Chapter 1 © Copyright 2025
John Wiley and Sons
Chapter 2 © Copyright 2025
Oxford University Press
All other materials © Copyright 2025
Marlena S. Bannick

Robust and flexible statistical methods for clinical trials with innovative and practical designs

Marlena S. Bannick

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

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Ting Ye, Chair

Thomas Fleming

Patrick Heagerty

Program Authorized to Offer Degree:

Department of Biostatistics

University of Washington

Abstract

Robust and flexible statistical methods for clinical trials
with innovative and practical designs

Marlena S. Bannick

Chair of the Supervisory Committee:
Ting Ye
Department of Biostatistics

Randomized clinical trials (RCTs) are widely viewed as the gold standard study design for determining the efficacy of health interventions. Given that RCTs are time consuming and resource intensive, efficiency is of paramount concern. However, it is crucial to ensure that in the design, conduct, and analysis of clinical trials, expediency does not come at the expense of statistical rigor. In this dissertation, we take a careful look at trial designs that aim to be efficient and practical. The main through line for all of our projects is to propose modifications to the design, conduct, and/or analysis of these trials so that they can still reliably answer their intended questions. We are motivated by real problems encountered in real trials with commonly-used designs. We prioritize solutions to these problems that are easy to implement and understand. We justify our proposals with comprehensive asymptotic theory and practical simulation studies.

TABLE OF CONTENTS

Page

List of Figures iii

Glossary v

Chapter 1: Inconsistent Use of Covariate Adjustment in Group Sequential Trials . 1
- 1.1 Introduction 2
- 1.2 Preliminaries 3
- 1.3 Proposed Methods for the Hybrid Design 8
- 1.4 Simulation Study 16
- 1.5 Conclusions 20

Chapter 2: A General Form of Covariate Adjustment in Randomized Clinical Trials 24
- 2.1 Introduction 25
- 2.2 Setup, Estimators, and Assumptions 27
- 2.3 Asymptotic Results 30
- 2.4 Joint Calibration 35
- 2.5 Robust Variance Estimation 37
- 2.6 Simulations 38
- 2.7 Application to Clinical Trial Data 41
- 2.8 Conclusion and Recommendation 43

Chapter 3: Estimands and Estimators of Treatment Effect in Master Protocol Trials with Time-to-Event Outcomes 45
- 3.1 Introduction 45
- 3.2 Estimand 49
- 3.3 Sufficient Adjustment Sets 54
- 3.4 Estimation 62

3.5 Simulations 71
3.6 Discussion 75

Bibliography 79

Appendix A: Supplemental Material for Chapter 1 90
A.1 Review of Covariate Adjustment in RCTs 91
A.2 The Hybrid Design 100
A.3 Supplemental Simulation Results 119

Appendix B: Supplemental Material for Chapter 2 123
B.1 Illustration of G-Computation and Prediction Bias 124
B.2 Asymptotic Linearity (AL) of AIPW Estimators under CAR 124
B.3 Asymptotic Normality (AN) of AIPW Estimators under CARs with Condition (B) 150
B.4 Conditions for Guaranteed Efficiency Gain 158
B.5 Joint Calibration 164
B.6 Augmenting Estimating Equations under CAR 179
B.7 Full Simulation Results 190
B.8 Considerations for Finite Sample Performance of Variance Estimators 192

Appendix C: Supplemental Material for Chapter 3 194
C.1 Preliminaries 195
C.2 Asymptotic Theory 202

LIST OF FIGURES

Figure Number Page

1.1 Diagram of group sequential trial with three stages. Figure (a) shows the standard group sequential trial design with the same analysis method to obtain Z_k^* at each stage $k = 1, ..., 3$. Figure (b) shows the modified group sequential trial design where covariate adjustment is only performed after crossing an interim monitoring boundary, or the end of the trial is reached. 5

1.2 Uniform-inflated Pocock and OBF bounds in a 4-stage trial with equally-sized stages. 15

1.3 End-inflated Pocock and OBF bounds in a 4-stage trial with equally-sized stages. 15

1.4 Type I error under scenarios (A)(i)-(ii) and (B)(i)-(iii) across 25,000 simulated trials. 19

1.5 Power under scenarios (A)(i)-(ii) and (B)(i)-(iii) with an effect size of 0.1 across 25,000 simulated trials. 19

3.1 Directed acyclic graph that illustrates the possible relationships between variables in a parallel group trial design (a), a platform trial design (b), and a platform trial design with a sub-study structure (c). 55

3.2 Four possible study designs (a) - (d) for a simple trial where two treatments are compared against one placebo group. Treatment 2 is stopped for futility one year after enrollment is complete. The DAGs to the left and right show the relationships between the variables for each specific design, as in Figure 3.1. Dashed lines indicate possible relationships, which are discussed in the text, and in the specific Settings 1 through 4 in Table 3.1. 57

3.3 Structure of the platform trial for simulations. At time 0, the placebo arm and treatment arm 1 open for enrollment. At time 1, treatment arm 2 open for enrollment. At time 2, treatment arm 1 ceases enrollment, and at time 3, treatment arm 2 ceases enrollment. The trial arms are all administratively censored at time 4 so that all participants have at least 1 year of follow-up, and individuals are censored at 3 years of follow-up. 73

3.4 Simulation results for Cases 1 and 2. 75

A.1 Type I error under scenario (B)(ii) with misspecified ρ at the design stage across 25,000 simulated trials. Stars indicate where the anticipated reduction and the true reduction in variance from using ANCOVA match up (correct specification), which represents the orange line in Figure 4. The dark purple line at the top of each subplot represents anticipating that there is no reduction in variance from using ANCOVA, equivalent to not modifying the boundaries, which represents the green line in Figure 4. 119

B.1 Treatment effect estimates for 1000 simulated two-arm trials with $n = 500$ based on a Poisson data generating process, under simple randomization with $\pi_1 = 1/3$. The response means conditional on $X = (X_c, X_b)^T$ are $E(Y_1 \mid X) = \log(1+X_c+6X_c^3+X_b)$ and $EY(Y_2 \mid X) = X_c+10X_c^2+X_b$, where X_c is uniformly distributed on the interval (-5, 5), X_b is binary with $P(X_b = 1) = 0.5$, and X_c and X_b are independent. The working model is negative binomial in covariate X including treatment-by-covariate interactions, with an unknown dispersion parameter. 125

GLOSSARY

AIPW: Augmented Inverse Probability Weighting

ANOVA: Analysis of Variance

ANCOVA: Analysis of Covariance

ASA: American Statistical Association

CAR: Covariate-Adaptive Randomization

CF: Cystic Fibrosis

ECE: Entire Concurrent and Eligible

FDA: U.S. Food and Drug Administration

GLM: Generalized Linear Model

JC: Joint Calibration

LC: Linear Calibration

NDA: New Drug Application

OBF: O'Brien-Fleming

RCT: Randomized Clinical Trial

SQNSCLC: Squamous Non-Small Cell Lung Cancer

ACKNOWLEDGMENTS

I would like to thank my wonderful advisors for their guidance on these projects. Thank you to Dr. Noah Simon for his supervision on Chapter 1, and thank you to Dr. Ting Ye for her supervision on Chapters 2 and 3. Their mentorship has been instrumental to my development as a biostatistician. I would like to thank the additional members of my reading committee, Drs. Thomas Fleming and Patrick Heagerty, for being excellent mentors, and for their collaboration and helpful insights, particularly during my general and final exams. Thank you also to Dr. Fleming for involving me in his Design of Medical Studies course, during which I learned so much about the design, conduct, and operation of clinical trials.

I would like to thank many other mentors in the University of Washington Biostatistics Department that I have had throughout my journey to earning this degree, starting with my first interaction with the department. I was first exposed to biostatistics by Dr. James Hughes. He gave me my first research projects as an undergraduate, inspired in me a passion for biostatistics, and encouraged me to pursue this degree. I will always be grateful to him for shaping my academic journey and career. Thank you to Dr. Fei Gao at Fred Hutchinson Cancer Center, who was my research advisor and provided my funding for the duration of this degree. Thank you to others in the department who have inspired me and advocated for me, including Drs. Holly Janes, Ruth Etzioni, Barbra Richardson, and Lurdes Inoue. Thank you also to all of the amazing department staff, who are endlessly helpful.

Lastly, I would like to thank my family and friends for supporting me on this journey. Thank you to my parents, my aunt, my in-laws, and all of the rest of my family. Most of all, thank you to my husband, Jonathan Bannick, who has tirelessly supported me and encouraged me to pursue this degree, especially as we became parents to our daughter, Clara.

DEDICATION

This dissertation is dedicated to Clara Nicolina Bannick.

Chapter 1

INCONSISTENT USE OF COVARIATE ADJUSTMENT IN GROUP SEQUENTIAL TRIALS

This chapter is based on the following article, by permission of John Wiley and Sons:

Marlena S Bannick, Sonya L. Heltshe, Noah Simon (2025). Accounting for Inconsistent Use of Covariate Adjustment in Group Sequential Trials. Statistics in Medicine 44 (10-12) e70082. DOI:10.1002/sim.70082

The version that is included here has been slightly modified so that the notation style matches the style of the other two chapters.

1.1 Introduction

Randomized clinical trials (RCTs) are often viewed as the gold standard study design for determining the efficacy of health interventions. It is important that RCTs make efficient use of resources. To that end, two techniques in the design and analysis of RCTs have become popular: group sequential designs, and covariate-adjusted analyses.

The group sequential trial design allows for early termination of a trial based on strong signals of efficacy or futility. Since investigators perform multiple analyses on the same data over time, standard group sequential designs set higher thresholds to reject the null hypothesis at each interim analysis. This carefully controls the type I error across the whole duration of the trial. Covariate adjustment refers to using pre-randomization variables in an analysis of the treatment effect between groups. When performed correctly, covariate adjustment can decrease the variance of the estimated treatment effect without changing the estimand. This increases estimate precision and power to reject the null hypothesis. Both group sequential designs and covariate adjustment, and their use simultaneously, have been studied extensively elsewhere (Yang and Tsiatis, 2001; Colantuoni and Rosenblum, 2015; Van Lancker et al., 2019; Wang et al., 2019, 2021; Van Lancker et al., 2022; Ye et al., 2023).

1.1.1 Hybrid Design Using ANOVA and ANCOVA

The hybrid design may arise from practical concerns. For example, there could be an administrative delay in making covariates available for inclusion in an analysis until the end of the study. Another reason could be a small sample size at an interim analysis. Two recent examples in cystic fibrosis (CF) research used this hybrid design. The GROW trial investigated the effect of oral glutathione on grown in children with CF (Bozic et al., 2020). The TEACH trial investigated the effect of adding oral azithromycin to inhaled tobramycin on lung function in people with CF who are also chronically infected with Pseudomonas aeruginosa (Nichols et al., 2021).

The classical critical values in a group sequential trial (e.g., Pocock (1977) or O'Brien and

Fleming (1979) boundaries) are predicated on conducting the same analysis at each stage of the trial. Changing the analysis strategy at the end of the trial (or any other time during the course of the trial) without proper adjustments may result in inflation of type I error, biased point estimates, and anti-conservative confidence intervals. Even with appropriate critical values, traditional point estimates and confidence intervals can lead to incorrect inference: point estimates may be biased (Kim, 1989), and confidence intervals may be anti-conservative (Kim and DeMets, 1987). Adjusted versions of these quantities have been proposed that take into account the standard group sequential design, but not the hybrid design (Kim, 1989; Kim and DeMets, 1987; Siegmund, 1978; Gernot and Brannath, 2016).

We address both of these problems. We propose adjustments for critical values to ensure correct type I error in the hybrid design. We also propose techniques to obtain median-unbiased point estimates and asymptotically correct confidence intervals in the hybrid design. These techniques mirror the traditional procedures for obtaining for point estimates and confidence intervals in group sequential designs. We justify our proposed methods theoretically. We show empirically through simulation studies that they control the type I error rate asymptotically and produce correct estimation and inference.

1.2 Preliminaries

Our contribution is presented in Section 1.3. In this section, we briefly review the aspects of group sequential designs and covariate adjustment that are critical building blocks for our proposed methods in Section 1.3. We emphasize that the technical results on covariate adjustment in Lemma 1.1 have been shown elsewhere (Yang and Tsiatis, 2001; Colantuoni and Rosenblum, 2015; Wang et al., 2019), but we re-prove them in the Supplemental Materials for the convenience of the reader. The reader familiar with group sequential designs and covariate adjustment may choose to skip ahead to Section 1.3 after reviewing the notation in our standard assumptions and definitions. We will refer to these assumptions and definitions when we propose valid methods for interim monitoring, estimation, and inference in the hybrid design.

1.2.1 Review of Group Sequential Designs

In a group sequential design, individuals are enrolled sequentially across stages of a trial. Hypothesis testing is also performed sequentially, at the end of each stage of enrollment. This procedure is depicted in Figure 1.1(a).

Assumption 1.1 formally defines the asymptotic regime that we consider for a group sequential design.

Assumption 1.1 (Group Sequential Asymptotic Regime)**.** *Let K be a fixed, maximum number of stages in a group sequential trial, and let n be the maximum sample size of the trial. Let $\{t_k; k = 1, ..., K\}$ be a set of fixed information fractions. Consider $n \to \infty$. Then,*

- $n_k = floor(n \cdot t_k)$ *is the number of individuals enrolled at at stage k*
- $n_k^* = \sum_{k'=1}^{k} n_k$ *is the cumulative number of individuals enrolled by stage k*
- $t_k^* = n_k^*/n_K$ *is the cumulative information fraction.*

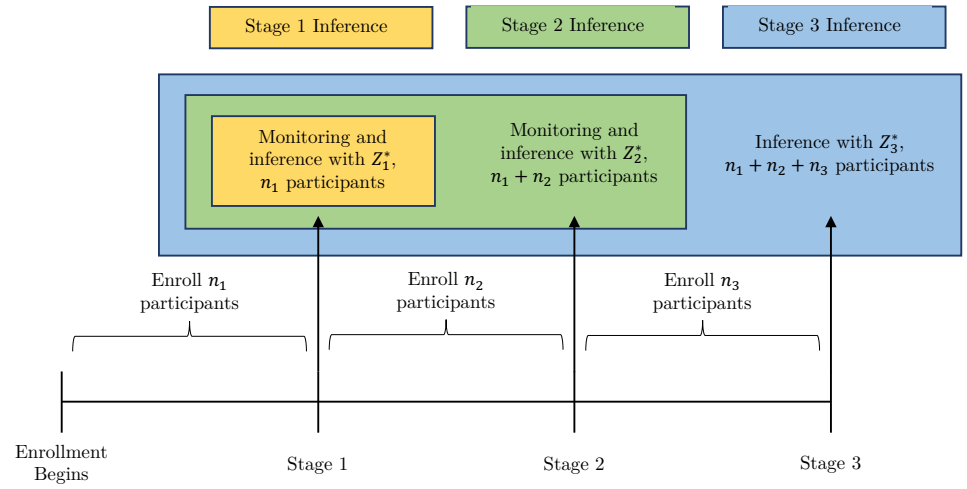
We perform a hypothesis test at stage k using a test statistic derived using n_k^* observations. The null hypothesis of no treatment effect is rejected if the test statistic exceeds stage-specific critical values $(l_k, u_k), k = 1, ..., K$. The test statistics that we consider for the hybrid design are discussed in detail in Section 1.3.

1.2.2 Review of Covariate Adjustment

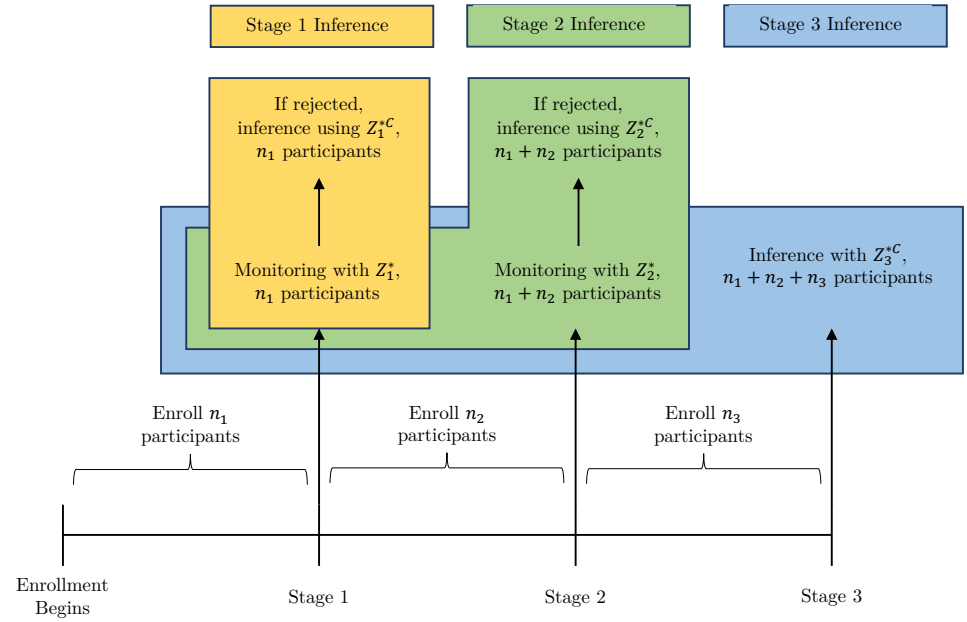
We first review classical results on covariate adjustment for estimating the average treatment effect in a randomized trial (Yang and Tsiatis, 2001; Colantuoni and Rosenblum, 2015; Wang et al., 2019). In Assumption 1.2 we describe the standard assumptions that we require of the data generating process. We note that some readers may be more familiar with treatment assignments labeled as $\{0, 1\}$ rather than $\{-1, 1\}$. All of the results in this and future sections still apply using $\{0, 1\}$. We choose to use $\{-1, 1\}$ as it is a common convention and

Figure 1.1: Diagram of group sequential trial with three stages. Figure (a) shows the standard group sequential trial design with the same analysis method to obtain Z_k^* at each stage $k = 1, ..., 3$. Figure (b) shows the modified group sequential trial design where covariate adjustment is only performed after crossing an interim monitoring boundary, or the end of the trial is reached.

(a) Standard group sequential trial design



(b) Group sequential design with delayed covariate adjustment



eases exposition of technical proofs. Furthermore, we note that the mean-zero requirement in Assumption 1.2 (iii) can be relaxed, but it also eases exposition.

Assumption 1.2 (Data Generation)**.** *We require the following assumptions of the data generating process. Let* $(Y, A, X) \sim P$*, where* P *satisfies the following:*

(i) $Y \in \mathbb{R}$ *is the continuous outcome of interest*

(ii) $A \in \{-1, 1\}$ *is the treatment assignment, and has a Rademacher distribution, i.e.,* $P(A = -1) = P(A = 1) = 0.5$

(iii) $X = (X_1, ..., X_p) \in \mathbb{R}^p$ *are* p *covariates such that* $A \perp X$*, and* p *is fixed*

Our target of inference is the average treatment effect (ATE), defined in Definition 1.1.

Definition 1.1 (Average Treatment Effect)**.** *We define the following as the average treatment effect (ATE):* $E_P[Y|A = 1] - E_P[Y|A = -1]$.

Consider the population minimizer of squared error loss for predicting Y with A, Δ^*:

$$(\Delta^*, \theta^*) = \arg \min_{\Delta, \theta \in \mathbb{R}} E_P[(Y - \theta - \Delta A)^2] \tag{1.1}$$

This is the population version of linear regression with an intercept and a treatment variable, with coefficients given by Δ and θ respectively. We give an interpretation for the population minimizer for Δ, Δ^*, in Lemma 1.1 under some identification assumptions. Now consider the case where we have explained some of the variation in the outcome Y with covariates X using a population line of best fit:

$$(\Delta^{*C}, \theta^{*C}, \gamma^{*C})^T = \arg \min_{\Delta, \theta \in \mathbb{R}, \gamma \in \mathbb{R}^p} E_P[(Y - \theta - \Delta A - X^T \gamma)^2]. \tag{1.2}$$

The assumptions listed in Assumption 1.2, as well as the following bounded variance assumption, are critical for identification of the ATE using the population minimizers in (1.1) and (1.2).

