. Table A.2 numerically shows

power across different values of N controlled at $\alpha = 0.1$.

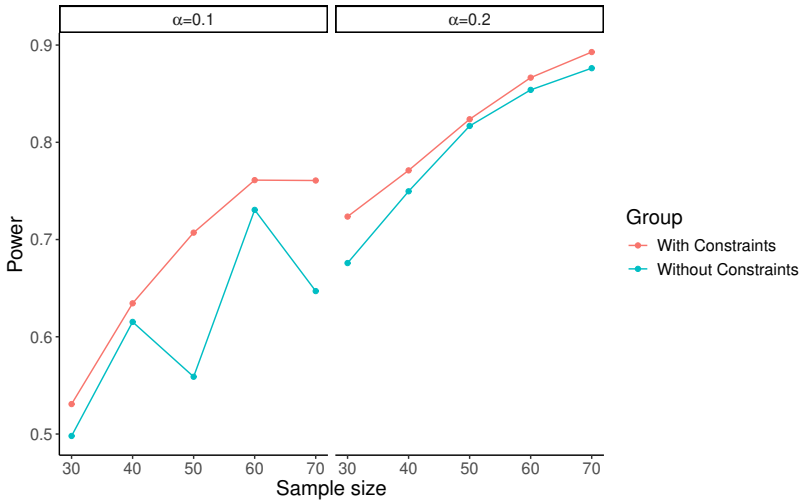

Figure A.2: Power of the screening trials for various N controlled at $\alpha = 0.1, 0.2$, fixing $\boldsymbol{S}_C = (0.35, 0.45)$, $\boldsymbol{\theta}^* = (0.2, 0.2)$.

Table A.2: Power of the screening trials for various N controlled at $\alpha = 0.1$, fixing $\boldsymbol{S}_C = (0.35, 0.45)$, $\boldsymbol{\theta}^* = (0.2, 0.2)$.

	With constraints		Without constraints	
N	α	$1-\beta$	α	$1-\beta$
30	0.080	0.531	0.081	0.498
40	0.090	0.634	0.090	0.615
50	0.099	0.707	0.083	0.559
60	0.099	0.761	0.088	0.731
70	0.070	0.761	0.096	0.647

A.2 Bias and variability of constrained estimators

Here we investigate the bias and variability of the estimators without constraints and with constraints for both binomial and time-to-event outcomes under 8,000 Monte Carlo simulations. In the simulations, we assume $G = 2$ strata in each treatment arm with patient proportions $w_{a1} = w_{b1} = 0.4$ and sample size for each treatment arm N. For binomial responses, the number of responders r_{jg} are independent binomial random variables with $r_{jg} \sim \text{Bin}(n_{jg}, \pi_{jg}), j = a, b, g = 1, 2$. The treatment difference for each stratum is $\boldsymbol{\theta}^* = (0.2, 0.2)$. E-constraint is $\pi_{j2} \geqslant \pi_{j1}, j = a, b$. The binomial response rate MLE without considering order is ${n_{jg}}^{-1} r_{jg}$. The results are in Table A.3.

Similar to a recent publication [Dai et al., 2020], we find that estimators under constraints are biased for the response rate of each stratum, compared with the unbiased binomial MLEs. On the other hand, for the treatment effect which is the difference of the estimators between two arms, the biases become almost negligible. We therefore take comfort in this fact for our proposed method as the treatment effect is of ultimate interest. In addition, we find that the constrained estimators have slightly smaller variance than the estimators without constraints. We also find positive correlations between estimated treatment effects across ordered strata, especially under the small sample size setting.

Table A.3: Estimation mean, variance and correlation for binary outcomes with and without constraints

Parameters	True values	Estimates - Mean		Estimates - Variance		Estimate - Correlation	
		w constr	w/o constr	w constr	w/o constr	w constr	w/o constr
			N = 50				
(π_{a1}, π_{a2})	(0.6, 0.7)	(0.588, 0.707)	(0.599, 0.700)	(0.010, 0.006)	(0.012, 0.007)	0.217	-0.014
(π_{b1}, π_{b2})	(0.4, 0.5)	(0.387, 0.508)	(0.399, 0.500)	(0.009, 0.007)	(0.012 0.008)	0.225	-0.003
(θ_1^*, θ_2^*)	(0.2, 0.2)	(0.201, 0.198)	(0.199,0.199)	(0.019, 0.0.013)	(0.024, 0.015)	0.220	0.006
			N=40				
(π_{a1}, π_{a2})	(0.55, 0.65)	(0.534, 0.659)	(0.548, 0.649)	(0.012, 0.008)	(0.015, 0.010)	0.247	0.023
(π_{b1}, π_{b2})	(0.35, 0.45)	(0.334, 0.460)	(0.349,0.450)	(0.011, 0.009)	(0.014, 0.010)	0.239	0.014
(θ_1^*, θ_2^*)	(0.2, 0.2)	(0.200, 0.199)	(0.199,0.199)	(0.022, 0.017)	(0.029, 0.0.019)	0.242	0.009
			N = 30				
(π_{a1}, π_{a2})	(0.6, 0.7)	(0.580,0.712)	(0.598,0.700)	(0.015, 0.010)	(0.020, 0.012)	0.270	0.016
(π_{b1}, π_{b2})	(0.4, 0.5)	(0.379,0.514)	(0.399,0.500)	(0.014, 0.012)	(0.020, 0.014)	0.286	-0.014
(θ_1^*, θ_2^*)	(0.2, 0.2)	(0.201,0.198)	(0.199,0.200)	(0.029, 0.021)	(0.039, 0.025)	0.275	0.012
			N=20				
(π_{a1}, π_{a2})	(0.55, 0.65)	(0.520,0.668)	(0.547,0.650)	(0.023, 0.015)	(0.031, 0.019)	0.290	0.003
(π_{b1}, π_{b2})	(0.35, 0.45)	(0.321,0.469)	(0.349,0.450)	(0.020, 0.017)	(0.028, 0.021)	0.303	-0.001
(θ_1^*, θ_2^*)	(0.2, 0.2)	(0.199,0.199)	(0.198,0.199)	(0.042, 0.032)	(0.059, 0.039)	0.307	0.010

For time-to-event responses, similar findings could be derived from the simulation results in Table A.4. The simulation setting is as follows. Suppose $S_{jg}(x)$ where $j = a, b$ and $g = 1, \ldots, G$ is the true survival probability at time x for Stratum g in Arm j. The treatment difference $\boldsymbol{\theta}^* = (0.2, 0.2)$. $S_{jg}(x)$ satisfies the E-constraints at month 6 that $S_{j2}(6) \geqslant S_{j1}(6)$. The survival probability NPMLE without considering order is the Kaplan-Meier estimator. Here we assume patients enroll according to a Poisson process with an accrual rate of 4 patients per month for each of the treatment arm stratum. The follow-up is for an additional 6 months after the last patient is enrolled for each stratum. The survival time follows an exponential distribution and we are constraining and comparing survival probabilities

at 6 months.

Table A.4: Estimation mean, variance and correlation for time-to-event outcomes with and without constraints