Assumption 1.3 (Bounded Variance)**.** *Define* $\tilde{\epsilon} = Y - \theta^* - \Delta^* A$ *and* $\epsilon = Y - \theta^{*C} - \Delta^{*C} A - X^T \gamma^{*C}$*. Also define* $\tilde{\sigma}^2 = \mathrm{Var}(\tilde{\epsilon})$ *and* $\sigma^2 = \mathrm{Var}(\epsilon)$*. We assume that* $\tilde{\sigma}^2 < \infty$ *and* $\sigma^2 < \infty$*.*

The sample analogs of (1.1) and (1.2) are given in the following definition, which are recognizable as linear regression estimators.

Definition 1.2 (ANOVA and ANCOVA Estimators)**.** *Consider* P *from Assumption 1.2. Suppose that we observe* n_k^* *data pairs* $(Y_i, A_i, X_i) \overset{iid}{\sim} P$*,* $i = 1, .., n_k^*$*, with* $X_i = (X_{i,1}, ..., X_{i,p})^T$*, and where each* $X_{\cdot,p}$ *has been centered at its mean, and where* $p < n_k^*$*. We define* $\hat{\Delta}_k$ *to satisfy the following:*

$$(\hat{\Delta}_k, \hat{\theta}_k) = \arg \min_{\Delta, \theta \in \mathbb{R}} \sum_{i=1}^{n_k^*} (Y_i - \theta - \Delta A_i)^2 \tag{1.3}$$

We refer to $\hat{\Delta}_k$ *as the ANOVA estimator. Similarly, we define* $\hat{\Delta}_k^C$ *to satisfy the following:*

$$(\hat{\Delta}_k^C, \hat{\theta}_k^C, \hat{\gamma}_k^C) = \arg \min_{\Delta, \theta \in \mathbb{R}, \gamma \in \mathbb{R}^p} \sum_{i=1}^{n_k^*} (Y_i - \theta - \Delta A_i - X_i^T \gamma)^2. \tag{1.4}$$

We refer to $\hat{\Delta}_k^C$ *as the ANCOVA estimator. The ANCOVA and ANOVA estimators are simply ordinary least squares (OLS) estimators of the treatment effect with and without adjusting for the covariates* X_i*, respectively. The OLS estimator for the ANOVA parameterization above is just the group-specific sample mean, i.e.,* $\hat{\theta}_k + \hat{\Delta}_k$ *and* $\hat{\theta}_k - \hat{\Delta}_k$ *are the sample means of* Y_i *within groups* $A_i = 1$ *and* $A_i = -1$*, respectively, for* $i = 1, ..., n_k^*$*.*

The following lemma summarizes well-known results about ANOVA and ANCOVA in randomized trials (Yang and Tsiatis, 2001; Colantuoni and Rosenblum, 2015; Wang et al., 2019).

Lemma 1.1 (Identification and Estimation)**.** *Consider* Δ^* *and* Δ^{*C} *from Definition 1.1 and* $\hat{\Delta}_k$ *and* $\hat{\Delta}_k^C$ *from Definition 1.2. Under Assumptions 1.2 and 1.3,*

(i) Identification. $\Delta^* = \Delta^{*C}$*, and that* $2\Delta^* = E_P[Y|A = 1] - E_P[Y|A = -1]$*.*

(ii) Consistency. $\hat{\Delta}_k \to \Delta^*$ *and* $\hat{\Delta}_k^C \to \Delta^*$. *More specifically,* $\sqrt{n_k^*}(\hat{\Delta}_k - \Delta^*) = O_P(1)$ *and* $\sqrt{n_k^*}(\hat{\Delta}_k^C - \Delta^*) = O_P(1)$.

(iii) Asymptotic Normality. $\sqrt{n_k^*}(\hat{\Delta}_k - \Delta^*) \Rightarrow N(0, \tilde{\sigma}^2)$ *and* $\sqrt{n_k^*}(\hat{\Delta}_k^C - \Delta^*) \Rightarrow N(0, \sigma^2)$, *where* $\Rightarrow$ *denotes convergence in distribution. Also,* $\sigma^2 \leq \tilde{\sigma}^2$.

Lemma 1.1(i) tells us that the population minimizers Δ^* and Δ^{*C} in the un-adjusted and adjusted problem are equivalent, and that they are a scalar multiple of the average treatment effect. Lemma 1.1(ii) tells us that under standard assumptions for randomized trials, both the ANOVA and ANCOVA estimators can be used to estimate the average treatment effect. Finally, Lemma 1.1(iii) confirms that, asymptotically, we will not lose efficiency by performing covariate adjustment. This point is subtle: the efficiency guarantees for ANCOVA only apply under equal randomization, which we assume in Assumption 1.2. These are classical results (Yang and Tsiatis, 2001; Lin, 2013; Colantuoni and Rosenblum, 2015; Wang et al., 2019; Ye et al., 2023). We build off of these results in the next section to propose valid methods for interim monitoring, estimation, and inference in the hybrid design.

1.3 Proposed Methods for the Hybrid Design

We now present our methodological contribution. Our work builds off of the results referenced in Section 1.2. Our proposal is presented in three parts. In Section 1.3.1, we derive the joint distribution of the ANOVA and ANCOVA estimators and their associated test statistics across stages $1, ..., K$ of a group sequential design. We formally define the monitoring procedure for the hybrid design in terms of a sequence of decisions based on the test statistics that we introduce. In Section 1.3.2, we use the joint distribution to form interim monitoring boundaries that have correct type I error in the hybrid design. Finally, in Section 1.3.3, we discuss how to use the joint distribution to form median unbiased point estimates and asymptotically correct confidence intervals in the hybrid design.

1.3.1 Joint Distribution of ANOVA and ANCOVA Estimators in a Group Sequential Design

We show that ANOVA and ANCOVA test statistics have a joint asymptotic Gaussian distribution across interim analyses. Additionally, their covariance matrix can be expressed as a relatively simple function of the information fraction (i.e., the relative fraction of observations accrued between two analyses) and the prognostic value of the covariates used in ANCOVA. These results are summarized in Theorem 1.1 and Corollary 1.1. These multivariate Gaussian distributions are leveraged to form corrected rejection boundaries, point estimates, and confidence intervals in the hybrid design.

Theorem 1.1 (Joint Distribution of Estimates)**.** *Consider the asymptotic regime in Assumption 1.1, and define* $\hat{\Delta}_k$ *and* $\hat{\Delta}_k^C$ *from Definition 1.2. Let*

$$\hat{\Delta} := \left(\hat{\Delta}_1, \hat{\Delta}_1^C, ..., \hat{\Delta}_K, \hat{\Delta}_K^C\right)^T.$$

In other words, $\hat{\Delta}$ *are stacked effect estimates across stages using ANOVA and ANCOVA. Then under Assumptions 1.2 and 1.3,*

$$\sqrt{n}(\hat{\Delta} - \Delta^*) \Rightarrow N\Big(0, \Sigma_{\hat{\Delta}}\Big), \quad \Sigma_{\hat{\Delta}} = T \otimes \left(\begin{smallmatrix} \tilde{\sigma}^2 & \sigma^2 \\ \sigma^2 & \sigma^2 \end{smallmatrix}\right).$$

where σ^2 *and* $\tilde{\sigma}^2$ *are defined in Assumption 1.3,* T *is a matrix with entries* $T_{(k,k')} = 1/t^*_{\max(k,k')}$*, and* $A \otimes B$ *is the Kronecker product of matrices* A *and* B*.*

The result of Theorem 1.1 is critical for deriving the joint distribution of the test statistics. For hypothesis testing in the hybrid design, we will use the following test statistics:

Definition 1.3 (ANOVA and ANCOVA Test Statistics)**.** *Consider* $(\hat{\Delta}_k, \hat{\theta}_k)$ *and* $(\hat{\Delta}_k^C, \hat{\theta}_k^C, \hat{\gamma}_k^C)$ *as the empirical minimizers from Definition 1.2. Let*

$$\hat{\sigma}_k = \left[\frac{1}{n_k^*}\sum_{i=1}^{n_k^*}(Y_i - \hat{\theta}_k - \hat{\Delta}_k A_i)^2\right]^{1/2} \quad \textit{and} \quad \hat{\sigma}_k^C = \left[\frac{1}{n_k^*}\sum_{i=1}^{n_k^*}(Y_i - \hat{\theta}_k^C - \hat{\Delta}_k^C A_i - X_i^T\hat{\gamma}_k^C)^2\right]^{1/2}.$$

We define $Z_k^* = \sqrt{n_k^*}\hat{\Delta}_k/\hat{\sigma}_k$ *and* $Z_k^{*C} = \sqrt{n_k^*}\hat{\Delta}_k^C/\hat{\sigma}_k^C$ *as the ANOVA and ANCOVA test statistics at stage* k*.*

Corollary 1.1 (Joint Distribution of Test Statistics)**.** *Consider the definitions of the ANOVA and ANCOVA test statistics* Z_k^* *and* Z_k^{*C} *from Definition 1.3, with* $p < n_1^*$*. Define the vector of stacked test statistics as* $Z^* = (Z_1^*, Z_1^{*C}, ..., Z_K^*, Z_K^{*C})^T$*, and let* T' *to be the matrix with entries* $T'_{(k,k')} = \sqrt{t^*_{\min(k,k')}/t^*_{\max(k,k')}}$*. Then under Assumptions 1.2 and 1.3, and the null hypothesis* $\Delta^* = 0$*,*

$$Z^* \Rightarrow N(0, \Sigma_{Z^*}) \quad \Sigma_{Z^*} = T' \otimes \left(\begin{smallmatrix} 1 & p \\ p & 1 \end{smallmatrix}\right)$$

where $\rho = \sigma/\tilde{\sigma}$ *from Assumption 1.3.*

Therefore, the joint distribution of the ANOVA and ANCOVA test statistics across stages of a group sequential trial is a function of the information fractions and the ratio of the asymptotic variances of ANOVA and ANCOVA. Theorem (1.1) and Corollary 1.1 are the primary results that we use to derive monitoring boundaries, adjusted estimates, and confidence intervals for the hybrid design, which we outline in Procedure 1.1. Procedure 1.1 differs from a standard group sequential trial in two ways: (a) the critical values in step (1) are modified to account for inconsistent covariate adjustment using the results from Theorem 1.1, and (b) the construction of unbiased point and confidence interval estimates constructed in step (3) are modified based on the results of Corollary 1.1, and the sample space ordering discussed in Section 1.3.3.

Procedure 1.1 (Hybrid Design of a Group Sequential Trial)**.**

1. *Select an asymptotically acceptable sequence* $\{u_k; k = 1, ..., K\}$ *of critical values (Definition 1.4) for a two-sided hypothesis test using the test statistics in Definition 1.3 (see Section 1.3.2).*

2. *For* $k = 1, ..., K$*:*

(a) *Enroll n_k participants and randomize with equal probability.*

(b) *Test to see if we should stop at the current stage and reject the null. If $k < K$, check if $|Z_k^*| \geq u_k$. If $k = K$, check if $|Z_K^{*C}| \geq u_K$.*

3. *After termination of the trial due to early stopping or continuing through stage K, use Theorem 1.1 to construct a median-unbiased point estimate and confidence interval that takes into account both covariate adjustment and the group sequential design (see Section 1.3.3).*

Details on step (1) are given in Section 1.3.2. Modifications for step (3) are discussed at a high level in Section 1.3.3. The details for step (3) are technical, and are given in the appendix. We primarily write the above procedure without those details to highlight that the conduct of the trial and skeleton of the analysis is straightforward and looks very similar to a trial which completely ignores the inconsistent covariate adjustment. Specifically, we can fix the issues with that overly simplistic analysis merely by changing decision boundaries and modifying our final point and confidence interval estimates (and associated p-values). These modifications are all based on knowing the joint distribution of ANOVA and ANCOVA test statistics and estimates, given in Theorem 1.1 and Corollary 1.1.

For clarity of exposition (and to match what is most likely done in practice), we focus on the hybrid design in Definition 1.1 and illustrated the design in Figure 1.1(b). However, we note that as long as we have the joint distribution of test statistics from (1.1), we are not restricted to only performing covariate adjustment at the end of a group sequential trial. We are also not restricted to having symmetric (or two-sided) monitoring boundaries. Everything that we discuss in the following sections can be modified to account for an arbitrary sequence of un-adjusted and covariate-adjusted estimators across stages of a trial, asymmetric monitoring boundaries, and one-sided monitoring boundaries.

1.3.2 Interim Monitoring in the Hybrid Design

Our goal in this section is to derive a stopping boundary $\{u_k; k = 1, ..., K\}$ such that we control the overall probability of making a type I error. In particular, we want to find a $\{u_k\}$ that satisfies Definition 1.4, given below. We emphasize that we do not propose a new, "hybrid" estimator to preserve type I error. Rather, we derive monitoring boundaries that preserve type I error when the well-known ANOVA and ANCOVA estimators from Definition 1.2 and their test statistics from Definition 1.3 are used in an hybrid trial design described in Definition 1.1.

Definition 1.4 (Asymptotically Acceptable Sequence of Critical Values)**.** *Consider the hybrid design in Definition 1.1. Let P_0 be the joint distribution of test statistics $(Z_1^*, ..., Z_{K-1}^*, Z_K^{*C})$ under the null hypothesis that $\Delta^* = 0$. We call $\{u_k; k = 1, ..., K\}$ an* asymptotically acceptable sequence of critical values *if*

$$\lim_{n\to\infty} P_0(|Z_1^*| \geq u_1 \text{ or } |Z_2^*| \geq u_2 \text{ or } ... \text{ or } |Z_{K-1}^*| \geq u_{K-1} \text{ or } |Z_K^{*C}| \geq u_K) = \alpha \qquad (1.5)$$

where α is a pre-specified type I error probability.

We have already derived the joint asymptotic distribution of the test statistics in Corollary 1.1. In addition to the information fractions, knowledge (or an accurate estimate) of $\rho \equiv \sigma/\tilde{\sigma}$ is the key quantity that determines the joint distribution. There are infinite choices of asymptotically acceptable sequences $\{u_k\}$. We focus on the O'Brien and Fleming (1979) style of boundary, but note that these can be modified for any shape that is desired (e.g., for Pocock (1977), the shape has c for each stage, rather than $c/\sqrt{k}$). We present two examples of asymptotically acceptable sequences of critical values for the hybrid design in Examples 1.1 and 1.2.

Example 1.1 (Uniform-Inflated OBF Critical Values)**.** *Let* $U^\rho \sim N_K(0, \Sigma_\rho)$ *with*

$$\Sigma(\rho) := \begin{pmatrix} 1 & \sqrt{t_1^*/t_2^*} & ... & \sqrt{t_1^*/t_{K-1}^*} & \rho\sqrt{t_1^*/t_K^*} \\ \sqrt{t_1^*/t_2^*} & ... & ... & ... & \rho\sqrt{t_2^*/t_K^*} \\ ... & ... & ... & ... & ... \\ \sqrt{t_1^*/t_{K-1}^*} & ... & ... & ... & \rho\sqrt{t_{K-1}^*/t_K^*} \\ \rho\sqrt{t_1^*/t_K^*} & \rho\sqrt{t_2^*/t_K^*} & ... & \rho\sqrt{t_{K-1}^*/t_K^*} & 1 \end{pmatrix} \tag{1.6}$$

Under Assumptions 1.1, 1.2, and 1.3, define c *as the value that satisfies the following:*

$$P(|U_1^\rho| \geq c \text{ or } |U_2^\rho| \geq c/\sqrt{2} \text{ or ... or } |U_{K-1}^\rho| \geq c/\sqrt{K-1} \text{ or } |U_K^\rho| \geq c/\sqrt{K}) = \alpha. \tag{1.7}$$

Then $\{c, c/\sqrt{2}, ..., c/\sqrt{K}\}$ *is an asymptotically acceptable sequence of critical values for* $(Z_1^*, ..., Z_{K-1}^*, Z_K^{*C})$ *in the hybrid design described in Procedure 1.1.*

Example 1.2 (End-Inflated OBF Critical Values)**.** *Let* $U^\rho \sim N_K(0, \Sigma(\rho))$ *and* $U^1 \sim N_K(0, \Sigma(1))$ *with* Σ *defined in Example 1.1. Note, by* $\Sigma(1)$ *we mean to replace all values of* ρ *in* (1.6) *with* 1*. Define* c *as the value that satisfies*

$$P(|U_1^1| \geq c \text{ or } |U_2^1| \geq c/\sqrt{2} \text{ or ... or } |U_{K-1}^1| \geq c/\sqrt{K-1} \text{ or } |U_K^1| \geq c/\sqrt{K}) = \alpha. \tag{1.8}$$

Subsequently, define c' *as the value that satisfies*

$$P(|U_1^\rho| \geq c \text{ or } |U_2^\rho| \geq c/\sqrt{2} \text{ or ... or } |U_{K-1}^\rho| \geq c/\sqrt{K-1} \text{ or } |U_K^\rho| \geq c'/\sqrt{K}) = \alpha. \tag{1.9}$$

Then $\{c, c/\sqrt{2}, ..., c/\sqrt{K-1}, c'/\sqrt{K}\}$ *is an asymptotically acceptable sequence of critical values for* $(Z_1^*, ..., Z_{K-1}^*, Z_K^{*C})$ *in the hybrid design described in Procedure 1.1.*

The critical values above do not have a closed form, but can be calculated numerically without too much difficulty (see Supplemental Materials). The critical values in Example 1.1 have the exact same shape as the classical OBF boundaries (Figure 1.2) but are uniformly

inflated to account for the decreased correlation between the last stage test statistic with previous stages. In contrast, the critical values in Example 1.2 follow the classical OBF boundaries up to the last stage. At the last stage of the trial, the critical value is inflated to a level that ensures exactly α type I error is spent (Figure 1.3). In practice, it is unlikely that we will know ρ during trial design. This makes the critical values in Example 1.1 less practical. However, the boundaries in Example 1.2 can be used by plugging in a consistent estimator of ρ at the end of the trial.

We have focused on the hybrid design where covariate adjustment is performed when the trial ends (Procedure 1.1). However, our methods are readily extended to the setting where covariate adjustment is performed at any stage(s) of a trial. In this case, we recommend using "alpha-spending"(Lan and DeMets, 1983): rather than pre-specifying the critical values at each stage, they are calculated on the fly based on a pre-specified amount of type I error to be spent at each stage. We provide some further details about how one might employ alpha-spending in the Supplementary Material.

1.3.3 Point Estimates, p-values, and Confidence Intervals in the Hybrid Design

Once the trial has ended, we want to compute point estimates and confidence intervals for the treatment effect and p-values for the test of the null hypothesis. Computing these quantities without taking into consideration the group sequential design may lead to incorrect inference: point estimates may be biased (Kim, 1989), and inference based on p-values and confidence intervals may be anti-conservative (Kim and DeMets, 1987). Methods have been proposed to adjust point estimates (median unbiased point estimates), p-values, and confidence intervals (Kim, 1989; Kim and DeMets, 1987; Siegmund, 1978; Gernot and Brannath, 2016). Briefly, these methods use the joint distribution (under the null and various alternative hypotheses) of the test statistics across stages and an ordering that allows us to compare two test statistics and define one as "more extreme" than the other. One of our proposed modifications is straightforward: to instead use the joint distribution of the test statistics and effect estimates from Corollary 1.1 and Theorem 1.1 that take into account inconsistent covariate adjustment.

Figure 1.2: Uniform-inflated Pocock and OBF bounds in a 4-stage trial with equally-sized stages.

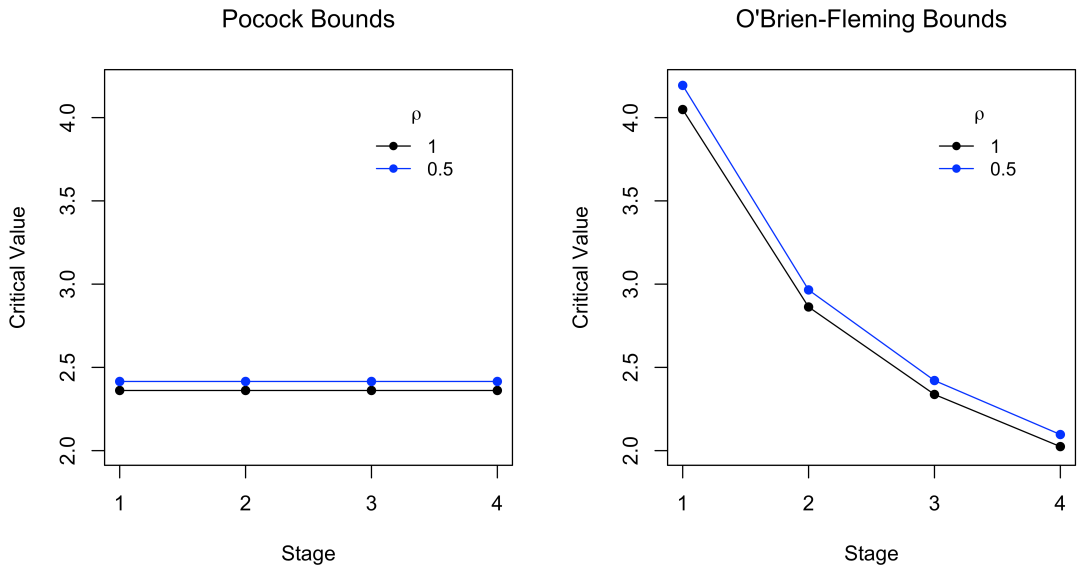
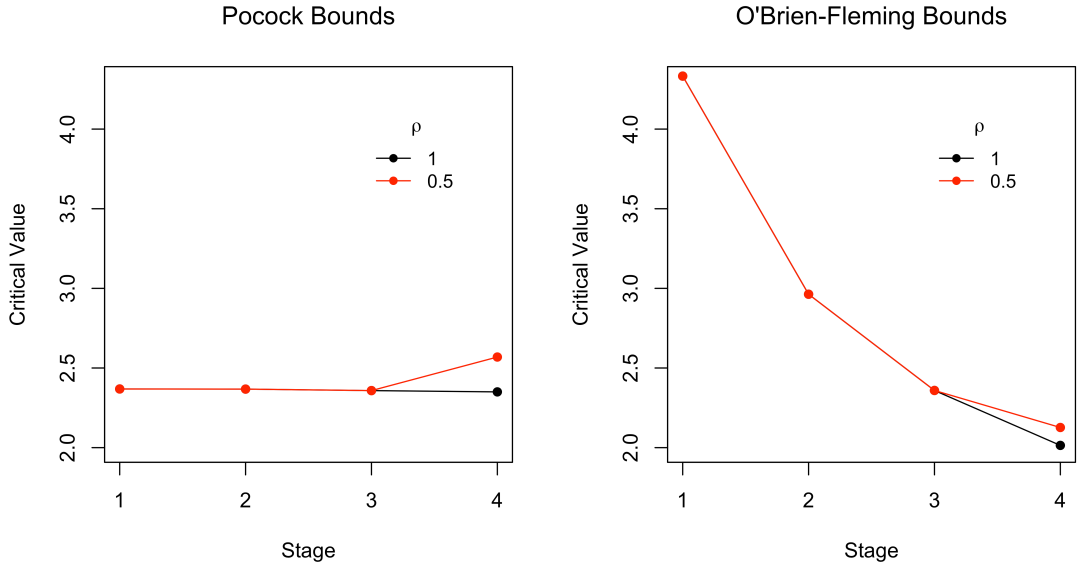


Figure 1.3: End-inflated Pocock and OBF bounds in a 4-stage trial with equally-sized stages.



A more subtle modification is required for the ordering of test statistics in the sample space and computing associated p-values (which we then use to compute median unbiased point estimates and confidence intervals). In a trial with only one stage, we order test statistics based on their position on the real number line. Group sequential trials have a two-dimensional test statistic: the observed statistic and the stage at which it was observed. There are several common sample space orderings for test statistics from group sequential trials, including stage-wise ordering and sample mean ordering (Gernot and Brannath, 2016). For the hybrid design described in Procedure 1.1, we use the same types of orderings, and we base the ordering on the ANCOVA statistic at the end of the trial (either through reaching the final stage or from early termination). Finally, the computation of p-values based on the chosen ordering must be modified to account for the possibility of both un-adjusted and covariate-adjusted analyses being performed within the same stage of a trial after an early trial termination. In the Supplemental Material, we give full details of our strategy for forming adjusted estimates, p-values, and confidence intervals using the key results from Theorem 1.1 and Corollary 1.1.

1.4 Simulation Study

We analyze the performance of our adjustments for interim monitoring boundaries, point estimates, and confidence intervals via simulation. We base the trial designs on two published trials in cystic fibrosis (CF).

1.4.1 Simulation Parameters

The GROW (Bozic et al., 2020) trial was an RCT of 58 children with CF between 2-10 years old to evaluate the effect of oral glutathione on their growth. The primary outcome was change in weight-for-age Z-score from visits 2-4. There was one interim look after 25 participants (about 50% accrual). The trial protocol specified ANOVA at the interim analysis and ANCOVA at the final analysis, using the weight-for-age z-score at visit 3 as a covariate. The TEACH (Nichols et al., 2021) trial was an RCT of 108 people over 12

years old with CF and chronic Pseudomonas aeruginosa infection. The goal was to evaluate the effect of adding oral azithromycin to inhaled tobramycin on lung function. The trial protocol specified ANOVA at two interim analyses (after 50% and 75% accrual) and a final analysis with ANCOVA using the randomization strata as a covariate. We explore the use of additional baseline covariates available for analysis in the TEACH and GROW trials as well.

Based on the GROW and TEACH trials, we consider sample sizes of 50, 100, and 250, and 1000 for a 3-stage trial with information fractions 0.33, 0.67, and 1.0. We fix $\tilde{\sigma}^2 = 1$, $E[X_j] = 0$, and $\text{Var}[X_j] = 1$, $j = 1, ..., p$, and consider $p = 1$. Given a ρ, it's then possible to solve for a vector γ where each entry is identical, i.e., all of the covariates having the same coefficient, that ensures that $\text{Var}[X^T\gamma] = 1 - \rho^2$. We consider effect sizes for Δ of 0.0, 0.1, 0.2, and 0.5, prognostic values of baseline covariates ρ of 0.25, and 0.5 (smaller values of ρ corresponding to higher prognostic value). We present results for Pocock-type boundaries. See Supplemental Material for 2-stage trials, two covariates, and OBF-type boundaries.

1.4.2 Type I Error and Power in the Hybrid Designs

We analyze type I error and power in a simulation study considering five scenarios of how a trial might be conducted, outlined below. Results are shown in Figures 1.4 and 1.5.

(A) The same analysis method is performed for interim monitoring and the final analysis, and critical values are computed during trial design.

  (i) ANOVA or (ii) ANCOVA is used for both monitoring and final analysis.

(B) Hybrid design: ANOVA is used for interim monitoring, ANCOVA is used in the final analysis

  (i) No adjustment of the monitoring boundaries.

  (ii) Example 1.1: uniform-inflated bounds with ρ known.

(iii) Example 1.2: end-inflated bounds with ρ estimated at the last stage of the trial.

At small sample sizes, the type I error is inflated for all trial procedures. As the sample size increases, most procedures have nominal type I error. Importantly, the type I error when switching from ANOVA to ANCOVA without adjusting the boundaries is uniformly high. Both types of adjustment (uniform- and end-inflated) correct this inflation in type I error. The implication is that boundaries need to be adjusted in the hybrid design, especially when highly prognostic covariates are used. In the Supplemental Material we explore the impact of incorrectly specifying ρ at the beginning of the trial when computing boundaries. As expected, more prognostic covariates increases the power of procedures that use covariate adjustment. Importantly, there is not a large difference between the power in the hybrid design scenarios. The implication is that adjusting the boundaries to control type I error in the hybrid design does not hurt power.

1.4.3 Performance of Adjusted Point Estimates and Confidence Intervals in the Hybrid Design

In Table 1, we show bias and coverage of several strategies to create point estimates and confidence intervals in the hybrid design. The first strategy is to compute point estimates and CIs ignoring the group sequential design ("simple"). The second strategy takes into account the group sequential design using stage-wise ordering, but not the hybrid design ("GS"). The third strategy takes into account the group sequential hybrid design with our modified stage-wise ordering ("GS + Adjust").

The simple point estimate and confidence interval generally has small median bias and acceptable coverage. However, there are several scenarios where "GS + Adjust" substantially out performs has the simple point estimate (e.g., $\Delta = 0.1$, $\rho = 0.5$, $n = 250, 1000$). In almost all scenarios, "GS + Adjust" has higher coverage than the simple confidence intervals. The point estimates for "GS + Adjust" are constructed so that they will be median unbiased; as a result, there is a small amount of mean bias in the point estimates.

Figure 1.4: Type I error under scenarios (A)(i)-(ii) and (B)(i)-(iii) across 25,000 simulated trials.

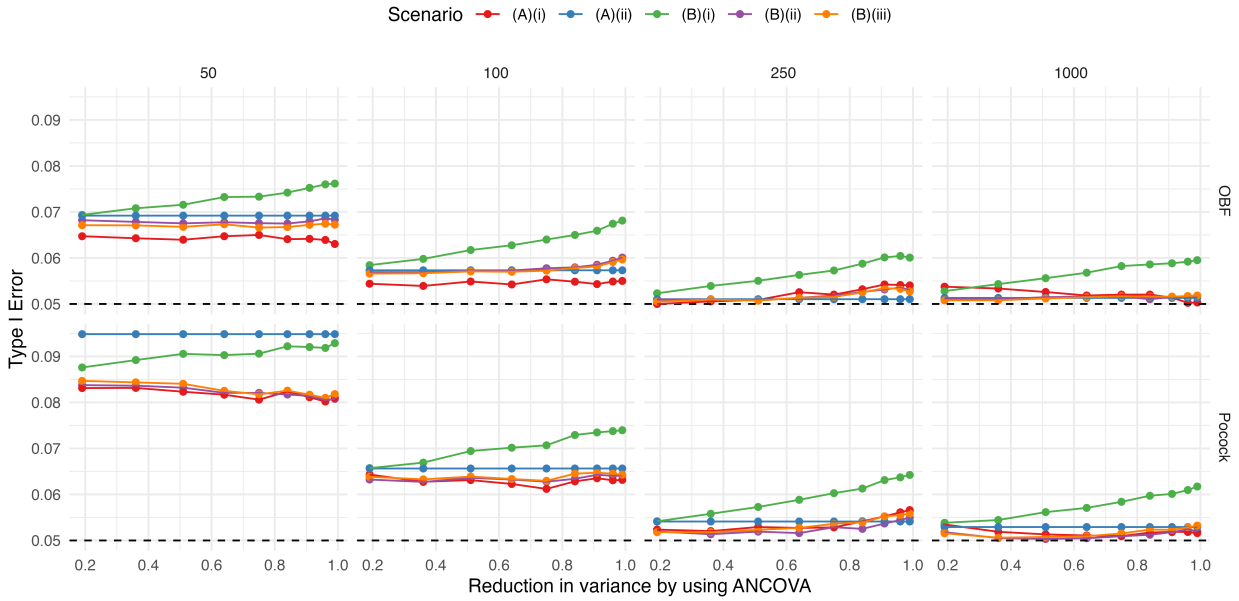
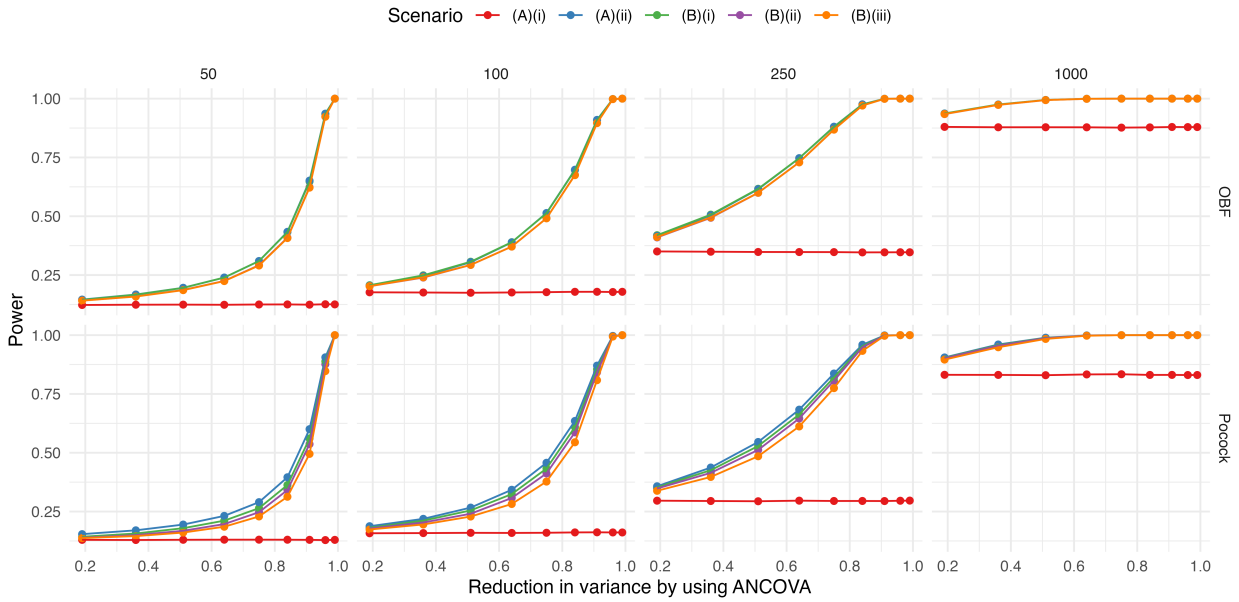


Figure 1.5: Power under scenarios (A)(i)-(ii) and (B)(i)-(iii) with an effect size of 0.1 across 25,000 simulated trials.



Importantly, ignoring the hybrid design when using the group sequential adjustment for point estimates and confidence intervals has serious consequences. Under some simulation parameters, the "GS" approach leads to large bias, particularly for covariate adjustment using highly prognostic covariates. The confidence intervals may have over- or under-coverage. In some cases, the under-coverage is extreme (e.g., 16%). The implication of these simulation results is that if point estimates and confidence intervals use standard group sequential adjustment methods, they should be modified for the hybrid design.

1.5 Conclusions

Covariate adjustment using ANCOVA can provide important efficiency gains over ANOVA. If covariate-adjusted estimation and inference will be used at the end of the trial, it is ideal to also perform covariate adjustment in the same way during interim monitoring (Van Lancker et al., 2019). Not only is this the most powerful approach, but standard group sequential trial methodology relies on the same hypothesis testing procedure that is performed at each stage of the trial. However, investigators may not be able to perform covariate adjustment at interim analyses due to practical constraints. Failing to account for this in construction of the monitoring boundaries, and in final estimation and inference, may result in incorrect decisions on rejection of the null hypothesis, and may lead to incorrect point estimates and confidence intervals. Inconsistent use of covariate adjustment changes the covariance between group sequential test statistics. We derived the covariance matrix of the test statistics in a group sequential trial under the hybrid design. We used it to construct group sequential monitoring boundaries that preserve type I error, and to obtain correct estimation and inference for treatment effects under a slightly modified stage-wise ordering of the test statistics.

Our simulations demonstrate two key implications of inconsistent covariate adjustment. If baseline covariates are expected to be prognostic, but will not be used until the end of the trial, then it is important to adjust the monitoring boundaries for the hybrid design. In some cases, failing to adjust the boundaries resulted in a 1-2% inflation in type I error, even at large sample sizes. A small inflation in type I error can have a big impact on the

Table 1.1: Bias (×100) and coverage results based on 10,000 simulated trials for each row, end-inflated Pocock boundaries for 3-stage trial. Bias is shown as median bias and mean bias in parentheses. Δ is the true effect size, ρ is the true ratio of standard error between the ANCOVA and ANOVA estimators, and n is the total sample size for the trial.

			Scaled Bias: Median (Mean)			Coverage		
Δ	ρ	n	Simple	GS	GS + Adjust	Simple	GS	GS + Adjust
0.00	0.25	50	-0.03 (0.01)	-0.14 (-0.16)	-0.04 (0.03)	0.93	0.94	0.92
0.00	0.25	100	0.01 (-0.00)	-0.05 (-0.11)	0.02 (-0.03)	0.94	0.95	0.94
0.00	0.25	250	0.02 (0.01)	-0.02 (-0.07)	0.02 (0.05)	0.95	0.96	0.94
0.00	0.25	1000	-0.00 (-0.00)	-0.02 (-0.04)	0.00 (0.02)	0.95	0.96	0.95
0.00	0.50	50	-0.05 (0.03)	-0.21 (-0.26)	-0.06 (-0.02)	0.92	0.94	0.92
0.00	0.50	100	0.02 (-0.01)	-0.11 (-0.23)	0.03 (0.01)	0.93	0.95	0.94
0.00	0.50	250	0.03 (0.01)	-0.02 (-0.11)	0.03 (0.01)	0.94	0.96	0.95
0.00	0.50	1000	-0.00 (-0.01)	-0.03 (-0.07)	0.00 (-0.01)	0.95	0.96	0.95
0.10	0.25	50	0.03 (0.15)	-1.63 (-1.95)	0.06 (1.99)	0.93	0.94	0.93
0.10	0.25	100	0.07 (0.09)	-3.38 (-3.16)	0.12 (1.87)	0.94	0.96	0.94
0.10	0.25	250	0.06 (0.09)	-5.68 (-4.92)	0.01 (1.58)	0.95	0.18	0.94
0.10	0.25	1000	0.03 (0.09)	-6.87 (-4.90)	-0.02 (0.35)	0.95	0.30	0.95
0.10	0.50	50	0.15 (0.58)	-0.30 (-0.75)	0.10 (1.78)	0.92	0.93	0.93
0.10	0.50	100	0.19 (0.41)	-0.39 (-1.08)	0.16 (1.56)	0.93	0.94	0.94
0.10	0.50	250	0.21 (0.40)	-1.85 (-1.73)	0.05 (1.35)	0.94	0.95	0.95
0.10	0.50	1000	0.16 (0.36)	-3.78 (-2.78)	-0.01 (0.47)	0.95	0.62	0.95
0.20	0.25	50	0.09 (0.26)	-10.24 (-8.74)	0.20 (3.56)	0.93	0.29	0.94
0.20	0.25	100	0.10 (0.24)	-12.97 (-10.45)	0.17 (2.87)	0.94	0.16	0.94
0.20	0.25	250	0.12 (0.19)	-13.80 (-9.74)	-0.05 (0.65)	0.95	0.31	0.95
0.20	0.25	1000	-0.00 (0.01)	-0.12 (-1.51)	-0.02 (-0.54)	0.95	0.86	0.95
0.20	0.50	50	0.39 (1.08)	-3.01 (-2.80)	0.35 (3.06)	0.92	0.93	0.93
0.20	0.50	100	0.43 (0.92)	-6.36 (-4.37)	0.36 (2.56)	0.93	0.94	0.94
0.20	0.50	250	0.43 (0.72)	-7.71 (-5.57)	0.03 (0.91)	0.95	0.60	0.95
0.20	0.50	1000	0.01 (0.08)	-0.15 (-1.06)	-0.03 (-0.34)	0.95	0.87	0.95
0.50	0.25	50	0.19 (0.48)	-34.85 (-21.51)	-0.61 (-0.01)	0.92	0.37	0.93
0.50	0.25	100	0.15 (0.24)	-1.75 (-11.15)	-0.12 (-1.81)	0.94	0.67	0.94
0.50	0.25	250	0.01 (0.02)	-0.02 (-0.54)	0.01 (-0.25)	0.95	0.94	0.95
0.50	0.25	1000	-0.02 (-0.01)	-0.02 (-0.01)	-0.02 (-0.01)	0.95	0.95	0.95
0.50	0.50	50	0.89 (1.82)	-20.22 (-12.68)	-0.79 (1.04)	0.92	0.56	0.92
0.50	0.50	100	0.43 (0.83)	-2.43 (-7.32)	-0.22 (-0.77)	0.94	0.67	0.94
0.50	0.50	250	0.03 (0.06)	-0.02 (-0.36)	0.00 (-0.15)	0.95	0.94	0.95
0.50	0.50	1000	-0.04 (-0.02)	-0.04 (-0.02)	-0.04 (-0.02)	0.95	0.95	0.95

Bayesian posterior probability of a beneficial treatment (Emerson et al., 2007). Modifying the boundaries using our strategies brings the type I error to the nominal level. Additionally, if group sequential adjustment for point estimates and confidence intervals is desired, these should further be adjusted for the hybrid design. Ignoring the hybrid design can result in high bias and extremely low coverage.

Though we focused on a particular trial design, the concepts and methods that we discuss here may be extended to other designs. Our approach could be extended to other trial settings, such as trials with multiple treatment arms or trials with unequal randomization probabilities. We focused on trials satisfying Assumption 1.2 because ANCOVA is a simple to understand and widely-used approach, and has robust and efficient properties in this setting (Lin, 2013; Wang et al., 2019; Ye et al., 2023). ANCOVA2 (or ANHECOVA) which uses linear models with interaction terms is an approach better suited for trials with multiple treatment arms, or unequal randomization probabilities. However, studying this estimator requires more involved influence function derivations (as we will see in Chapter 2), and it would potentially have a more complicated covariance matrix in the hybrid design.