Parameters	True values	Estimates - Mean		Estimates - Variance		Estimates - Correlation	
		w constr	w/o constr	w constr	w/o constr	w constr	w/o constr
			N=50				
$(S_{a1}(6), S_{a2}(6))$	(0.75, 0.85)	(0.741, 0.854)	(0.748, 0.850)	(0.008 0.004)	(0.010 0.004)	0.166	-0.008
$(S_{b1}(6), S_{b2}(6))$	(0.55, 0.65)	(0.538, 0.656)	(0.550, 0.648)	(0.010 0.007)	(0.012 0.008)	0.205	-0.001
(θ_1^*, θ_2^*)	(0.2, 0.2)	(0.203, 0.198)	(0.198, 0.202)	(0.018, 0.011)	(0.021, 0.012)	0.217	-0.003
			N=40				
$(S_{a1}(6), S_{a2}(6))$	(0.55, 0.65)	(0.534, 0.658)	(0.549, 0.648)	(0.012 0.008)	(0.015, 0.010)	0.231	-0.007
$(S_{b1}(6), S_{b2}(6))$	(0.35, 0.45)	(0.336, 0.460)	(0.351, 0.450)	(0.011 0.009)	(0.014, 0.010)	0.225	-0.006
(θ_1^*, θ_2^*)	(0.2, 0.2)	(0.198,0.198)	(0.198,0.198)	(0.023, 0.0170)	(0.030, 0.020)	0.248	0.025
			N=30				
$(S_{a1}(6), S_{a2}(6))$	(0.75, 0.85)	(0.736, 0.858)	(0.749, 0.849)	(0.013, 0.006)	(0.016, 0.007)	0.138	0.019
$(S_{b1}(6), S_{b2}(6))$	(0.55, 0.65)	(0.531, 0.663)	(0.552, 0.649)	(0.016, 0.010)	(0.021, 0.013)	0.266	0.003
(θ_1^*, θ_2^*)	(0.2, 0.2)	(0.205, 0.195)	(0.198, 0.200)	(0.028, 0.017)	(0.036, 0.020)	0.249	-0.003
			N=20				
$(S_{a1}(6), S_{a2}(6))$	(0.55, 0.65)	(0.522,0.668)	(0.550, 0.649)	(0.023, 0.015)	(0.031, 0.019)	0.302	0.023
$(S_{b1}(6), S_{b2}(6))$	(0.35, 0.45)	(0.321,0.469)	(0.349, 0.450)	(0.020, 0.017)	(0.028, 0.021)	0.316	-0.019
(θ_1^*, θ_2^*)	(0.2, 0.2)	(0.200,0.199)	(0.201, 0.198)	(0.042, 0.033)	(0.057, 0.041)	0.315	0.016

A.3 Checking the monotonicity assumption

Tibshirani et al. [2011] investigated the problem of nearly isotonic regression where the order constraint might be violated at some change-points. Specifically, for N normal observations $x_i \sim \mathcal{N}(\mu_i, \sigma^2)$ for $i = 1, ..., N$, the

problem could be formulated as the regularized optimization as following:

$$\hat{\mu}_\lambda = \underset{\mu}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{N} (x_i - \mu_i)^2 + \lambda \sum_{i=1}^{N-1} (\mu_i - \mu_{i+1})_+ ,$$

where $\lambda > 0$ is the regularization parameter and $(\cdot)_+ = \max(\cdot, 0)$.

Matsuda and Miyatake [2022] extended the nearly isotonic regression for general one-parameter exponential families for binomial responses. For stratum i with responses $r_i \sim \mathrm{Bin}(n_i, \pi_i)$ for $i = 1, ..., N$, the regularized optimization could be formulated as

$$\begin{aligned} \hat{c}_\lambda &= \underset{c}{\operatorname{argmin}} - \sum_{i=1}^{N} \log p(r_i|c_i) + \lambda \sum_{i=1}^{N-1} (c_i - c_{i+1})_+ \\ &= \underset{c}{\operatorname{argmin}} \sum_{i=1}^{N} w_i(-c_i \frac{r_i}{w_i} + b(c_i)) + \lambda \sum_{i=1}^{N-1} (c_i - c_{i+1})_+ , \end{aligned} \tag{A.5}$$

where $c_i = \log \frac{\pi_i}{1-\pi_i}$, $b_i(c_i) = w_i b(c_i) = -n_i \log(1 - \pi_i) = n_i \log(1 + \exp(c_i))$, $w_i = n_i$. The paper shows that this optimization problem is efficiently solved by modified Pool Adjacent Violators Algorithm (mPAVA) [Matsuda and Miyatake, 2022].

The selection of regularization parameter λ is based on Akaike information criterion (AIC) defined as

$$\mathrm{AIC}(\lambda) = -2 \sum_{i=1}^{n} \log p(x_i|c_\lambda)_i + 2K_\lambda . \tag{A.6}$$

where K_λ is the number of joined pieces used as an unbiased estimate of the degrees of freedom of nearly isotonic regression.