Our approach could also be extended to other covariate-adjusted estimators that are asymptotically linear (as others have noted (Van Lancker et al., 2022)), since the covariance matrix for the hybrid design can be characterized using the estimators' influence functions. This would enable the study of estimators tailored for binary outcomes, like logistic regression. However, covariate adjustment using non-linear models in clinical trials, even without the group sequential design, is not always straightforward. Covariate-adjusted estimators do not inherently possess the robustness and efficiency properties like ANCOVA. For example, covariate adjustment may change the estimand and the null hypothesis under model misspecification (Robinson and Jewell, 1991). Even if the covariate-adjusted estimator targets the same estimand as the un-adjusted estimator, it does not necessarily gain efficiency, which may make it undesirable (Guo and Basse, 2021; Cohen and Fogarty, 2024). There are several "fixes" like standardization (Rosenblum and Steingrimsson, 2016) and calibration (Cohen and Fogarty, 2024), but these are more complicated than a simple estimator like

ANCOVA. Integrating covariate adjustment with group sequential designs considering these nuances should be the subject of future work.

Constructing monitoring boundaries and performing group sequential-adjusted estimation and inference requires computing probabilities of regions of multivariate Gaussians. Standard group sequential methodology uses recursive integration techniques which simplify the calculation of probabilities of a multivariate Gaussian to a sequence of probabilities from univariate Gaussians (Armitage et al., 1969). Recursive integration relies on "independent increments," i.e., that incremental changes to test statistics across stages in a trial are independent (Kim and Tsiatis, 2020). The hybrid design breaks the independent increments property. Therefore, rather than use recursive integration techniques, we use numerical integration across regions of a multivariate Gaussian. Using standard statistical software, this integration is feasible for trials that have a reasonable number of stages (e.g., ≤ 5). Others have pursued strategies that orthogonalize estimators across consecutive stages in order to recover independent increments and avoid integrating multivariate Gaussians (Van Lancker et al., 2022). However, it is not necessary for the settings we have discussed in this paper.

Future work may consider extensions to the covariate-adaptive randomization setting. Using a working model for covariate adjustment that includes treatment-by-covariate interactions achieves a guaranteed efficiency gain and universal applicability under a variety of covariate-adaptive randomization strategies where the ANCOVA procedure we have discussed here does not (Ye et al., 2023). Extensions of our work to the covariate-adaptive randomization setting would be particularly useful for trials where categorical covariates are used for randomization, but where estimates adjusted for additional clinically relevant baseline covariates are desired at the end of a trial. Finally, as mentioned, our results – and existing standard group sequential monitoring methodology – are asymptotic. Future work could consider developing parallel results for the finite sample regime.

Chapter 2

A GENERAL FORM OF COVARIATE ADJUSTMENT IN RANDOMIZED CLINICAL TRIALS

This chapter is based on the following article, by permission of Oxford University Press:

Marlena S Bannick, Jun Shao, Jingyi Liu, Yu Du, Yanyao Yi, and Ting Ye (2025). A General Form of Covariate Adjustment in Clinical Trials under Covariate-Adaptive Randomization. Biometrika, 112 (3) DOI:10.1093/biomet/asaf029.

2.1 Introduction

In randomized clinical trials, the credibility and precision of estimated treatment effects are of utmost importance. Prognostic baseline covariates can be utilized – called *covariate adjustment* – to increase the precision of estimated treatment effects and power of hypothesis tests (ICH, 2019; U.S. Food and Drug Administration, 2024; EMA, 2015). Based on extensive research and discussion in recent decades, the U.S. Food and Drug Administration (FDA) recently released guidance for sponsors and trial practitioners on best practices for covariate adjustment (U.S. Food and Drug Administration, 2024). Among them is the importance of *model-assisted* analyses. That is, although a working model between the responses and covariates may be used to assist estimating treatment effect, the resulting estimator must be valid even if the working model is misspecified and "under approximately the same minimal statistical assumptions that would be needed for unadjusted estimation" (U.S. Food and Drug Administration, 2024).

Covariate-Adaptive Randomization (CAR), such as stratified permuted block randomization (Zelen, 1974) and Pocock and Simon's minimization (Taves, 1974; Pocock and Simon, 1975), is widely applied for balancing important prognostic baseline factors across treatment groups to enhance the credibility of trial results (EMA, 2015). CAR is a complementary approach to covariate adjustment at the analysis stage since it decreases the variance of the estimated treatment effect. Unlike simple randomization, however, CAR generates a dependent sequence of treatment assignments. For instance, in stratified permuted block randomization, participants within each stratum are randomized in "blocks" with a predetermined block size to achieve the desired allocation ratio by the end of each block. If the block size is four in a two-arm trial with equal allocation, then precisely two out of four participants in a block are assigned to each arm in a random fashion. Thus, to fully realize the power of CAR, analysis with or without further covariate adjustment should account for this dependence (U.S. Food and Drug Administration, 2024).

Adjusting for baseline covariates at the analysis stage using linear models has a long

practical history and strong theoretical justification (Yang and Tsiatis, 2001; Zhang et al., 2008; Tsiatis et al., 2008; Lin, 2013; Wang et al., 2019, 2021; Ye et al., 2023; Liu and Yang, 2020, among others). Results can also be found under generalized linear models for binary or count outcomes (Freedman, 2008; Moore and van der Laan, 2009; Rosenblum and van der Laan, 2010; Steingrimsson et al., 2017; Guo and Basse, 2021; Cohen and Fogarty, 2024). Many proposals for model-assisted covariate adjustment, including those listed above, are the same as or asymptotically equivalent to the *augmented inverse propensity weighted* (AIPW) method (Robins et al., 1994; Zhang et al., 2008), which is the focus of this article. To facilitate its practical uses, there is a need for a comprehensive theory of AIPW estimators with arbitrary – and potentially nonparametric – working models under CAR. We establish theoretical results that show a complete picture of the asymptotic properties of AIPW estimators of the vector of population mean responses under different treatments, in the context of CAR. Our theory covers the use of machine learning working models and cross-fitting, which is the first rigorous theoretical justification of these methods under CAR that produces dependent data. Beyond asymptotic normality, we provide conditions and results for *universal applicability* (i.e., the asymptotic distribution of AIPW estimator is invariant to CAR), and *guaranteed efficiency gain* (i.e., the AIPW estimator is asymptotically at least as efficient as the unadjusted estimator, even in cases of working model misspecification).

Furthermore, we provide insights into the conditions for guaranteed efficiency gain and universal applicability. Our theoretical results motivate our proposal of a joint calibration estimator that achieves both guaranteed efficiency gain and universal applicability regardless of the initial working model. For assessment of accuracy and inference, we derive estimators of asymptotic covariance matrices that are both robust to model misspecification and correct under CAR, bridging the gap between the practical guidelines from the FDA (U.S. Food and Drug Administration, 2024) and the theoretical underpinnings in statistical literature. All estimators for our simulation study and data application are computed using our RobinCar R package, available on both CRAN and GitHub.

2.2 Setup, Estimators, and Assumptions

Consider a trial with K treatments, where treatment $a = 1, ..., K$ has a pre-specified assignment proportion $\pi_a > 0$ and $\sum_{a=1}^{K} \pi_a = 1$. Let $Y_1, ..., Y_K$ be potential responses of a participant and $\theta_a = E(Y_a)$, the population mean of Y_a. We consider estimation and inference on a given function of the parameter vector $\theta = (\theta_1, \ldots, \theta_K)^T$, where c^T denotes the transpose of c. Let X be the vector of baseline covariates used for adjustment at the analysis stage and Z be a discrete baseline covariate vector used in CAR, where Z has a fixed L joint levels $z_1, ..., z_L$ and X and Z may share some components. The distribution of X and Z is unaffected by treatment, which is the reason why we can improve efficiency by adjusting for baseline covariates.

A random sample of n participants is obtained. For the ith participants, let $Y_{1,i}, ..., Y_{K,i}$, X_i, Z_i be realizations of $Y_1, ..., Y_K$, X, Z, respectively, $i = 1, ..., n$. We assume that $Y_{1,i}..., Y_{K,i}, X_i, Z_i$, $i = 1, ..., n$, are independent and identically distributed with finite second order moments. Based on a given randomization scheme, participant i is assigned to a single treatment denoted by A_i and has the observed response $Y_{A_i,i}$ and baseline covariates X_i and Z_i, $i = 1, ..., n$. Let P_n be the empirical measure placing weight $1/n$ on each ith observation, and P_0 population measure which governs the distribution of $(Y_1, \ldots, Y_K, X, Z)$. For example, $P_n X = \frac{1}{n} \sum_{i=1}^{n} X_i$, and $P_0 X = E(X)$. We use the notation $\mathcal{A}_n = (A_1, \ldots, A_n)$ and $\mathcal{Z}_n = (Z_1, \ldots, Z_n)$ to indicate the vector of treatment indicators and strata variables across all participants.

We treat simple randomization as a special case of CAR, in which A_i's are assigned without regard to Z_i's and they are independent across observations $i = 1, ..., n$ with $P(A_i = a) = \pi_a$. Under CAR other than simple randomization, A_i depends on the treatment assignments and covariates for all previous participants. Thus, the sequence of treatment assignments, $A_1, ..., A_n$, is interrelated and depends on Z_i's used in randomization. All commonly used CAR schemes satisfy the following mild assumption (Antognini and Zagoraiou, 2015; Hu et al., 2023).

Assumption 2.1 (Treatment Assignments)**.** *(i) Given covariate* $(Z_1, ..., Z_n)$*, the assignment vector* $(A_1, ..., A_n)$ *is conditionally independent of* $\{Y_{1,i}, ..., Y_{K,i}, X_i,\ i = 1, ..., n\}$*; (ii) For each* a*,* $P(A_i = a \mid Z_1, ..., Z_n) = \pi_a$*; (iii) For every level* z *of* Z*, the sequence* $\sqrt{n}\{n_a(z)/n(z) - \pi_a\}$ *is bounded in probability as* $n \to \infty$*, where* $n(z)$ *is the number of participants with* $Z = z$ *and* $n_a(z)$ *is the number of participants with* $Z = z$ *and assigned to treatment* a*.*

Assumption 2.1(i) means that the treatment assignment vector $\mathcal{A}_n$ depends solely on the covariate $\mathcal{Z}_n$. Assumption 2.1(ii) ensures that the allocation ratio is preserved at π_a for every step; although this implies that A_i is independent of $\mathcal{Z}_n$, the whole vector $\mathcal{A}_n$ is not independent of $\mathcal{Z}_n$ under CAR other than simple randomization. Assumption 2.1(iii) requires a rate at which $n_a(z)/n(z)$ converges to π_a and is satisfied by most commonly used CAR schemes. For example, it holds for simple randomization; it holds for stratified permuted block randomization, as $|n_a(z) - n(z)\pi_a|$ is bounded by the block size; for Pocock and Simon's minimization, Assumption 2.1(iii) holds (see Hu et al. (2022) (Hu et al., 2023)).

Without covariate adjustment, the unadjusted estimator of θ is the sample mean vector $\bar{Y} = (\bar{Y}_1, ..., \bar{Y}_K)^T$, where $\bar{Y}_a$ is the sample mean of $Y_{a,i}$'s for those assigned to treatment a. Throughout, $\bar{Y}$ is considered as the benchmark for comparison with covariate adjustment.

We consider model-assisted adjustment for covariate X using a working model for the conditional response mean $E(Y_a \mid X = x)$. Working models can be arbitrary and include parametric or nonparametric machine learning methods such as random forest. Let $\hat{\mu}_a(x)$ be the estimated value of $E(Y_a \mid X = x)$ obtained by fitting a working model under treatment a. To estimate θ, we utilize the AIPW estimator (Robins et al., 1994; Zhang et al., 2008) given by

$$\hat{\theta}_{\text{AIPW}} = \left(\bar{Y}_a - \frac{1}{n_a} \sum_{i:A_i=a} \hat{\mu}_a(X_i) + \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}_a(X_i), \quad a = 1, ..., K \right)^T, \tag{2.1}$$

where n_a is the number of participants in treatment arm a. The AIPW estimator in (2.1) is a general form of covariate adjustment and robust against working model misspecification, because $\bar{Y}_a$ is asymptotically correct for estimating θ_a and $n_a^{-1} \sum_{i:A_i=a} \hat{\mu}_a(X_i) - n^{-1} \sum_{i=1}^{n} \hat{\mu}_a(X_i)$

has asymptotic mean zero, as the distribution of X is not affected by treatments. If $\bar{Y}_a = n_a^{-1} \sum_{i:A_i=a} \hat{\mu}_a(X_i)$ almost surely, which is referred to as prediction unbiasedness (Guo and Basse, 2021), then $\hat{\theta}_{\text{AIPW}}$ is the same as the g-computation or generalized Oaxaca-Blinder estimator (Freedman, 2008). Without prediction unbiasedness, the g-computation estimator may be seriously biased and $\hat{\theta}_{\text{AIPW}}$ is a de-biased g-computation estimator. A simple simulation example comparing the g-computation and AIPW estimators is shown in the Supplementary Material.

It is often of interest to estimate a known differentiable function f of θ like a treatment contrast $f(\theta) = \theta_2 - \theta_1$ or a relative risk $f(\theta) = \theta_2/\theta_1$. We estimate $f(\theta)$ using the plug-in estimator $f(\hat{\theta}_{\text{AIPW}})$. This is an "augment-then-contrast" approach as it first augments the response mean estimators and then takes a contrast of the augmented estimators. A discussion about a "constrast-then-augment" approach, which can be applied when θ is defined as a solution of estimating equation (Zhang et al., 2008), is given in the Supplementary Material Section F. There, we show that if an augmented estimating equation-based estimator in the form of Zhang et al. (2008) (Zhang et al., 2008) is consistent and asymptotically normal under simple randomization for a given estimand, then under very mild regularity conditions, it is also consistent and asymptotically normal for that estimand under CAR.

In addition to Assumption 1 for CAR, our asymptotic results in Sections 2.3 and 2.4 require the following two assumptions standard for establishing the asymptotic properties of AIPW estimators.

Assumption 2.2 (Stability)**.** *For every a, there is a function $\mu_a(X)$ with $E\{\mu_a^2(X)\} < \infty$ such that, as $n \to \infty$, $\|\hat{\mu}_a - \mu_a\|_{L_2} = \{\int |\hat{\mu}_a(x) - \mu_a(x)|^2 dF_X\}^{1/2} \to 0$ in probability with respect to the randomness of $\hat{\mu}_a$ as a function of observed data, where F_X is the distribution of X.*

Assumption 2.3 (Donsker condition)**.** *For every a, there exists a class $\mathcal{F}_a$ of functions of X such that (i) $\mu_a \in \mathcal{F}_a$, where μ_a is the function in Assumption 2; (ii) $P(\hat{\mu}_a \in \mathcal{F}_a) \to 1$ as $n \to \infty$; (iii) $\int_0^1 \sup_Q \{\log N(\mathcal{F}_a, \|\cdot\|_{L_2(Q)}, s)\}^{1/2} \, ds < \infty$, where Q is any finitely supported*

probability distribution on the range of X, $N(\mathcal{F}_a, \|\cdot\|_{L_2(Q)}, s)$ *is the* s*-covering number of the metric space* $(\mathcal{F}_a, \|\cdot\|_{L_2(Q)})$*, defined as the size of the smallest collection* $\mathcal{G}_{a,s} \subset \mathcal{F}_a$ *such that every* $f \in \mathcal{F}_a$ *satisfies* $\|f - g\|_{L_2(Q)} \leq s$ *for some* $g \in \mathcal{G}_{a,s}$*, and* $\|\cdot\|_{L_2(Q)}$ *is the* L_2 *norm with respect to* Q*.*

The function μ_a in Assumption 2 can be viewed as a limit of $\hat{\mu}_a$, and it is not equal to the conditional mean $E(Y_a \mid X = x)$ when the working model is misspecified. Under Assumption 2.3, $\hat{\mu}_a$ and μ_a fall in a function class $\mathcal{F}_a$ with small complexity. Examples of models that meet this assumption include smooth parametric models of fixed dimension and nonparametric regression models with sufficiently many bounded derivatives (van der Vaart and Wellner, 1996). If $\hat{\mu}_a$ is obtained from complex machine learning methods, Assumption 2.3 may not hold (Chernozhukov et al., 2017). In Section 2.3.2, we relax Assumption 2.3 using cross-fitting.

2.3 Asymptotic Results

2.3.1 Asymptotic Linearity and Normality

The following theorem for asymptotic linearity and normality of AIPW estimator $\hat{\theta}_{\text{AIPW}}$ in (2.1) is the first established result allowing for nonparametric $\hat{\mu}_a(X)$, under CAR.

Theorem 2.1. *Suppose that Assumptions 2.1-2.3 hold.*

(i) $\hat{\theta}_{\text{AIPW}}$ *has the following asymptotic linear form, regardless of the CAR scheme used:*

$$\hat{\theta}_{\text{AIPW}} - \theta = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{I_{a,i}}{\pi_a}\left\{Y_{a,i} - \theta_a - \varepsilon_a(X_i)\right\} + \varepsilon_a(X_i),\ a = 1, ..., K\right)^T + o_P(1/\sqrt{n}) \quad (2.2)$$

where $I_{a,i}$ *is the indicator function of* $A_i = a$*,* $\varepsilon_a(X) = \mu_a(X) - E\{\mu_a(X)\}$*, and* $o_P(1/\sqrt{n})$ *denotes a term that, when multiplied by* $\sqrt{n}$*, goes* $\to 0$ *in probability as* $n \to \infty$*.*

(ii) Assume further that, almost surely,

$$\sqrt{n}\Big(\tfrac{n_a(z_l)}{n(z_l)} - \pi_a,\ a = 1, ..., K,\ l = 1, ..., L\Big)^T \mid Z_1, ..., Z_n \xrightarrow{d} N\left(0, V_\Omega\right), \tag{B}$$

where $\xrightarrow{d}$ *denotes convergence in distribution as* $n \to \infty$, $V_\Omega = \text{diag}\big[\frac{\Omega(z_l)}{P(Z=z_l)},\ l = 1, ..., L\big]$, $\Omega(z)$ *is a* $K \times K$ *covariance matrix depending on the CAR scheme, and* $\text{diag}\big[C_h,\ h = 1, ..., H\big]$ *denotes the block diagonal matrix whose* h*th block is the matrix* C_h, $h = 1, ..., H$. *Then,*

$$\sqrt{n}(\hat{\theta}_{\text{AIPW}} - \theta) \xrightarrow{d} N(0, V), \tag{2.3}$$

$$\begin{aligned} V = {} & \text{diag}\big[\pi_a^{-1}\text{Var}\left\{Y_a - \mu_a(X)\right\}, a = 1, ..., K\big] + \text{Cov}\{Y - \mu(X), \mu(X)\} \\ & + \text{Cov}\{\mu(X), Y\} - E\left[\{R_Y(Z) - R_X(Z)\}\left\{\Omega_{\text{SR}} - \Omega(Z)\right\}\{R_Y(Z) - R_X(Z)\}\right], \end{aligned}$$

where Y *is the vector of all potential responses,* $\mu(X)$ *is the vector of* $\mu_a(X)$*'s from Assumption 2.2,* $\Omega_{\text{SR}} = \text{diag}[\pi^T] - \pi\pi^T$, $\pi^T = (\pi_1, ..., \pi_K)$, $R_Y(Z) = \text{diag}\big[\pi_a^{-1}E(Y_a - \theta_a \mid Z),\ a = 1, ..., K\big]$, *and* $R_X(Z) = \text{diag}\big[\pi_a^{-1}E[\mu_a(X) - E\{\mu_a(X)\} \mid Z],\ a = 1, ..., K\big]$.

The proofs of all theorems are given in the Supplemental Material. A novel technique in proving that (I) is negligible under Assumptions 2.2 and 2.3 is that $n^{-1/2}\sum_{i=1}^n I_{a,i}[\{\hat{\mu}_a(X_i) - \mu_a(X_i)\} - E\left\{\hat{\mu}_a(X) - \mu_a(X)\right\}] = o_P(1)$ for a given a. This requires a modification to the standard empirical process result to account for the fact that A_i's are dependent under CAR. We can then show asymptotic normality by breaking up term (III) into three terms: the first one is mean zero conditioning on all treatment indicators and strata $(\mathcal{A}_n, \mathcal{Z}_n)$, the second is mean zero conditioning on all strata $\mathcal{Z}_n$, and the last is mean zero unconditionally. We show that each of these three terms converges in distribution to a multivariate normal random variable using the Lindeberg-Feller Central Limit Theorem and Condition (B). We further show that they are all asymptotically independent, which allows us to add together their variances to find the overall asymptotic variance of $\hat{\theta}$.

Besides Assumptions 2.1, 2.2, and 2.3, the asymptotic normality of $\hat{\theta}_{\text{AIPW}}$ under CAR requires condition (B) on the limiting distribution of the allocation fractions by stratum and

treatment group in randomization. Condition (B) is satisfied under simple randomization with $\Omega(z) = \Omega_{\text{SR}}$ and stratified permuted block randomization with $\Omega(z) = 0$ for all z. It is also satisfied under most commonly used CAR schemes that are implemented separately for each stratum. Pocock and Simon's minimization is a notable exception under which condition (B) is not satisfied – it assigns participants by minimizing a balance measure across the marginal levels of covariates, thereby generating correlations across different strata. Later in Section 2.3.3, we will provide a strategy for valid inference under Pocock and Simon's minimization.

The sample mean $\bar{Y}$ can be treated as a special case of AIPW with $\hat{\mu}_a = \mu_a = 0$ for all a. Thus, Theorem 2.1 tells us that, under Assumption 1 and (B), (2.3) holds for $\bar{Y}$ with

$$V = \text{diag}\left[\pi_a^{-1}\text{Var}(Y_a), a = 1, ..., K\right] - E\left[R_Y(Z)\left\{\Omega_{\text{SR}} - \Omega(Z)\right\}R_Y(Z)\right].$$

The matrix $\Omega_{\text{SR}} - \Omega(Z)$ is typically positive definite for CAR other than simple randomization. As a result, using CAR reduces the asymptotic variance of $\bar{Y}$ compared to simple randomization. This also highlights the important point that using conventional formulas for inference based on $\bar{Y}$ under simple randomization can be conservative.

2.3.2 Machine Learning and Cross-Fitting

If $\hat{\mu}_a(X)$ is constructed using machine learning methods such as random forest, then $\hat{\mu}_a$ and its limit may be too complicated to satisfy Assumption 2.3 (Donsker condition). Cross-fitting – separating the data used for fitting the machine learning models and the data used for making predictions – is a technique that allows us to bypass Assumption 2.3 (Chernozhukov et al., 2017). Specifically, let $\mathcal{I}_1, ..., \mathcal{I}_J$ be a random partition of $\{1, ..., n\}$ with a fixed integer J, where each set in the partition has nearly equal size, and let $\mathcal{I}_j^c$ be the set containing indices in $\{1, ..., n\}$ but not in $\mathcal{I}_j$. For each $j = 1, ..., J$, we use data with indices in $\mathcal{I}_j^c$ to obtain $\hat{\mu}_{a,j}(X)$ as an estimate of $E(Y_a \mid X)$. The cross-fitted estimator of θ is

$$\hat{\theta}_{\text{CF}} = \left(\frac{1}{J} \sum_{j=1}^{J} \sum_{i \in \mathcal{I}_j} \left[\frac{I_{a,i}}{n_{a,j}} \{ Y_{a,i} - \hat{\mu}_{a,j}(X_i) \} + \frac{\hat{\mu}_{a,j}(X_i)}{n_j} \right], \quad a = 1, ..., K \right)^T, \tag{2.4}$$

where $n_{a,j}$ is the number of indices in $\mathcal{I}_j$ with $A_i = a$ and n_j is the number of indices in $\mathcal{I}_j$. With cross-fitting, we use the following assumption to replace Assumptions 2.2 and 2.3.

Assumption 2.4 (Cross-Fitting)**.** *For every a, there exists a function μ_a of X with finite L_2 norm such that $P(\|\hat{\mu}_{a,j} - \mu_a\|_{L_2} \leq \delta_n) \to 1$ as $n \to \infty$, where $\hat{\mu}_{a,j}$ is given in* (2.4) *for a fixed j and $\{\delta_n\}$ is a sequence of positive numbers converging to 0.*

Assumption 2.4 is stronger than Assumption 2.2, since it requires $\|\hat{\mu}_{a,j} - \mu_a\|_{L_2} \to 0$ at rate δ_n though δ_n can be any sequence tending to 0. We have the following theorem for $\hat{\theta}_{\text{CF}}$ in (2.4) under CAR.

Theorem 2.2. *Suppose that Assumptions 2.1 and 2.4 hold. Then all results in Theorem 2.1 hold with $\hat{\theta}_{\text{AIPW}}$ replaced by $\hat{\theta}_{\text{CF}}$ defined in* (2.4)*.*

Under simple randomization, $(A_i, Y_{A_i,i}, X_i)$'s are independent and identically distributed so that Assumption 2.4 holds for machine learning methods such as random forests (Chernozhukov et al., 2017). Under CAR other than simple randomization, $(A_i, Y_{A_i,i}, X_i)$'s conditioned on A_i's and Z_i's are independent and thus Assumption 2.4 likely still holds for machine learning methods, which is also supported by empirical results from our comprehensive simulations. A formal proof is left for future research. Alternatively, we can estimate $\hat{\mu}_{a,j}$ using data in $\mathcal{I}_j^c$ stratified by levels of A and Z (Rafi, 2023). Since data within each stratum are independent and identically distributed, Assumption 2.4 is satisfied and Theorem 2.2 is applicable for this stratified $\hat{\theta}_{\text{CF}}$. The result in Rafi (2023) (Rafi, 2023) (Theorem 4.1) shows semiparametric efficiency of $\hat{\theta}_{\text{CF}}$, under the assumption that $\hat{\mu}_{a,j}(X)$ has limit $\mu_a(X) = E(Y_a \mid X)$. Our Theorem 2.2 is more general since it also covers scenarios where working models are incorrect for machine learning methods and V in (2.3) does not reach the semiparametric bound, as long as Assumption 2.4 is satisfied.

2.3.3 Guaranteed Efficiency Gain and Universal Applicability

Condition (B) is required for Theorems 2.1 and 2.2. More importantly, it requires knowing the form of $\Omega(z)$ for a particular CAR scheme and it is not satisfied for Pocock and Simon's minimization. To bypass condition (B), we establish the following theorem for $\hat{\theta}_{\text{AIPW}}$ and $\hat{\theta}_{\text{CF}}$ under a different condition (U). The result also strengthens the asymptotic normality in Theorems 2.1 and 2.2, i.e., the asymptotic distribution of $\hat{\theta}_{\text{AIPW}}$ or $\hat{\theta}_{\text{CF}}$ is invariant across CAR schemes, which is the universal applicability discussed in Section 2.1.

Theorem 2.3. *Assume Assumptions 1-3 for* $\hat{\theta}_{\text{AIPW}}$ *or Assumptions 1 and 2* for* $\hat{\theta}_{\text{CF}}$*. If*

$$E\{Y - \mu(X) \mid Z\} = \theta - E\{\mu(X)\} \quad \textit{almost surely,} \tag{U}$$

then results (B.1) *and* (2.3) *hold for* $\hat{\theta}_{\text{AIPW}}$ *or* $\hat{\theta}_{\text{CF}}$ *with*

$$V = \text{diag}\left[\pi_a^{-1}\text{Var}\left\{Y_a - \mu_a(X)\right\}, a = 1, ..., K\right] + \text{Cov}\{Y - \mu(X), \mu(X)\} + \text{Cov}\{\mu(X), Y\},$$

i.e., the asymptotic distribution of $\hat{\theta}_{\text{AIPW}}$ *or* $\hat{\theta}_{\text{CF}}$ *is invariant to CAR schemes. Furthermore, a sufficient condition for* $\hat{\theta}_{\text{AIPW}}$ *or* $\hat{\theta}_{\text{CF}}$ *to gain efficiency over* $\bar{Y}$ *under CAR is*

$$\text{Cov}\{Y - \mu(X), \mu(X)\} = 0. \tag{G}$$

Theorem 2.3 facilitates inference about $f(\theta)$ using a single and universally valid inference procedure applicable to all CAR schemes satisfying Assumption 1. Notably, this includes Pocock-Simon's minimization for which condition (B) is not satisfied. Both conditions (U) and (G) put some restrictions on the limit function $\mu(X)$. Condition (U) says that in order to have a universal variance formula, the residual vector $Y - \mu(X)$ should be mean-independent of the strata. A simple way to achieve this is to have mean-zero residuals for all strata (i.e., for all Z, $E(Y|Z) = E(\mu(X)|Z)$), which typically amounts to including all joint levels of Z and its interactions with treatment in the working model. In other words, the variables

used in CAR must be explicitly modeled at the analysis stage. On top of (U), condition (G) states that if the residual vector is uncorrelated with the vector of predicted values from the working model, covariate adjustment increases efficiency. A situation under which both (U) and (G) hold is when the working model is correctly specified, i.e., $\mu(X) = E(Y \mid X)$, and X includes Z. We now discuss two strategies to achieve (U) and (G) when the working model is incorrect.

To restore universal applicability if condition (U) does not hold, we can replace $\hat{\mu}(X) = (\hat{\mu}_1(X), ..., \hat{\mu}_K(X))^T$ with $\hat{\mu}^{(U)}(X) = \hat{\mu}(X) + \big(\sum_{j:A_j=a, Z_j=Z_i}\{Y_{aj} - \hat{\mu}_a(X_j)\}/n_a(Z_i),\ a = 1, ..., K\big)^T$. Since the limit of $\hat{\mu}^{(U)}(X)$ is $\mu(X) + E\{Y - \mu(X) \mid Z\}$, condition (U) is satisfied with $\hat{\mu}(X)$ replaced by $\hat{\mu}^{(U)}(X)$. This is called Z-calibration.

When the initial $\hat{\mu}_a$ does not satisfy (G), under simple randomization, Cohen and Fogarty (2024) (Cohen and Fogarty, 2024) proposed a linear calibration strategy that replaces $\hat{\mu}_a(X)$ with $\tilde{\gamma}_a^T \hat{\mu}(X)$, where $\tilde{\gamma}_a$ is the coefficient vector of $\hat{\mu}(X)$ from a linear regression of Y_a on "covariate" $\hat{\mu}(X)$ plus an intercept with data under treatment a. However, this strategy only works under simple randomization.

2.4 Joint Calibration

We aim to derive an estimator of θ that has both universal applicability and guaranteed efficiency gain. Two calibration techniques are introduced in the end of Section 2.3: the Z-calibration for universal applicability and linear calibration for guaranteed efficiency gain under simple randomization. It seems to suggest that we can first achieve universal applicability through Z-calibration to obtain $\hat{\mu}^{(U)}(X)$, and then, to obtain guaranteed efficiency gain, we apply a second linear calibration (Cohen and Fogarty, 2024) with $\hat{\mu}(X)$ replaced by $\hat{\mu}^{(U)}(X)$. However, this second calibration step may lose universal applicability and thus does not necessarily lead to guaranteed efficiency gain under CAR.

To simultaneously achieve both universal applicability and guaranteed efficiency gain, we propose the following joint calibration with regressor $\hat{W} = \big(Z^T, \hat{\mu}(X)^T\big)^T$ under each treatment $a = 1, ..., K$. We assume that the regressors have been appropriately adjusted to

avoid collinearity such that $\Sigma_W = \mathrm{Var}(W)$ is positive definite, where $W = \big(Z^T, \mu(X)^T\big)^T$. Define $\hat{\mu}_a^*(\hat{W}_i) = \hat{\gamma}_a^T \hat{W}_i,\ a = 1, ..., K$, where $\hat{W}_i$ is the value of $\hat{W}$ from participant i and $\hat{\gamma}_a$ is the coefficient vector of $\hat{W}$ from a linear regression of Y_a on covariate $\hat{W}$ plus an intercept with data under treatment a. The joint calibration estimator $\hat{\theta}_{\text{JC}}$ is defined as the AIPW estimator in (2.1) or the cross-fitted estimator in (2.4) with $\hat{\mu}_a(X_i)$ replaced by $\hat{\mu}_a^*(\hat{W}_i)$ for $a = 1, ..., K$. Theorem 2.4 summarizes the asymptotic properties of $\hat{\theta}_{\text{JC}}$, showing that joint calibration can simultaneously achieve universal applicability and guaranteed efficiency gain.

Theorem 2.4. *Assume Assumptions 2.1–2.3 for $\hat{\theta}_{\text{AIPW}}$ or Assumptions 2.1 and 2.4 for $\hat{\theta}_{\text{CF}}$. Then result (2.3) holds for $\hat{\theta}_{\text{JC}}$ with V replaced by $V_{\text{JC}} = \mathrm{diag}\big[\pi_a^{-1}\mathrm{Var}(Y_a - \gamma_a^T W), a = 1, ..., K\big] + \Gamma^T \Sigma_W \Gamma$, regardless of which CAR scheme is used, where $\gamma_a = \Sigma_W^{-1}\mathrm{Cov}(W, Y_a)$ and $\Gamma = (\gamma_1, ..., \gamma_K)$, and $\hat{\theta}_{\text{JC}}$ has guaranteed efficiency gain over the sample mean $\bar{Y}$ under simple randomization.*

An application of joint calibration and Theorem 2.4 is illustrated in the following example where a generalized linear model (GLM) is used as a working model.

Example 2.1. *Consider a GLM with a canonical link g as a working model for $E(Y_a \mid X)$, $a = 1, ..., K$. In this case, $\hat{\mu}_a(X) = g^{-1}(\hat{\alpha}_a + \hat{\beta}_a^T X)$, where $\hat{\alpha}_a$ and $\hat{\beta}_a$ are fitted values using the GLM and g^{-1} is the inverse of g. The specific form of g depends on the problem being considered. For example, if Y_a is binary, then $g^{-1}(t) = \mathrm{expit}(t) = \mathrm{e}^t/(1+\mathrm{e}^t)$ (logistic working model); if Y_a is non-negative integer valued, then $g^{-1}(t) = \mathrm{e}^t$ (Poisson working model); if Y_a is continuous, then $g^{-1}(t) = t$ (linear working model). Since the working model for each a is a finite-dimensional parametric model, Assumptions 2.2-2.3 are satisfied. The limit of $\hat{\mu}_a(X)$ is $\mu_a(X) = g^{-1}(\alpha_a + \beta_a^T X)$, where α_a and β_a are the limits of $\hat{\alpha}_a$ and $\hat{\beta}_a$, respectively, satisfying*

$$E\left[\{Y_a - g^{-1}(\alpha_a + \beta_a^T X)\}\binom{1}{X}\right] = 0, \qquad a = 1, ..., K. \tag{2.5}$$

The first line of (2.5) implies that $E\{\mu(X)\} = \theta$. If all joint levels of Z are included in X, then the second line of (2.5) implies that $E[\{Y_a - \mu_a(X)\}Z] = 0$ and thus $E\{Y_a - \mu_a(X) \mid$

$Z\} = 0$ for every a. Consequently, condition (U) in Theorem 2.3 is satisfied. When the working model is misspecified and g^{-1} is nonlinear, however, (G) is not satisfied. Thus, the AIPW estimator $\hat{\theta}_{\text{AIPW}}$ that uses the GLM conditional mean estimate $\hat{\mu}_a(X)$ may not have guaranteed efficiency gain over the benchmark sample mean vector $\bar{Y}$ (Guo and Basse, 2021; Cohen and Fogarty, 2024).

To achieve both guaranteed efficiency gain over $\bar{Y}$ and universal applicability, we can use $\hat{\theta}_{\text{JC}}$ by applying the joint calibration with $\hat{\mu}_a(X)$ replaced by $\hat{\mu}_a^*(X) = \hat{\gamma}_a^T \hat{W}_i$. More details for joint calibration in this example and formulas of asymptotic variances of $\hat{\theta}_{\text{AIPW}}$ and $\hat{\theta}_{\text{JC}}$ are given in the Supplementary Material.

2.5 Robust Variance Estimation

The results in Theorems 2.1–2.4 allow us to make asymptotically valid inference about θ such as constructing confidence intervals for functions of θ, provided that we can derive an estimator of V in (2.3) that is consistent and robust against misspecification of working models.

Let S_a^2 be the sample variance of $Y_{a,i}$ for participants in treatment arm a, $\hat{Q}$ be the $K \times K$ matrix whose (a,b)th component is the sample covariance of $Y_{a,i}$ and $\hat{\mu}_b(X_i)$ for participants in arm a, $\hat{\Sigma}$ be the sample covariance matrix of $(\hat{\mu}_1(X_i), \ldots, \hat{\mu}_K(X_i))^T$ based on the entire sample, $\hat{Q}_{aa}$ and $\hat{\Sigma}_{aa}$ be respectively the (a,a)th component of $\hat{Q}$ and $\hat{\Sigma}$, $\hat{R}_Y(z) = \text{diag}\big[\hat{\pi}_a^{-1}\{\bar{Y}_a(z) - \hat{\theta}_a\}, a = 1, \ldots, K\big]$, $\hat{R}_X(z) = \text{diag}\big[\hat{\pi}_a^{-1}\{\bar{\mu}_a(z) - \bar{\mu}_a\}, a = 1, \ldots, K\big]$, $\bar{Y}_a(z)$ be the sample mean of $Y_{a,i}$ for participants in treatment arm a and stratum z, $\hat{\theta}_a$ be the ath component in $\hat{\theta}_{\text{AIPW}}$ or $\hat{\theta}_{\text{CF}}$, $\bar{\mu}_a(z)$ be the sample mean of $\hat{\mu}_a(X_i)$'s for participants in stratum z, and $\bar{\mu}_a$ be the sample mean of $\hat{\mu}_a(X_i)$'s for all participants. Then, V in (2.3) can be consistently estimated by

$$\begin{aligned}\hat{V} = \text{diag}\left[\hat{\pi}_a^{-1}(S_a^2 - 2\hat{Q}_{aa} + \hat{\Sigma}_{aa}),\, a = 1, ..., K\right] + \hat{Q} + \hat{Q}^T - \hat{\Sigma} \\ - \sum_z \frac{n(z)}{n}\{\hat{R}_Y(z) - \hat{R}_X(z)\}\{\Omega_{\text{SR}} - \Omega(z)\}\{\hat{R}_Y(z) - \hat{R}_X(z)\},\end{aligned}$$

when the randomization scheme satisfies condition (B), regardless of whether the working model is misspecified or not. This estimator requires knowing $\Omega(z)$.

When condition (U) is satisfied, $\hat{V}$ can be simplified by dropping the last term, without requiring condition (B) and knowing $\Omega(z)$. We provide two alternative forms of robust variance estimators in Section B.8 of the Supplemental Materials, which may have different finite sample performance, though they are asymptotically equivalent. $\hat{V}$ above is what is used for the simulation study and case study below. When cross-fitting is used, quantities involving $\hat{\mu}_a$'s are also cross-fitted. For the joint calibration estimator $\hat{\theta}_{\text{JC}}$, V_{JC} in Theorem 2.4 can be consistently estimated using $\hat{V}$ with the last term dropped, $\hat{\mu}_a$ replaced by $\hat{\mu}_a^*$, and $\hat{\theta}_a$ replaced by the ath component of $\hat{\theta}_{\text{JC}}$. For estimation and inference about a differentiable function of θ, a robust standard error can be obtained based on $\hat{V}$ and the delta method.

2.6 Simulations

We perform a simulation study to assess and compare the finite-sample performance of all aforementioned estimators, in a two-arm trial with binary responses. We consider two cases.