However, there is no existing work considering nearly isotonic regres-

sion modeling for time-to-event outcomes. Therefore we performed a simulation to examine the lack of fit for the monotonicity assumption in binomial responses using the algorithm from Matsuda and Miyatake [2022]. The simulation is set up as following. First, suppose $G = 5$ strata with monotonically increasing true response rates $\pi_g = (0.2, 0.3, 0.4, 0.5, 0.6)$ and sample size $n_g = (20, 20, 30, 30, 40)$ with response $r_g \sim \text{Bin}(n_g, \pi_g)$ for $g = 1, ..., 5$. We calculate $\text{AIC}(\lambda)$ under each value of $\lambda \in \{0, 0.1, 0.2, ..., 5\}$ and each with 3,000 Monte Carlo simulations. Then, we changed the response rates to be non-monotonic $\pi_g = (0.2, 0.3, 0.4, 0.32, 0.24)$ with the same sample sizes and calculate $\text{AIC}(\lambda)$ again. Figure A.3 plots $E_c\{\text{AIC}(\lambda)\}$ with 95% confidence interval for each value of λ. As expected, when the monotonicity assumption holds, AIC level decreases to gradual stabilization whereas under monotonicity assumption violation, AIC first decreases and then increases. This can serve as a visual tool to assess the monotonicity assumption.

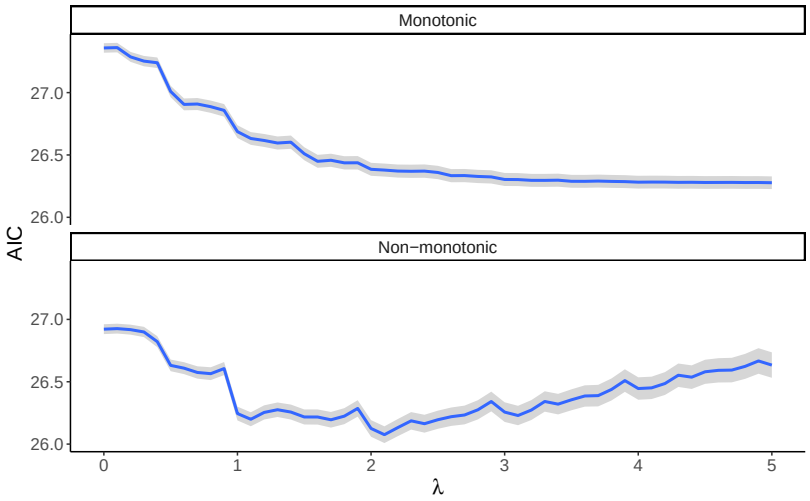

Figure A.3: $E_c\{AIC(\lambda)\}$ with 95% confidence interval for monotonic and non-monotonic binomial responses

A.4 Violation of the monotonicity assumption

Here are the details for the simulation setting that examines the violation of monotonicity constraints. The general strata, treatment, and accrual settings are the same as Section A.2 above. For both binomial and time-to-event responses, with the constraint ordering being $\pi_{j2} \geqslant \pi_{j1}$ and $S_{j2}(6) \geqslant S_{j1}(6), j = a, b$, we set the strata response rate difference $\tau = \pi_{j2} - \pi_{j1} = S_{j2}(6) - S_{j1}(6)$ to be either 0.1 or -0.1 for $j = a, b$. For $\tau = 0.1$, the monotonicity assumption holds, whereas for $\tau = -0.1$, it is violated.

We compare the estimated P_{corr} and P_{amb} between designs with and without order constraints under either concordant or discordant treatment effect $\boldsymbol{\theta}^*$. In particular, for concordant $\boldsymbol{\theta}^*$, we use $(0.2, 0.2)$ or $(0.1, 0.2)$. For discordant $\boldsymbol{\theta}^*$, we use $(-0.1, 0.2), (-0.2, 0.2), (0.2, -0.1)$, and $(0.2, -0.2)$.

Under the concordant $\boldsymbol{\theta}^*$, Arm a should be recommended and therefore we would like to see a larger P_{corr} and smaller P_{amb}. On the other hand, for discordant $\boldsymbol{\theta}^*$, we would like to see a smaller P_{corr} and larger P_{amb}.

From Tables A.5 and A.6, it can be seen that, with concordant treatment effects, P_{corr} could increase when order constraints are applied to the ordering violation case, which means that the probability of correct or definitive recommendation for the superior treatment could be gained by ignoring the order constraints among various strata.