Case I. $X = (X_c, X_b)^T$*, where* X_c *is uniform on* $(-5, 5)$ *and* X_b *is binary independent of* X_c *with* $P(X_b = 1) = 0.5$*. The true models for potential binary responses* Y_1 *and* Y_2 *are*

$$E(Y_1 \mid X) = \text{expit}\{0.77 + 0.5X_c + 0.5X_b - 0.2X_c^2\}$$
$$E(Y_2 \mid X) = \text{expit}\{0.2 + 0.5X_c + 0.5X_b\}.$$

We set $\pi_1 = \pi_2 = 0.5$ *using three randomization methods: simple randomization (SR), stratified permuted block randomization with a block size 6 (PB), and Pocock and Simon's minimization (Mini). The* Z *used in CAR has joint levels of* X_b *and* X_c *dichotomized at its median.*

Case II. $X = (X_{c1}, X_{c2}, X_{c3}, X_b)^T$*, where* X_{c1}, X_{c2}*, and* X_{c3} *are uniformly distributed on* $(-5, 5)$*,* X_b *is binary with* $P(X_b = 1) = 0.5$*, and* X_{c1}*,* X_{c2}*,* X_{c3}*, and* X_b *are independent.*

The true models for potential binary responses Y_1 and Y_2 are

$$\begin{aligned} E(Y_1 \mid X) &= \text{expit}\{0.77 - 0.5X_{c1} + 0.5X_{c2} + X_{c3} + 0.2X_b \\ &\qquad + X_{c1}(X_{c2} + X_{c3}) - 0.2X_{c1}^2 X_b - 0.02X_{c1}^2(1 - X_b)\}, \\ E(Y_2 \mid X) &= 1 - 0.02X_{c1}^2 - 0.02X_{c2}^2. \end{aligned}$$

We set $\pi_1 = 1/3$ and $\pi_2 = 2/3$, under the same randomization methods given in Case I. The Z used in CAR has joint levels of X_{c1} and X_{c2} dichotomized at their medians.

We consider the following estimators of the treatment contrast $\theta_2 - \theta_1$. First, the sample mean $\bar{Y}$. Second, the AIPW, linear calibration (LC), and joint calibration (JC) estimators, using the logistic model described in Example 2.1 as a working model for each treatment arm, i.e., $\hat{\mu}_a(X) = \text{expit}(\hat{\alpha}_a + \hat{\beta}_a^T X)$ for each a. In Case I, this logistic working model is correct for treatment arm 2 but incorrect for arm 1; in Case II, the working models are incorrect for both arms, with deviations more substantial than those in Case I. Third, the AIPW and JC estimators, as well as the stratified random forest approach (Rafi, 2023) using 5-fold cross-fitting and the SuperLearner R package (see Table G1 of the Supplementary Material for results using non-stratified random forests). Lastly, the AIPW estimator based on the true response model, which is an oracle estimator since the true response model is unknown in practice, is considered for benchmark purposes.

The simulation results with $n = 500$ based on 5,000 simulations are reported in Table 2.1. The results contain the simulation average of bias, simulation standard deviation (SD), simulation averages of the robust estimated standard error (SE) developed in Section 2.5, and simulation coverage probability (CP) of the 95% confidence intervals based on asymptotic normality and the SE. We also include power to reject the null hypothesis that $\theta_2 = \theta_1$ using the two different SEs. In Table G1 of the Supplementary Material, we include the SE, CP, and power columns using the naive SE obtained by treating CAR as simple randomization. In less than 0.5% of simulations in both Case I and Case II, a simulated dataset does not

have enough individuals in some strata to compute stratified $\hat{\theta}_{\text{CF}}$. These datasets are omitted from the simulation averages.

The following is a summary for the results in Table 2.1.

1. All estimators have negligible biases compared to their SDs. Our proposed robust SEs are accurate in all cases and CPs of the resulting confidence intervals are all close to the nominal level 95%. All these results support our asymptotic theory.

2. In terms of SD and power, the performance of an estimator is primarily influenced by the proximity of its working model to the true model of conditional mean response given covariates. The oracle estimator performs the best since it uses the true model. In Case II, where the working logistic model significantly deviates from the true model, estimators employing random forest outperforms estimators relying on the logistic working model. The estimators using the logistic working model and stratified random forests are comparable in Case I, where the logistic working model is either identical to or not significantly different from the true model. Although the sample mean consistently performs the worst, it can sometimes be comparable to the AIPW estimators under poor working models.

3. The use of CAR improves SD, but the effect is not as large as modeling since CAR uses dichotomized covariates. Under CAR, if universal applicability does not hold, the naive SE tends to overestimate, leading to an inflated CP (see Table G1 of the Supplementary Material). Thus, it is important to use the correct SE that accounts for CAR.

4. The JC estimator performs almost identically under simple randomization, stratified permuted block, and Pocock and Simons minimization, which confirms our asymptotic theory on its universal applicability. Although applying JC usually reduces SD, the magnitude of the reduction is not substantial when Z is already used either in AIPW

or in CAR, indicating that modeling and CAR are more useful than JC in efficiency gain.

2.7 Application to Clinical Trial Data

We analyze data from a phase 3 randomized clinical trial comparing a new basal insulin (insulin peglispro) to insulin glargine as the control in insulin-naive participants with type-2 diabetes (Davies et al., 2016). A total of 1516 participants were randomized in a 2:1 ratio to insulin peglispro or insulin glargine using permuted block randomization (with a block size of 6), stratified according to country (23 levels), baseline hemoglobin A1C (HbA1C) ($\leq$ 8.5 and $>$ 8.5 %), low-density lipoprotein cholesterol (LDL; $<$100 and $\geq$100 mg/dL), baseline sulfonylurea or meglitinide use. The endpoint we consider is an indicator of having an average of more than two hypoglycemic events per 30 days, during weeks 52-78 of follow-up. The estimand that we are interested in is the linear contrast of treatment group means.

We report the unadjusted analysis based on sample means. In addition, we consider two methods to adjust for the following covariates: age, sex, body mass index (BMI), country, baseline HbA1c, baseline LDL, baseline sulfonylurea or meglitinide use, and baseline hypoglycemic events count. As we did in the simulation section, the first method uses the logistic model described in Example 2.1 as a working model for each treatment arm. The second method fits a separate ensemble of random forests for each arm using the SuperLearner R package. Unlike our simulations, we only include non-stratified random forests because the large number of strata in this dataset renders stratified random forests infeasible. We also perform linear and joint calibration on the logistic working model, and we perform joint calibration on the random forest.

Our results in Table 2.2 show that participants receiving insulin peglispro were significantly less likely on average to experience more than two hypoglycemic events per 30 days within the 52-78 week follow-up period. All the point estimators are close to each other (with negligible differences compared to the standard errors). Covariate adjustment using the logistic working model has a 5.6% variance reduction compared to the unadjusted sample

Table 2.1: Simulation results for estimating $\theta_2 - \theta_1 = 0.123$ under case I and $\theta_2 - \theta_1 = 0.132$ under case II, based on $n = 500$ and 5,000 replicates. Entries are multiplied by 100.

						Correct		
	CAR	Working model	Method	Bias	SD	SE	CP	Power
Case I	SR		Mean	0.13	4.49	4.44	94.64	78.39
		Oracle	AIPW	0.04	3.82	3.75	94.60	89.61
		Logistic	AIPW	0.04	3.96	3.89	94.56	87.81
			LC	0.01	3.91	3.84	94.60	88.05
			JC	-0.00	3.87	3.77	94.34	88.97
		Stratified	CF	0.06	3.92	3.90	94.64	88.11
		random forest	JC	0.05	3.90	3.87	94.86	88.35
	PB		Mean	-0.01	4.01	3.98	95.14	85.87
		Oracle	AIPW	0.00	3.79	3.75	94.90	89.53
		Logistic	AIPW	-0.02	3.89	3.88	94.94	87.67
			LC	-0.03	3.83	3.81	94.92	88.63
			JC	-0.05	3.82	3.76	94.64	89.15
		Stratified	CF	-0.00	3.90	3.89	95.08	87.21
		random forest	JC	0.00	3.88	3.86	94.86	87.53
	Mini		Mean	0.18	4.00	Not available		
		Oracle	AIPW	0.16	3.74	3.75	95.04	90.42
		Logistic	AIPW	0.16	3.88	Not available		
			LC	0.13	3.82	Not available		
			JC	0.11	3.79	3.76	94.98	89.88
		Stratified	CF	0.17	3.86	Not available		
		random forest	JC	0.17	3.83	3.86	95.08	89.26
Case II	SR		Mean	-0.10	4.68	4.65	94.64	81.06
		Oracle	AIPW	-0.05	3.74	3.53	93.88	95.55
		Logistic	AIPW	-0.07	4.65	4.57	94.44	82.48
			LC	-0.07	4.66	4.57	94.32	82.46
			JC	-0.08	4.50	4.35	93.94	84.85
		Stratified	CF	-0.12	4.05	4.01	94.74	90.45
		random forest	JC	-0.11	4.05	3.99	94.72	90.65
	PB		Mean	0.09	4.41	4.41	94.85	85.79
		Oracle	AIPW	0.04	3.67	3.53	93.71	95.90
		Logistic	AIPW	0.09	4.40	4.35	94.63	86.13
			LC	0.08	4.40	4.35	94.57	86.23
			JC	0.10	4.40	4.34	94.59	86.47
		Stratified	CF	0.08	3.97	4.00	95.08	92.04
		random forest	JC	0.08	3.97	3.99	95.20	92.02
	Mini		Mean	-0.00	4.44	Not available		
		Oracle	AIPW	-0.03	3.69	3.53	93.78	95.69
		Logistic	AIPW	0.02	4.41	Not available		
			LC	0.02	4.42	Not available		
			JC	0.03	4.42	4.34	94.38	85.88
		Stratified	CF	-0.04	3.98	Not available		
		random forest	JC	-0.04	3.98	3.98	94.89	91.50

mean, which roughly corresponds to saving 5.6% sample size. This saving is amplified to 23% if the naive variance estimator is used for unadjusted sample mean, showing the importance of using correct variance estimators. The random forest method does not yield any efficiency gain over the use of logistic working model, which indicates that the logistic model is not far away from the true model.

The joint calibration enjoys universality, and it is the only method for which the naive standard errors developed under simple randomization are correct. For all other methods, the naive standard errors are too conservative because stratified permuted block randomization was applied. Note that the unadjusted sample mean is actually based on stratified permuted block randomization and is much better than the one under simple randomization, which explains why it cannot be substantially further improved by covariate adjustment in the analysis stage.

Table 2.2: Analysis results in the real data example. Entries (except for the p-values) are multiplied by 100. The naive SE and p-value use the method under simple randomization.

			Correct		Naive	
Working model	Method	Estimate	SE	p-value	SE	p-value
	Mean	-8.58	2.46	<0.001	2.73	0.002
Logistic	AIPW	-8.75	2.39	<0.001	2.50	<0.001
	LC	-8.73	2.39	<0.001	2.50	<0.001
	JC	-8.94	2.39	<0.001		
Random forest	CF	-8.51	2.45	0.001	2.70	0.002
	JC	-8.26	2.42	0.001		

2.8 Conclusion and Recommendation

From our theoretical and empirical findings, we draw the following conclusions and recommendations. When adjusting for baseline covariates, while AIPW with any misspecified working model can still lead to valid inference of the treatment effect, using a working model that closely mirrors the true model can significantly improve efficiency compared to methods that do not adjust for covariates. Machine learning methods are a suitable option when

not much is known about the functional form that governs how the response is related to the covariates. In these situations, parametric working models may be considerably wrong, and thus have adverse effects on efficiency. Joint calibration provides a simple strategy that ensures universal applicability (allowing for the use of a universal inference procedure) and provides a guaranteed efficiency gain over the unadjusted method. In all cases, it is critical to use variance estimators which are robust to model misspecification and take the covariate-adaptive randomization scheme into account.

Chapter 3

ESTIMANDS AND ESTIMATORS OF TREATMENT EFFECT IN MASTER PROTOCOL TRIALS WITH TIME-TO-EVENT OUTCOMES

3.1 Introduction

Master protocol trials can expedite and improve the efficiency of drug development by employing a shared infrastructure at the design and conduct phase, and leveraging shared controls at the analysis phase (U.S. Food and Drug Administration, 2023). Master protocol trials can be categorized as umbrella trials (multiple products for a single disease), basket trials (a single product for multiple diseases or disease sub-types), and platform trials (multiple products for a single disease that can enter and leave the platform) (Woodcock and LaVange, 2017). These three types are not exclusive and it is common for a master protocol trial to have features of multiple types. In this work, we use platform trial interchangeably with master protocol trials, for which umbrella and basket trials can be considered special cases. Platform trials have historically been utilized in oncology, including the BATTLE (Kim et al., 2011) and LungMAP (Redman et al., 2020) trials in lung cancer, the I-SPY2 trial in breast cancer (Barker et al., 2009), and the STAMPEDE trial in prostate cancer (Parker et al., 2013). Other disease areas where platform trials have been successfully employed are in infectious disease outbreak settings such as in ebola (Berry et al., 2015) and coronavirus disease 2019 (COVID-19) (The Accelerating Covid-19 Therapeutic Interventions and Vaccines (ACTIV)-6 Study Group, 2023), and neurology such as in Alzheimer's disease (Bateman et al., 2017) and amytrophic lateral sclerosis (ALS) (Paganoni et al., 2022).

Many important statistical issues arise when designing and analyzing data from platform trials. At the design stage, there is considerable interest in adapting and optimizing allocation

ratios to improve power, and better identify efficacious treatments (Ventz et al., 2018). Examples include a Bayesian phase 2 trial design to adaptively randomize individuals to the most promising and least toxic cancer immunotherapies based on accruing data (Yuan et al., 2016), using Bayesian modeling to identify whether past data from the platform can be informative for adapting allocation ratios in infectious disease outbreaks (Kaizer et al., 2018), and modifying treatment entry times into the platform in order to minimize the variance of treatment effect estimates (Bofill Roig et al., 2024). At the analysis stage, it is important to account for the use of shared control data, for example in making cross-intervention relative efficacy comparisons (Huang et al., 2023). Several of the previously cited works highlight the problem of population drift when using shared controls: since platform trials are intended to run over a long period of time, there can be considerable differences in the characteristics of participants enrolled at different stages of the trial. As such, a critical issue at the analysis stage is how to address these time trends to avoid confounding the treatment effect, when these characteristics are prognostic for the outcome. For example, can we use information from controls that were enrolled before the treatment of interest was available (Elm et al., 2012; Lee and Wason, 2020; Santacatterina et al., 2025)? Borrowing information from these so-called *non-concurrent controls* is a rich research topic, with some successful examples using Bayesian approaches such as in HEALY-ALS, a platform trial for therapeutics to slow the progression of ALS (Quintana et al., 2023). The use of non-concurrent controls is generally not recommended by regulatory agencies, primarily due to time-trends in the prognostic characteristics of the control patients, so the non-concurrent controls may not be comparable to the treated population (Koenig et al., 2023).

In this paper, we focus specifically on the use of *concurrent controls*, that is, controls that had the chance to be randomized to the treatments under consideration. This makes the treatments arm populations comparable with regards to both eligibility criteria, and, importantly, enrollment period. The choice of analytical methods for any clinical trial, including platform trials, should be driven first and foremost by the *estimand* of interest (ICH, 2019). Estimands have several attributes, including the outcome definition and handling of

intercurrent events. Two estimand attributes of notable importance in the platform trial design is the *population* and the *population summary*. Recent landmark work from Qian et al. (2025) grounds platform trials in the estimand framework, unambiguously defining the population of interest for a treatment comparison within platform trials as the *entire concurrent and eligible (ECE) population*. That is, everyone who had a chance to be randomized to the treatment groups being compared, even if they were not assigned to one of those groups. In a traditional parallel group trial, a marginal summary measure of treatment effect is typically evaluated over the entire trial population, as all individuals have positive probability to receive all treatments in the trial. However, in a platform trial, the population can be highly heterogeneous. Furthermore, the diversity of patients may change over time, as master protocol trials are often designed to be long-running enrollment platforms where treatments can enter or leave based on efficacy or futility (Webb and Derde, 2025). The population summaries considered in (Qian et al., 2025) are marginal treatment effects with respect to the ECE population, which mirrors the definition of the population in a traditional parallel group trial. Qian et al. (2025) show how to use inverse probability weighting or stratification to estimate the marginal treatment effect in the ECE population when the trial has continuous outcomes.

However, many platform trials have right-censored outcomes. In oncology platform trials, overall survival (OS) and progression-free survival (PFS) are the outcomes of interest, such as in STAMPEDE for prostate cancer, and Lung-MAP for non-small cell lung cancer. In ACTIV-6, a platform trial investigating the efficacy of repurposed medicines for the treatment of COVID-19, the primary endpoint was time to sustained recovery, or time to the first occurrence of hospitalization or death (The Accelerating Covid-19 Therapeutic Interventions and Vaccines (ACTIV)-6 Study Group, 2023). With all time-to-event outcomes, but particularly with outcomes observed on a longer time scale like OS and PFS, we may not observe outcome times for all individuals. In a traditional parallel group trial, all individuals still at risk are censored at the end of the study. Alternatively, if a participant discontinues treatment, they could withdraw study consent which precludes further follow-up. Survival

analysis methods commonly require a random censoring model that specifies variables which confer conditional independence between the outcome and censoring times. In the traditional parallel group trial design, it usually suffices to consider treatment assignment, and possibly some other variables based on the disease and treatment setting, in this variable set.

The design of platform trials can introduce new features of administrative censoring that are not present in the traditional parallel group design. In a platform design, one option is to use one long-running shared control group that is always open for enrollment even as treatment arms enter and leave the platform, as in STAMPEDE. Individuals on treatment arms could be administratively censored when those treatments show efficacy or futility when compared with the standard of care on the shared control arm. If a treatment is shown to be efficacious, perhaps it replaces the current control as the standard of care, at which point at-risk individuals receiving the old standard of care on the shared control arm may be administratively censored as they switch to the new standard of care. An alternative option is to first randomize individuals to available sub-studies for which they are eligible, and then randomize them to treatment versus standard of care within that sub-study, as in Lung-MAP. If the implementation of the standard of care across sub-studies is comparable (we will revisit this issue in the discussion), then these groups could be pooled at the analysis phase, to mimic the shared control group of the previous design. However, the censoring mechanisms in this sub-study design may differ from the shared control design, as entire sub-studies, including the control group, would be administratively censored when the sub-study is terminated. Which variables, then, confer a random censoring model? How can we use survival analysis methods to target the ECE population and address both confounding dependent censoring?

In our work paper, we address these questions and fill an important gap that bridges platform trials, estimands, and survival analysis. In Section 3.2, we define the ECE population in our setting and introduce an appropriate and clinically relevant summary measure for time-to-event outcomes based on this population. We take care to enumerate the assumptions necessary for identification of this estimand, including which follow-up times are valid

to consider for a given treatment comparison, and what variables may induce dependent censoring. In Section 3.3, we walk through how our framework applies to both stylistic and real examples of platform trials. In Section 3.4, we present estimators that are consistent and asymptotically normal for the treatment effects outlined in Section 3.2. In Section 3.5, we provide a simulation study that justifies our proposed methods.

3.2 Estimand

We now describe the general framework outlined in Qian et al. (2025) and highlight key issues for time-to-event estimands. Consider a platform trial with treatment assignment A taking on $j = 1, ..., J$ treatment groups within the trial. Let X be a vector of observed baseline variables and U be a vector of unobserved baseline covariates. Let $T^{(j)}$ be the potential failure time and $C^{(j)}$ be the potential censoring time under treatment j. Let $Y^{(j)} = \min(T^{(j)}, C^{(j)})$ and $\delta^{(j)} = I(T^{(j)} \leq C^{(j)})$ be the observed follow-up time and event indicator. Let $Z \subseteq X$ be a vector of observed variables that determine treatment assignment probabilities. These may include eligibility to different sub-studies, disease sub-type, participant preference, and enrollment window. We denote S as the sub-study assignment for trials that utilize a sub-study randomization structure. For full generality in what follows, we will always refer to the sub-study structure. However, for studies that do not have a sub-study structure, we can consider the variable S to be a constant value.

Assumption 3.1 (Randomization)**.** *There exists a sub-vector Z of X such that*

$$A \perp (T^{(1)}, ..., T^{(J)}, C^{(1)}, ..., C^{(j)}, X, U)|Z$$

and that $P(A = j|Z) = \pi_j(Z)$, where $\pi_j(Z)$ are known (but random) probabilities with $\sum_{j=1}^{J} \pi_j(Z) = 1$.

Assumption 3.1 presents the design of the platform trial as it relates to randomization. The function $\pi_j(z) : z \to P(A = j|Z = z)$ captures how randomization probabilities differ

by participant characteristics. For most platform trials, the vector Z will include someone's window of enrollment. The windows of enrollment are defined as periods within which randomization probabilities across the platform are constant. As a simple example, consider a platform trial with two arms initially open, and one additional arm open one year later. This platform trial has two windows of enrollment: one during the first year when there are two arms available, and one after one year when there are three arms available. Randomization probabilities may also depend on other participant-level characteristics, like having a particular biomarker, or having advanced disease. These variables would also be included in Z. We will discuss specific examples of platform trials and what constitutes their Z in Section 3.3.2. Importantly, the value of $\pi_j(Z)$ may be zero, which indicates that someone with variable(s) Z was not able to be randomized to treatment j, either because it was not available when they enrolled, or because they did not meet the eligibility criteria for treatment j. The full data vector is given by $(A, Y^{(1)}, \ldots, Y^{(J)}, \delta^{(1)}, \ldots, \delta^{(J)}, X, U, S)$.

3.2.1 Population of Interest

Let $\mathcal{J}$ be a set of treatments to compare (e.g., $\mathcal{J} = \{1, 2\}$ is for comparisons of treatments 1 and 2). In a traditional trial with two treatment arms, this comparison is straightforward, and the population of interest is typically defined as the entire trial population. In the platform trial setting, Qian et al. (2025) introduces the *Entire Concurrently Eligible (ECE)* population, which is the population of individuals within the trial who had the chance to be randomized to all of the treatments contained within set $\mathcal{J}$. This accounts for treatment arms within the platform having different windows of enrollment, and different eligibility criteria. In a traditional parallel group trial, the ECE population is the entire trial population.

Definition 3.1 (Entire Concurrently Eligible Population)**.** *The Entire Concurrently Eligible (ECE) population for a given treatment comparison set $\mathcal{J}$ is the population of individuals for which $\pi_j(Z) > 0$ for all $j \in \mathcal{J}$.*

The ECE population makes unambiguous the population attribute of the estimand: all

concurrent and eligible individuals for the treatments being compared in $\mathcal{J}$. This is a natural place to start because this is the corresponding definition of the population in a traditional parallel group trial. However, there could be other populations of interest for particular clinical questions. We do not discuss those alternative population definitions further and focus on the specific ECE population.

3.2.2 Summary Measure

Our estimand is the set of arm-specific survival functions in the ECE population, for each $j \in \mathcal{J}$:

$$S_j(t) := P(T^{(j)} > t | \pi_k(Z) > 0, k \in \mathcal{J}) \equiv P_{\mathcal{J}}(T^{(j)} > t); \quad j \in \mathcal{J} \tag{3.1}$$

where we use the notation $P_{\mathcal{J}}(\cdot) = P(\cdot|\pi_k(Z) > 0, k \in \mathcal{J})$ to denote that the probability is conditional on membership to the ECE population. When referencing (3.1) in the text, we often suppress the $\mathcal{J}$ notation for conciseness; throughout it is assumed that we are estimating survival functions with a particular ECE population defined by $\mathcal{J}$ in mind. This is similar to the definition of the Treatment Effect in Qian et al. (2025), with a key difference: in the time-to-event setting, we must place restrictions on which times t we can include in our estimand in (3.1).

3.2.3 Censoring Assumption

Assumption 3.2 (Random Censoring in the ECE Population)**.** *Let $I_{\mathcal{J}}(Z) = I(\pi_k(Z) > 0, k \in \mathcal{J})$ indicate membership to the ECE population. Then we assume that there is an observed sub-vector V of X that, for all $j \in \mathcal{J}$, among individuals in the ECE population:*

$$T^{(j)} \perp C^{(j)} | V, I_{\mathcal{J}}(Z) = 1. \tag{i}$$

If V is non-empty, then we also assume that:

$$A \perp (T^{(j)}, C^{(j)}) | V, I_{\mathcal{J}}(Z) = 1. \tag{ii}$$

Assumption 3.2 (i) states that, among individuals eligible and concurrent for randomization to arms within $\mathcal{J}$, the potential failure and censoring times are independent given V, which is referred to as the random censoring model. If V is empty, this is an independent censoring assumption, i.e., the potential failure and censoring times are independent. If V is non-empty, then we require an additional condition in (ii): that there is no confounding by treatment within levels of V. In practice, this will mean that if censoring is not independent (i.e., non-empty V), then we need to enrich V with all of the variables also in Z in order to satisfy Assumption 3.2 (ii). Since $V \subseteq X$, $V = (Q, Z)$, where Q is a subset of variables within X but not Z. Thus, $A \perp V | Z$ is implied by Assumption 3.1. The reason that we require (ii) is because we will need to estimate the censoring distribution conditional on (A, V) in a separate, (stratified) Cox proportional hazards model, which we detail in Assumption 3.3. Assumption (i) and (ii) together imply the more commonly-used form of random censoring, which is stated in terms of the observed failure and censoring times (Robins and Rotnitzky, 1992; Schaubel and Wei, 2011): $(T \perp C) | A, V, I_{\mathcal{J}}(Z) = 1$. We assume the following model for censoring in Assumption 3.3.

Assumption 3.3 (Model for Censoring)**.** *Let $V^P, V^S \subseteq V$, with $V^P \cup V^S = V$ and $V^P \cap V^S = \emptyset$. In other words, V^S and V^P are mutually exclusive, but together they contain all of the variables in V, the variable defined in Assumption 3.2. Assume the following stratified Cox proportional hazards model for the potential censorsing time $C^{(j)}$:*

$$\lambda_j^C \left(t | V^P, V^S, I_{\mathcal{J}}(Z) = 1\right) = \lambda_{j,0}^C \left(t | V^S, I_{\mathcal{J}}(Z) = 1\right) \exp\left(\beta^T V^P\right).$$

When $V^P = \emptyset$, the model for censoring is non-parametric since the baseline hazard is stratified by $V^S = V$, all variables in the random censoring model. Otherwise, it is semi-

parametric, as we make the proportional hazards assumption using covariates $V^P \subseteq V$ from the random censorship model.

3.2.4 Valid Follow-Up Times

Definition 3.2 outlines which times can be considered as valid follow-up times for the estimand in (3.1). That is, what times t are appropriate to compare $S_j(t)$ across $j \in \mathcal{J}$?

Definition 3.2 (Valid Follow-Up Times)**.** *Consider times t such that, for a V given in Assumption 3.2 and satisfying the model in 3.3, the following conditions hold:*

$$P_{\mathcal{J}}(\min(T^{(j)}, C^{(j)}) > t | V^S = v) > 0 \quad \forall v = 1, \dots S \tag{A}$$

$$P_{\mathcal{J}}(C^{(j)} > t | V) > 0 \ \textit{w.p.1} \tag{B}$$

where S is the number of joint levels of V^S from Assumption 3.3, and where w.p.1 indicates "with probability 1", taken over the support of V in the population $\mathcal{J}$. Define the function $\tau_j : V \to \max\{t \geq 0 : t$ satisfies Conditions (A) and (B) with $V\}$. Then the valid follow-up times in the ECE population are $t \in [0, \tau_j(V)]$.

If V is empty, Definition 3.2 is a standard assumption that is required to be able to identify the survival functions, and trivially, Condition (A) implies Condition (B). If V is non-empty, then to employ estimation strategies that use inverse probability of censoring weighting based on a model in Assumption 3.3, (A) and (B) are required: (A) ensures that the baseline hazard function for censoring can be estimated within each strata level v, and (B) ensures that the predicted probability of censoring is bounded away from zero, so that there is no division by zero (Robins and Rotnitzky, 1992). If $V^P = \emptyset$, then Condition (A) implies Condition (B). Since V will include Z in the case where there is *not* independent censoring, and since Z will typically include enrollment window, this means that the maximum follow-up time will typically be no longer than the follow-up time for the window most recently enrolled in the trial and eligible for treatments in the set $\mathcal{J}$. The valid follow-up times in

Definition 3.2 will in most cases be the longest potential follow-up across the estimation strategies that we present in Section 3.4. For post-stratification (PS) estimators, there is a different condition on the maximum follow-up time, that will usually be more restrictive than (A). Specifically, we need (A) to hold conditional on the joint levels of (H, V), where H is a variable that contains all of the information about π_j, even if V is empty. We provide more details in Section 3.4.3.

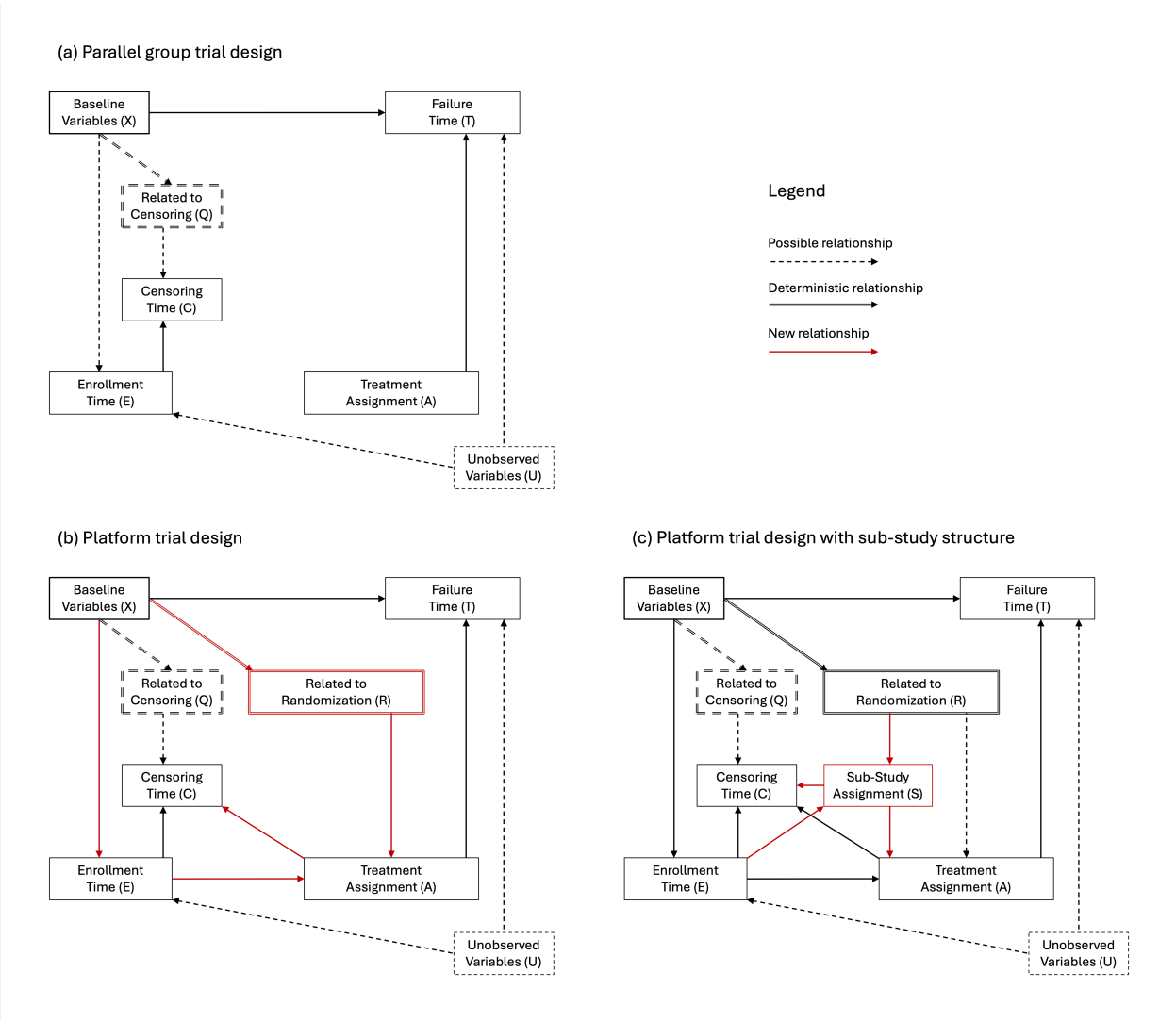
3.3 Sufficient Adjustment Sets

Figure 3.1 shows directed acyclic graphs (DAGs) that depict the variables to consider in the master protocol framework. Using a DAG allows us to be explicit with our assumptions, and construct sufficient sets Z and V to satisfy Assumptions 3.1 and 3.2. In this section, we do this exercise for both stylistic and real examples. We first present the stylistic examples, which illustrate how even in very simple settings, these issues require careful attention. We then walk through applications to real examples which are more complex, including the STAMPEDE, Lung-MAP, and ACTIV-6 platform trials.

3.3.1 Stylistic Examples

In Figure 3.2, we present four different types of master protocol designs that could be conducted to investigate the efficacy of two different treatments compared to placebo. Our comparison of interest for this exercise is treatment 1 compared to placebo. Designs (a) and (b) use a shared or common control group, whereas Designs (c) and (d) use a sub-study structure where the control group for each sub-study receives the same placebo treatment. In Designs (a) and (c), treatments 1 and 2 have the same enrollment period, whereas in Designs (b) and (c), they only have overlapping enrollment. Notably, these four examples each have different designs but they have the same ECE population for the treatment 1 to placebo comparison (except in the case where treatment 1 has more restrictive eligibility criteria than treatment 2). Two study designs which we did not consider, but could have been included in Table 3.1 are where treatment 1 and 2 have mutually exclusive enrollment

Figure 3.1: Directed acyclic graph that illustrates the possible relationships between variables in a parallel group trial design (a), a platform trial design (b), and a platform trial design with a sub-study structure (c).



windows, i.e., where treatment 1 enrolls during window 1 and treatment 2 enrolls during window 2. In these cases, the ECE population is only those in enrollment window 1 because those individuals are the only ones that were eligible for enrollment to treatment 1 and placebo. Since the enrollment window is restricted, these designs have similar considerations to Designs (a) and (c).

For each of these designs, we additionally consider four settings related to a prognostic binary variable R. In Setting 1, the prognostic variable is not used for treatment assignments. In Setting 2, we allow for treatment assignment propensities (in Designs (a) and (b)) or sub-study assignment propensities (in Designs (c) and (d)) to be a function R, which, at its extreme, could include treatment 2 being restricted to only to those with $R = 1$. Because everyone still has the potential to receive treatment 1 or placebo, the ECE population does not change in this setting. Setting 3 builds on Setting 2 by allowing the distribution of $R = 1$ versus $R = 0$ in enrollment windows 1 and 2 to differ, i.e., there is a temporal drift in R among participants enrolled. Finally, Setting 4 includes the example of restricting eligibility to treatment 1 to those with $R = 1$.

Figure 3.2: Four possible study designs (a) - (d) for a simple trial where two treatments are compared against one placebo group. Treatment 2 is stopped for futility one year after enrollment is complete. The DAGs to the left and right show the relationships between the variables for each specific design, as in Figure 3.1. Dashed lines indicate possible relationships, which are discussed in the text, and in the specific Settings 1 through 4 in Table 3.1.

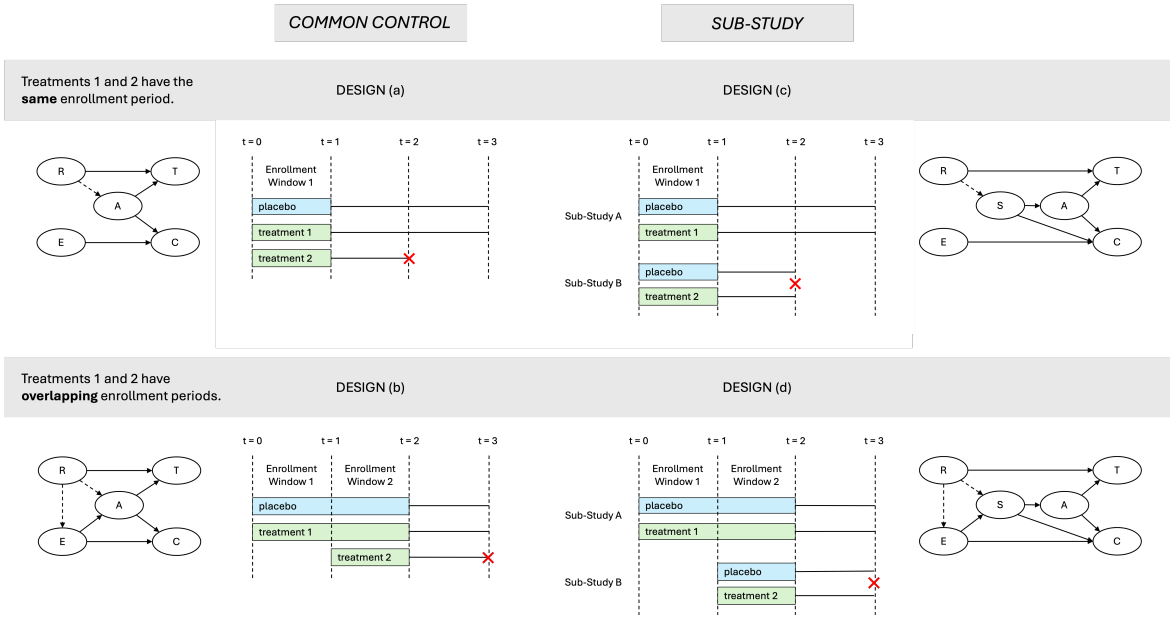


Table 3.1: Characteristics of Designs (a) through (d) depicted in Figure 3.2, for four different settings relating to a binary prognostic variable R. We focus on the specific case of comparing treatment 1 to placebo (pbo.) to define the ECE population, and on estimating survival under placebo for this population. Z: variables defining randomization probabilities; V: sufficient adjustment sets for censoring; SS: Sub-study. "–" indicates no variable is needed.

Design	ECE Pop.	Z	V	Valid Times	Pbo. Analysis Set	Pbo. Risk Set
Setting 1: R does not affect treatment/sub-study probabilities						
(a)	All	–	–	0-3	Assigned to pbo.	0-3: Pbo.
(b)	All	E	–	0-3	Assigned to pbo.	0-3: Pbo.
(c)	All	–	–	0-3	Assigned to pbo.	0-2: Pbo., 2-3: SS-A pbo.
(d)	All	E	–	0-3	Assigned to pbo.	0-2: Pbo., 2-3: SS-A pbo.
Setting 2: Treatment or SS assignments differ by R, or Treatment 2 is only available to those with R=1						
(a)	All	R	–	0-3	Assigned to pbo.	0-3: Pbo.
(b)	All	R, E	–	0-3	Assigned to pbo.	0-3: Pbo.
(c)	All	R	R	0-3	Assigned to pbo.	0-2: Pbo., 2-3: SS-A pbo.
(d)	All	R, E	R, E	0-2	Assigned to pbo.	0-2: Pbo.
Setting 3: Setting 2, plus drift in R by enrollment window[1]						
(a)	All	R	–	0-3	Assigned to pbo.	0-3: Pbo.
(b)	All	R, E	R, E	0-2	Assigned to pbo.	0-2: Pbo.
(c)	All	R	R	0-3	Assigned to pbo.	0-2: Pbo., 2-3: SS-A pbo.
(d)	All	R, E	R, E	0-2	Assigned to pbo.	0-2: Pbo.
Setting 4: Treatment 1 is only available to R=1, possibly with drift in R by enrollment window						
(a)	R=1	–	–	0-3	R=1 assigned to pbo.	0-3: R=1 pbo.
(b)	R=1	E	–	0-3	R=1 assigned to pbo.	0-3: R=1 pbo.
(c)	R=1	–	–	0-3	R=1 assigned to pbo.	0-2: R=1 pbo., 2-3: SS-A R=1 pbo.
(d)	R=1	E	–	0-3	R=1 assigned to pbo.	0-2: R=1 pbo., 2-3: SS-A R=1 pbo.

[1] Since the enrollment in (a) and (c) is only during one window, the table entries for Setting 3 are the same as Setting 2 for Designs (a) and (c).

Table 3.1 lists pertinent pieces of information to define and identify the estimand in the different Designs and Settings (we will discuss the non-estimand columns of Table 3.1 in Section 3.4). The four trials are quite simple in their structure, which means that the ECE population in Designs (a) - (d), and Settings 1 through 3, are the entire trial population, as everyone enrolled in the trial is eligible for treatment 1 and placebo, regardless of enrollment window or prognostic marker R. Setting 4 is slightly different in that treatment 1 has restricted eligibility criteria to only those with $R = 1$, which explicitly defines the ECE population to only those with $R = 1$, for all Designs (a) - (d).