On the other hand, with discordant treatment effects, the violation resulted in a larger P_{corr} and a smaller P_{amb}, which is worse than without constraints. We have added this findings to our discussion section in the main text.

Table A.5: Evaluation of binary response strata order violation constraints between with and without constraints with $\pi_{b1} = 0.35$, $\boldsymbol{\theta} = (0.05, 0.05)$ and $N = 30$.

		With constraints		Without constraints	
τ	$\boldsymbol{\theta}^*$	P_{corr}	P_{amb}	P_{corr}	P_{amb}
0.1	(0.2,0.2)	0.726	0.263	0.674	0.317
	(0.1,0.2)	0.570	0.408	0.528	0.453
	(-0.1,0.2)	0.202	0.745	0.190	0.759
	(-0.2,0.2)	0.070	0.861	0.067	0.865
	(0.2,-0.1)	0.281	0.556	0.170	0.754
	(0.2,-0.2)	0.147	0.575	0.057	0.845
-0.1	(0.2,0.2)	0.832	0.151	0.683	0.309
	(0.1,0.2)	0.694	0.276	0.535	0.448
	(-0.1,0.2)	0.275	0.658	0.192	0.763
	(-0.2,0.2)	0.095	0.821	0.068	0.872
	(0.2,-0.1)	0.363	0.408	0.130	0.790
	(0.2,-0.2)	0.173	0.421	0.015	0.879

Table A.6: Evaluation of time-to-event response strata order violation constraints between with and without constraints with $S_{b1}(6) = 0.55$, $\boldsymbol{\theta} = (0.05, 0.05)$ and $N = 30$.

		With constraints		Without constraints	
τ	$\boldsymbol{\theta}^*$	P_{corr}	P_{amb}	P_{corr}	P_{amb}
0.1	(0.2,0.2)	0.757	0.236	0.707	0.288
	(0.1,0.2)	0.588	0.396	0.543	0.442
	(-0.1,0.2)	0.224	0.739	0.210	0.756
	(-0.2,0.2)	0.107	0.847	0.097	0.858
	(0.2,-0.1)	0.278	0.554	0.170	0.759
	(0.2,-0.2)	0.167	0.551	0.066	0.843
-0.1	(0.2,0.2)	0.832	0.147	0.687	0.303
	(0.1,0.2)	0.692	0.273	0.529	0.450
	(-0.1,0.2)	0.298	0.626	0.204	0.744
	(-0.2,0.2)	0.139	0.771	0.094	0.840
	(0.2,-0.1)	0.410	0.356	0.170	0.760
	(0.2,-0.2)	0.302	0.333	0.056	0.852

BIBLIOGRAPHY

Rui Chen, Guanhua Chen, and Menggang Yu. Entropy balancing for causal generalization with target sample summary information. *Biometrics*, 79 (4):3179–3190, 2023. doi: https://doi.org/10.1111/biom.13825. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13825.

L. Rubinstein, J. Crowley, P. Ivy, M. Leblanc, and D. Sargent. Randomized phase II designs. *Clin Cancer Res*, 15(6):1883–1890, Mar 2009.

Manish R. Sharma, Walter M. Stadler, and Mark J. Ratain. Randomized Phase II Trials: A Long-term Investment With Promising Returns. *JNCI: Journal of the National Cancer Institute*, 103(14):1093–1100, 06 2011. ISSN 0027-8874. doi: 10.1093/jnci/djr218. URL https://doi.org/10.1093/jnci/djr218.

Peter F. Thall, J. Kyle Wathen, B. Nebiyou Bekele, Richard E. Champlin, Laurence H. Baker, and Robert S. Benjamin. Hierarchical bayesian approaches to phase II trials in diseases with multiple subtypes. *Statistics in Medicine*, 22(5):763–780, 2003. doi: 10.1002/sim.1399.

Wendy B. London and Myron N. Chang. One- and two-stage designs for stratified phase II clinical trials. *Statistics in Medicine*, 24(17):2597–2611, 2018/06/14 2005. doi: 10.1002/sim.2139.

J Kyle Wathen, Peter F Thall, John D Cook, and Elihu H Estey. Accounting

for patient heterogeneity in phase II clinical trials. *Statistics in Medicine*, 27(15):2802–2815, 07 2008. doi: 10.1002/sim.3109.

Sin-Ho Jung, Myron N. Chang, and Sun J. Kang. Phase II cancer clinical trials with heterogeneous patient populations. *Journal of Biopharmaceutical Statistics*, 22(2):312–328, 2012. doi: 10.1080/10543406.2010.536873. PMID: 22251176.

Myron N Chang, Jonathan J Shuster, and Wei Hou. Improved two-stage tests for stratified phase II cancer clinical trials. *Statistics in Medicine*, 31 (16):1688–1698, 07 2012. doi: 10.1002/sim.5314.

Myron Chang, Sin-Ho Jung, and Samuel S. Wu. Two-stage designs with additional futility tests for phase II clinical trials with heterogeneous patient populations. *Sequential Analysis*, 30(3):338–349, 2011. doi: 10.1080/07474946.2011.593924.

Richard Sposto and Paul S. Gaynon. An adjustment for patient heterogeneity in the design of two-stage phase II trials. *Statistics in Medicine*, 28(20):2566–2579, 2009. doi: 10.1002/sim.3624.

Menghao Xu, Ting Ye, Jun jun Zhao, and Menggang Yu. Sample size determination for stratified phase II cancer trials with monotone order constraints. *Statistics in Biopharmaceutical Research*, 13(4):425–434, June 2020. doi: 10.1080/19466315.2020.1764863. URL https://doi.org/10.1080/19466315.2020.1764863.

P. Y. Liu, Steve Dahlberg, and John Crowley. Selection designs for pilot studies based on survival. *Biometrics*, 49(2):391–398, 1993. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2532552.

D. J. Sargent and R. M. Goldberg. A flexible design for multiple armed screening trials. *Stat Med*, 20(7):1051–1060, Apr 2001.

R. Simon, R. E. Wittes, and S. S. Ellenberg. Randomized phase II clinical trials. *Cancer Treat Rep*, 69(12):1375–1381, Dec 1985.

Y. Park, J. M. Taylor, and J. D. Kalbfleisch. Pointwise nonparametric maximum likelihood estimator of stochastically ordered survivor functions. *Biometrika*, 99(2):327–343, Jun 2012a.

R. E. Barlow and H. D. Brunk. The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337):140–147, 1972. ISSN 01621459. URL http://www.jstor.org/stable/2284712.

Lawrence V. Rubinstein, Edward L. Korn, Boris Freidlin, Sally Hunsberger, S. Percy Ivy, and Malcolm A. Smith. Design issues of randomized phase II trials and a proposal for phase II screening trials. *Journal of Clinical Oncology*, 23(28):7199–7206, 2005. doi: 10.1200/JCO.2005.01.149. URL https://doi.org/10.1200/JCO.2005.01.149. PMID: 16192604.

Hajime Uno, Janet Wittes, Haoda Fu, Scott D. Solomon, Brian Claggett, Lu Tian, Tianxi Cai, Marc A. Pfeffer, Scott R. Evans, and Lee-Jen Wei. Alternatives to hazard ratios for comparing the efficacy or safety of

therapies in noninferiority studies. *Annals of Internal Medicine*, 163(2): 127–134, July 2015. doi: 10.7326/m14-1741. URL https://doi.org/10.7326/m14-1741.

Hajime Uno, Brian Claggett, Lu Tian, Eisuke Inoue, Paul Gallo, Toshio Miyata, Deborah Schrag, Masahiro Takeuchi, Yoshiaki Uyama, Lihui Zhao, Hicham Skali, Scott Solomon, Susanna Jacobus, Michael Hughes, Milton Packer, and Lee-Jen Wei. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology*, 32(22):2380–2385, 2014. doi: 10.1200/JCO.2014.55.2208. PMID: 24982461.

Yongseok Park, John D. Kalbfleisch, and Jeremy M. G. Taylor. Constrained nonparametric maximum likelihood estimation of stochastically ordered survivor functions. *Canadian Journal of Statistics*, 40(1):22–39, 2012b. doi: https://doi.org/10.1002/cjs.10143.

Sin-Ho Jung and Stephen L. George. Between-arm comparisons in randomized phase II trials. *Journal of Biopharmaceutical Statistics*, 19(3): 456–468, 2009. doi: 10.1080/10543400902802391. URL https://doi.org/10.1080/10543400902802391. PMID: 19384688.

Ryan J. Tibshirani, Holger Hoefling, and Robert Tibshirani. Nearly-isotonic regression. *Technometrics*, 53(1):54–61, 2011. doi: 10.1198/TECH.2010.10111. URL https://doi.org/10.1198/TECH.2010.10111.

Takeru Matsuda and Yuto Miyatake. Generalized nearly isotonic regression. *arXiv preprint arXiv:2108.13010*, 2022.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL https://www.R-project.org/.

Microsoft and Steve Weston. *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*, 2022a. URL https://CRAN.R-project.org/package=doParallel. R package version 1.0.17.

Microsoft and Steve Weston. *foreach: Provides Foreach Looping Construct*, 2022b. URL https://CRAN.R-project.org/package=foreach. R package version 1.5.2.

Renaud Gaujoux. *doRNG: Generic Reproducible Parallel Backend for 'foreach' Loops*, 2023. URL https://CRAN.R-project.org/package=doRNG. R package version 1.8.6.

Berwin A. Turlach and Andreas Weihs. *quadprog: Functions to Solve Quadratic Programming Problems*, 2019. URL https://CRAN.R-project.org/package=quadprog. R package version 1.5-8.

Terry M. Therneau. *survival: Survival Analysis*, 2023. URL https://CRAN.R-project.org/package=survival. R package version 3.5-7.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2023. URL https://CRAN.R-project.org/package=dplyr. R package version 1.1.4.

Irina Degtiar and Sherri Rose. A review of generalizability and transportability. *Annual Review of Statistics and Its Application*, 10(1):501–524, March 2023. ISSN 2326-831X. doi: 10.1146/annurev-statistics-042522-103837. URL http://dx.doi.org/10.1146/annurev-statistics-042522-103837.

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review, 2023.

Stephen R. Cole and Elizabeth A. Stuart. Generalizing Evidence From Randomized Clinical Trials to Target Populations: The ACTG 320 Trial. *American Journal of Epidemiology*, 172(1):107–115, 06 2010. doi: 10.1093/aje/kwq084. URL https://doi.org/10.1093/aje/kwq084.

Colm O'Muircheartaigh and Larry V. Hedges. Generalizing from Unrepresentative Experiments: A Stratified Propensity Score Approach. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 63(2):195–210, 11 2013. ISSN 0035-9254. doi: 10.1111/rssc.12037. URL https://doi.org/10.1111/rssc.12037.

Kara E. Rudolph and Mark J. van der Laan. Robust estimation of encouragement-design intervention effects transported across sites. *J. R. Stat. Soc. Series B Stat. Methodol.*, 79(5):1509–1525, November 2017.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 540–547, 2011. doi: 10.1109/ICDMW.2011.169.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proc. Natl. Acad. Sci. U. S. A.*, 113(27):7345–7352, July 2016.

Ellen Graham, Marco Carone, and Andrea Rotnitzky. Towards a unified theory for semiparametric data fusion with individual-level data, 2025. URL https://arxiv.org/abs/2409.09973.

Sijia Li and Alex Luedtke. Efficient estimation under data fusion. *Biometrika*, 110(4):1041–1054, 02 2023. ISSN 1464-3510. doi: 10.1093/biomet/asad007. URL https://doi.org/10.1093/biomet/asad007.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.*, 8:985–1005, dec 2007. ISSN 1532-4435.

Issa J. Dahabreh, James M. Robins, Sebastien J.-P. A. Haneuse, Iman Saeed, Sarah E. Robertson, Elizabeth A. Stuart, and Miguel A. Hernán. Sensitivity analysis using bias functions for studies extending inferences

from a randomized trial to a target population. *Statistics in Medicine*, 42(13):2029–2043, 2023. doi: https://doi.org/10.1002/sim.9550. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9550.

Lin Dong, Shu Yang, Xiaofei Wang, Donglin Zeng, and Jianwen Cai. Integrative analysis of randomized clinical trials with real world evidence studies. *arXiv preprint arXiv:2003.01242*, 2020.

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

Qingyuan Zhao and Daniel Percival. Entropy balancing is doubly robust. *Journal of Causal Inference*, 5(1):20160010, 2016.

Kevin P Josey, Fan Yang, Debashis Ghosh, and Sridharan Raghavan. A calibration approach to transportability with observational data. *arXiv preprint arXiv:2008.06615*, 2020.

Ambarish Chattopadhyay, Eric R. Cohn, and José R. Zubizarreta. One-step weighting to generalize and transport treatment effect estimates to a target population. *The American Statistician*, 78(3):280–289, 2024. doi: 10.1080/00031305.2023.2267598. URL https://doi.org/10.1080/00031305.2023.2267598.

Donald B. Rubin. Estimating causal effects of treatments in randomized

and nonrandomized studies. *Journal of Educational Psychology*, 66(5): 688–701, 1974.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 04 1983. ISSN 0006-3444. doi: 10.1093/biomet/70.1.41. URL https://doi.org/10.1093/biomet/70.1.41.

Issa J. Dahabreh, Sarah E. Robertson, Jon A. Steingrimsson, Elizabeth A. Stuart, and Miguel A. Hernán. Extending inferences from a randomized trial to a new target population. *Statistics in Medicine*, 39(14):1999–2014, April 2020. ISSN 1097-0258. doi: 10.1002/sim.8426. URL http://dx.doi.org/10.1002/sim.8426.

M. I. Parzen, L. J. Wei, and Z. Ying. A resampling method based on pivotal estimating functions. *Biometrika*, 81(2):341–350, 06 1994. ISSN 0006-3444. doi: 10.1093/biomet/81.2.341. URL https://doi.org/10.1093/biomet/81.2.341.

Feifang Hu and John D. Kalbfleisch. The estimating function bootstrap. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 28(3): 449–481, 2000. ISSN 03195724. URL http://www.jstor.org/stable/3315958.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of*

Statistics, 7(1):1 – 26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Bradley Efron. Bayesian inference and the parametric bootstrap. *The Annals of Applied Statistics*, 6(4):1971 – 1997, 2012. doi: 10.1214/12-AOAS571. URL https://doi.org/10.1214/12-AOAS571.

Yixin Wang and Jose R Zubizarreta. Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. *Biometrika*, 107(1):93–105, 10 2019. ISSN 0006-3444. doi: 10.1093/biomet/asz050. URL https://doi.org/10.1093/biomet/asz050.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. doi: 10.1198/016214506000000735.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-Wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Sci. Data*, 3(1):160035, May 2016.

Daniel J. Stekhoven and Peter Bühlmann. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 10 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr597. URL https://doi.org/10.1093/bioinformatics/btr597.

Anqi Fu, Balasubramanian Narasimhan, and Stephen Boyd. CVXR: An R package for disciplined convex ptimization. *Journal of Statistical Software*, 94(14):1–34, 2020. doi: 10.18637/jss.v094.i14.

MOSEK-ApS. Mosek optimization suite, 2024. URL https://www.mosek.com/. Version 10.0.

Tim Hesterberg. *resample: Resampling Functions*, 2022. URL https://CRAN.R-project.org/package=resample. R package version 0.6.

Paul E. Johnson. *rockchalk: Regression Estimation and Presentation*, 2022. URL https://CRAN.R-project.org/package=rockchalk. R package version 1.8.157.

R M Simon, S M Steinberg, M Hamilton, A Hildesheim, S Khleif, L W Kwak, C L Mackall, J Schlom, S L Topalian, and J A Berzofsky. Clinical trial designs for the early clinical development of therapeutic cancer vaccines. *J. Clin. Oncol.*, 19(6):1848–1854, March 2001.

E L Korn, S G Arbuck, J M Pluda, R Simon, R S Kaplan, and M C Christian. Clinical trial designs for cytostatic agents: are new approaches needed? *J. Clin. Oncol.*, 19(1):265–272, January 2001.

Ran Dai, Hyebin Song, Rina Foygel Barber, and Garvesh Raskutti. The bias of isotonic regression. *Electronic Journal of Statistics*, 14(1):801 – 834, 2020. doi: 10.1214/20-EJS1677. URL https://doi.org/10.1214/20-EJS1677.