Even though the ECE population in Settings 1 - 3 and Designs (a) - (d) are the same, the specific Design and Setting combinations necessitate different adjustment variables Z and V to identify the estimand in that population. Note that for these simple examples, we only consider the necessary V that results from the platform design. There could be other baseline covariates which jointly influence censoring and failure times, which in practice would be included in V as well. The randomization probabilities could differ by enrollment window (E) and/or prognostic variable (R). Furthermore, R and/or E may also be needed in order to block pathways of dependent censoring. Next to each study schema in Figure 3.2, we include a DAG similar to Figure 3.1, but tailored to the specific Designs, and Setting (through the dashed arrows). First, consider Setting 2 comparing Designs (a) and (b) to Designs (c) and (d). In Designs (a) and (b), there is no pathway for dependent censoring between failure time T and censoring time C once we have conditioned on treatment assignment A. However, in Designs (c) and (d), the initial randomization to sub-study, and then to treatment arm, opens up a pathway between $T \leftarrow R \rightarrow S \rightarrow C$ if sub-studies have different administrative censoring. In these designs, the variable V must include R, because it cannot include sub-study indicator which is a post-randomization variable that does not satisfy Assumption 3.2. Setting 4 provides the interesting, and reasonable case where treatment 1 has restricted enrollment criteria relative to treatment 2. Since our censoring assumption is conditional on the ECE population – which in this case is only those with $R = 1$ – there is no dependent censoring in any of the Designs (a) - (d).

3.3.2 Real Examples

In this section, we will walk through increasingly complicated DAGs and motivate these increasing complexities using real trials. The three trials that we will use as examples are STAMPEDE, Lung-MAP, and ACTIV-6; we will describe them here briefly before using them to illustrate our approach. STAMPEDE was a multi-arm, multi-stage platform trial evaluating treatment additions to hormone therapy for men with high-risk, locally advanced, metastatic, or recurrent prostate cancer (Parker et al., 2013). Lung-MAP was a biomarker-driven sub-study designed platform trial to evaluate therapies for advanced or relapsed squamous non-small cell lung cancer (sqNSCLC) (Redman et al., 2020). ACTIV-6 was a decentralized platform trial evaluating the efficacy of COVID-19 treatment regimes in the outpatient setting (The Accelerating Covid-19 Therapeutic Interventions and Vaccines (ACTIV)-6 Study Group, 2023).

In a traditional parallel group trial design, shown in Figure 3.1(a), we are primarily interested in how the treatment assignment influences failure time ($A \to T$). Individuals will have heterogeneous and prognostic baseline variables X, but those variables are unrelated to treatment assignment ($X \not\to A$). Some of those variables X may be related to when an individual chooses to enroll in the trial E ($X \to E$), but enrollment is unrelated to treatment assignment since there are a fixed number of treatments under consideration ($E \not\to A$). Administrative censoring is indicated by the arrow between E and censoring time C: everyone's calendar time at censoring is the same, but their time until censoring is a function of enrollment calendar time. Other baseline variables Q may impact censoring as well ($Q \to C$). In addition to observed baseline variables X influencing both enrollment time and failure times, there may be similar, but unobserved baseline variables U. However, we make the assumption that unobserved variables do not directly influence censoring ($U \not\to C$), which is standard in survival analysis.

The STAMPEDE trial motivates the changes needed when moving from a parallel group trial design to a platform trial design in Figure 3.1(b). In STAMPEDE, there are multiple

treatments under consideration with one shared control group, and distinct enrollment windows that determine which of the treatment regimens are available to participants ($E \to A$) (Sydes et al., 2009, 2012). The primary outcome is overall survival (T). In this population, a type of hormone therapy called androgen deprivation therapy (ADT) was the standard of care (SoC), and this was designated for the placebo group. Initially, five experimental arms received treatments on top of ADT with a fixed allocation ratio of 2:1:1:1:1:1 so that the ADT arm had two times as many individuals as each of the following: (1) zoledronic acid, (2) docetaxel, (3) celecoxib, (4) zoledronic acid + docetaxel, and (5) celecoxib + zoledronic acid. Over a period of ten years, another six experimental treatments were added in separate arms (Schiavone et al., 2019). One of these was ADT + radiotherapy, which was only available to men with metastatic disease (Parker et al., 2013), which makes metastatic disease a key prognostic variable that influences randomization ratios ($R \to A$ and $R \to T$). For this example, the treatment propensities are determined by $Z = (R, E)$ with R being metastatic versus localized disease. To account for dependent censoring, we need $V = (R, E, Q)$, where Q are the prognostic variables that directly influence censoring.

A second type of platform trial is shown in Figure 1(c). In the sub-study structure, individuals are first randomized to a sub-study that they are eligible for, and then to the active arm versus the placebo within that sub-study. Lung-MAP exemplifies this structure: individuals with sqNSCLC are first randomized to the sub-studie(s) testing targeted therapies for the specific tumor biomarker that they have. There is an additional sub-study for individuals for which there is no active sub-study that matches their biomarker. Individuals with multiple matching biomarkers are randomized to relevant sub-studies weighted by the relative prevalence of those biomarkers, favoring biomarkers with lower prevalence (to ensure sufficient accrual for the sub-studies with lower prevalence biomarkers). Sub-study assignment can also influence censoring time because each individual sub-study may be ended administratively at different times based on completing recruitment targets, or stopping for efficacy or futility. In this example, $Z = (E, R)$, where R is the biomarker(s) that characterize the tumor,a nd $V = (R, E, Q)$ again, which accounts for unobserved prognostic variables

influencing E.

The ACTIV-6 trial also has a sub-study structure. In ACTIV-6, the repurposed drugs under investigation over the course of the study at some point included Ivermectin 400 and 600, Fluticasone, Fluvoxamine 50 and 100, Montelukast, and Metformin. Each drug had a window during which it was available in general for the study. The study was decentralized, with over 100 participating study sites, and there was a high level of complexity in the local review and approval process for a drug to be available at a specific site. This led to, within the study-wide window of availability, variable site-specific windows of enrollment for a particular drug. Additionally, individuals could choose which of the available drugs at their study site they were comfortable taking. For each individual, it is known how many sub-studies they were eligible for based on site availability and personal preference, and they were randomized with equal probability to these sub-studies. The variable dictating the randomization probabilities $Z = (N)$, where N is the number of sub-studies for which the individual was eligible and consented. In ACTIV-6, administrative censoring acted at the site level, rather than the sub-study level. As such, geographic site is also included in Q, the variable that influences censoring. It is also likely a proxy for important prognostic variables that could influence both enrollment time and failure time like local epidemiology and circulating SARS-CoV-2 variants. Therefore, it suffices to consider $V = (N, E, Q)$.

3.4 Estimation

Let i be the index for the ith person. We observe n independent and identically distributed copies of the full data vector $(A_i, Y_i^{(1)}, \ldots, Y_i^{(J)}, \delta_i^{(1)}, \ldots, \delta_i^{(J)}, X_i, U_i, S_i)$, where n is the sample size for the entire platform trial. Let $Y_i = \sum_{j=1}^{J} I(A_i = j) Y_i^{(j)}$ and $\delta_i = \sum_{j=1}^{J} I(A_i = j)\delta_i^{(j)}$ be the observed follow-up time and event indicator. Define the set of individuals $\mathcal{I}_{\mathcal{J}} = \{i : \pi_k(Z_i) > 0, k \in \mathcal{J}\}$ to be those who are random samples from the ECE population. That is, they had a non-zero probability of being randomized to all of the treatments in the comparison set $\mathcal{J}$. Let $\mathcal{I}_{\mathcal{J}}(j) = \{i \in \mathcal{I}_{\mathcal{J}} : I(A_i = j) = 1\}$ be the subset of those individuals assigned to treatment arm j. For a generic set $\mathcal{I}$, let $n(\mathcal{I}) = |\mathcal{I}|$ be the number

of individuals included in that set. For a random variable X_i and its possible value x, let $n(x) : x \to \sum_{i=1}^{n} I(X_i = x)$ and $n_j(x) = x \to \sum_{i=1}^{n} I(X_i = x)I(A_i = j)$.

3.4.1 Analysis and Risk Sets

In Section 3.3, we thoroughly discussed different study designs and scenarios and how they impact the variables that constitute sufficient adjustment sets. In subsequent sections, we will discuss different estimation strategies for (3.1) based on the sufficient adjustment sets of Z and V. Now, for our proposed estimators, we need to define an analysis set: that is, which individuals contribute to the estimation? This is a similar discussion as in Qian et al. (2025). For time-to-event outcomes, another important concept is the composition of the risk set at different follow-up times.

For this discussion, we return to the stylistic examples in Figure 3.2 and Table 3.1 that we discussed in Section 3.3.1, and we ask: who should be included in the analysis set and the risk set in order to estimate survival under *placebo*? As we previously discussed, Settings 1, 2, and 3 all have the same ECE population: the whole trial population. Therefore, the analysis set is every individual. The placebo analysis set is individuals in the analysis set and assigned to placebo. These are people whose outcome information is used; other people's information on X can still be used, as we will see for the stratified estimators in Section 3.4.3. Importantly, the estimand is *not* survival among those assigned to placebo, because this is not the ECE population. At the estimation stage, each individual in the placebo analysis set (Table 3.1) will be re-weighted according to their treatment propensities defined by the corresponding Z variable for that Design and Setting combination. This ensures that the proposed estimators in Section 3.4 target the entire trial population. Setting 4 is analogous: the ECE population is those with $R = 1$, and so the placebo analysis set consists of participants with $R = 1$ who were assigned to the placebo group(s).

The analysis set is the group of individuals that are used in estimation for a particular estimand. The risk set is a function of the time since enrollment. In Designs (a) and (b), this is straightforward, because there is only one placebo group. So, the risk set at any time

t is all individuals assigned to placebo who have not had the event or been censored by time t. The sub-study structure in Designs (c) and (d) complicates this. The risk set at any time $t \in [0, 2)$ consists of all individuals who were assigned to placebo and who have survived until time t. But at two time-units of follow-up, all individuals who were assigned to the placebo group in sub-study B were administratively censored. So the risk set at any time $t \in [2, 3)$ consists only of individuals in sub-study A who have not had the event or been censored by time t. The reason that this does not cause an issue is because of Assumption 3.2. Under the censoring at random assumption, after conditioning on V, those who remain in the risk set have the same event time distribution as those who have potentially survived, regardless of their censoring time.

3.4.2 Naive Estimators

To estimate $S_j(t)$ for a $j \in \mathcal{J}$, the first naive method is to use the Kaplan-Meier estimator to account for right-censoring:

$$\widehat{S}_j^{\text{KM}}(t) := \prod_{t_k \leq t} \left[1 - \frac{\sum_{i:Y_i=t_k} \delta_i I(A_i = j)}{\sum_{i:Y_i \geq t_k} I(A_i = j)} \right] \tag{3.2}$$

where $t_1 < t_2 < \ldots$ are the unique observed failure times for treatment group j, which we refer to as t_k. This estimator uses all individuals *assigned* to treatment group j to draw conclusions about the potential survival experience under treatment j. As another naive method, we can subset the observations to those who were both *assigned* to treatment group j, and *eligible* for randomization to all other arms in the comparison set $\mathcal{J}$:

$$\widehat{S}_j^{\text{KMCE}}(t) := \prod_{t_k \leq t} \left[1 - \frac{\sum_{i \in \mathcal{I}_\mathcal{J}:Y_i=t_k} \delta_i I(A_i = j)}{\sum_{i \in \mathcal{I}_\mathcal{J}:Y_i \geq t_k} I(A_i = j)} \right]. \tag{3.3}$$

This approach ensures that the individuals included are concurrent and eligible for all arms in the comparison set, which is in line with FDA guidance. However, the unique structure of platform trials results in constrained and non-uniform treatment assignment, i.e., some

participants have zero probability to receive certain treatments, and the non-zero treatment assignment probabilities can vary based on participant characteristics, enrollment time, and treatment availability. As a result, the estimator in (3.3) does not necessarily target the survival function estimand in the ECE population. Also, it cannot directly handle the dependent censoring described in Section 3.3.2 that arises as a result of the platform trial design. Instead, it can be shown that $\widehat{S}_j^{\text{KMCE}}(t)$ is consistent for the survival function in a different population, which we will denote as the "assigned, concurrently eligible" population, or ACE. Contrasting this with the ECE population estimand of $P_{\mathcal{J}}(T^{(j)} > t)$, $\widehat{S}_j^{\text{KMCE}}(t)$ is consistent instead for $P_{\mathcal{J}}(T^{(j)} > t | A = j)$. That is, it is the potential survival function among those who are concurrently eligible for randomization to all treatments under consideration in the set $\mathcal{J}$, *and* assigned to treatment arm j. It is possible that the potential survival function in the ACE and ECE populations is equivalent. In order for a treatment contrast to then be equivalent, the ACE and ECE populations would need to match up for each treatment being compared in the set $\mathcal{J}$. A key advantage of the ECE population is that it is a characteristic of the set $\mathcal{J}$, whereas the ACE population depends implicitly on randomization ratios.

3.4.3 Proposed Estimators for the ECE Population Summary

We now turn our attention to our proposed estimators which are consistent for the estimand in (3.1). It will be convenient to work with a new variable H, which we introduce in Definition 3.3. The variable H may be more or less granular than Z, but it contains all of the treatment propensity information from Z.

Definition 3.3. *Let H be a discrete random variable that deterministically maps $Z \to H(Z)$ onto $L < \infty$ levels that confer the same treatment propensity as Z, i.e.,*

$$\pi_j(H) = \pi_j(Z) \quad \forall j \in \mathcal{J}.$$

Let H_i be the realization of H for person i. Consider the asymptotic regime where, for each level h of H, $n(h)/n$ converges in probability to come constant $c > 0$. In other words, the

number of participants within each strata grows at the same rate as the sample size.

First, we consider a general form of an *inverse probability-weighted* (IPW) estimator extension to the Kaplan-Meier estimator, tailored to the ECE population, which is given by

$$\widehat{S}_j^{\text{IPW}}(t) := \prod_{t_k \leq t} \left[1 - \frac{\sum_{i \in \mathcal{I}_{\mathcal{J}}: Y_i = t_k} \frac{\delta_i I(A_i = j)}{w_{ij}(t_k)}}{\sum_{i \in \mathcal{I}_{\mathcal{J}}: Y_i \geq t_k} \frac{I(A_i = j)}{w_{ij}(t_k)}} \right]. \tag{3.4}$$

An important difference between (3.2) and (3.4) is that (3.4) only uses individuals that are random samples from the ECE population, whereas the entire population assigned to treatment j is used in (3.2). The other key difference is the inclusion of weights $w_{ij}(t)$. There are several forms that the weights $w_{ij}(t)$ can take. The simplest IPW estimator is (3.4) using $w_{ij}(t) = \pi_j(Z_i)$. These weights are constant over time, which has the form of the adjusted Kaplan-Meier estimator (AKME) introduced in Xie and Liu (2005). The treatment propensities based on variables Z_i are known by the design of the platform trial (Assumption 3.1), so we can directly use their known value.

It is generally more efficient to estimate treatment propensities rather than use their known value. A second form of the weights that we can consider is $w_{ij} = \widehat{\pi}_j(H_i) := n_j(H_i)/n(H_i)$, Constructing the weights with a discrete H from Definition 3.3 in this way allows us to do IPW adjustment with estimated propensities completely non-parametrically. In practice, randomization probabilities are almost always based on Z as a discrete variable; and if they are not, then collapsing the resulting probabilities of treatment into the variable H suffices for non-parametric IPW adjustment. One obvious choice for H if Z is discrete is to choose H to be the joint levels of Z. Alternatively, we could set H to be the values of the allocation proportions. Let $\Pi_{\mathcal{J}}(Z) = (\pi_j(Z) : j \in \mathcal{J})$ be a random vector of the treatment arm assignment probabilities for a given comparison; i.e., if $\mathcal{J} = \{1, 2\}$, then $\Pi_{\mathcal{J}}(Z) = (\pi_1(Z), \pi_2(Z))$. Then $H = \Pi_{\mathcal{J}}(Z)$ also satisfies Definition 3.3.

Under dependent censoring (Assumption 3.2 with non-empty V), the IPW estimator lacks consistency. Setting $w_{ij}(t) = \pi_j(Z_i)\widehat{p}_j^C(t|V_i)$ or $w_{ij}(t) = \widehat{\pi}_j(H_i)\widehat{p}_j^C(t|V_i)$ combines in-

verse probability weighting with inverse probability of censoring weighting, with $\widehat{p}_j^C(t|V_i)$ the estimated probability of being un-censored by time t given covariate profile V_i based on the assumed censoring model in Assumption 3.3 (Robins and Rotnitzky, 1992; Robins and Finkelstein, 2000; Cole and Hernán, 2004; Schaubel and Wei, 2011). We use $\widehat{p}_j^C(t|V_i) = \exp(-\widehat{\Lambda}_j^C(t|V_i))$ where $\widehat{\Lambda}_j^C(t|V_i)$ is the estimated cumulative hazard function for censoring variable profile V_i using Cox proportional hazards model with the specification in Assumption 3.3 and the Breslow estimator of the baseline hazard function.

Theorem 3.1 provides the asymptotic behavior of these inverse probability-weighted estimators.

Theorem 3.1 (Inverse Probability-Weighted Kaplan-Meier Estimators)**.** *Under Assumption 3.1, if Assumption 3.2 holds with $V = \emptyset$, then:*

(i) $\widehat{S}_j^{\mathrm{IPW}(1)}(t)$, the estimator from (3.4) *with $w_{ij} = \pi_j(Z_i)$, is uniformly consistent for $S_j(t)$, for $t \in [0, \tau(\emptyset)]$. Furthermore,*

$$\sqrt{n(\mathcal{I}_{\mathcal{J}})}\left(\widehat{S}_j^{\mathrm{IPW}(1)} - S_j\right) \rightsquigarrow \mathbb{G}\left(\Sigma_j^{\mathrm{IPW}(1)}\right), \tag{3.5}$$

where $\rightsquigarrow$ denotes weak convergence, and $\mathbb{G}$ is a mean-zero Gaussian process with variance function at t:

$$\Sigma_j^{\mathrm{IPW}(1)}(t) = S_j(t)^2 \int_0^t \frac{d\Lambda_j(s)}{\{1 - \Delta\Lambda_j(s)\} P_{\mathcal{J}}(R_j(s))} \cdot \frac{E_{\mathcal{J}}\left(I(A_i = j)\pi_j(Z)^{-2} R_j(s)\right)}{P_{\mathcal{J}}(R_j(s))},$$

where $\Lambda_j(s) = E_{\mathcal{J}} N_j(s) / E_{\mathcal{J}} R_j(s)$ is the cumulative hazard, $N_j(s) = I(Y^{(j)} \leq t, \delta^{(j)} = 1)$ is the event counting process, and $R_j(s) = I(Y^{(j)} \geq s)$ is the at-risk process.

(ii) $\widehat{S}_j^{\mathrm{IPW}(2)}(t)$, the estimator from (3.4) *with $w_{ij} = \widehat{\pi}_j(H_i) = n_j(H_i)/n(H_i)$, is uniformly consistent for $S_j(t)$, for $t \in [0, \tau(\emptyset)]$, where H_i satisfies Definition 3.3. Furthermore,* (3.5)

holds for $\widehat{S}_j^{\mathrm{IPW}(2)}$ *with variance function:*

$$\Sigma_j^{\mathrm{IPW}(2)}(t) = \Sigma_j^{\mathrm{IPW}(1)}(t) - [S_j(t)]^2 \sum_{h=1}^{L} p(h) \left[\frac{1-\pi_j(h)}{\pi_j(h)}\right] \left[\int_0^t \frac{E_{\mathcal{J}}(R_j(s)|H=h)\,[d\Lambda_{j,h}(s) - d\Lambda_j(s)]}{P_{\mathcal{J}}(R_j(s))\,[1-\Delta\Lambda_j(s)]}\right]^2,$$

where $\Lambda_{j,h}(s)$ *is the cumulative hazard function conditional on strata level* $H = h$*, and* $p(h) = P_{\mathcal{J}}(H = h)$*.*

If Assumptions 3.2 and 3.3 hold with non-empty V*, then:*

(iii) $\widehat{S}_j^{\mathrm{IPW}(3)}(t)$*, the estimator from* (3.4) *with* $w_{ij}(t) = \pi_j(Z_i)\widehat{p}_j^C(t|V_i)$*, and (iv)* $\widehat{S}_j^{\mathrm{IPW}(4)}(t)$*, the estimator from* (3.4) *with* $w_{ij}(t) = \widehat{\pi}_j(H_i)\widehat{p}_j^C(t|V_i)$*, are both uniformly consistent for* $S_j(t)$*, for* $t \in [0, \tau(V)]$*, and weakly converge to mean-zero Gaussian processes as in* (3.5)*, with influence functions given in the Supplemental Materials.*

From Theorem 3.1, we see that $\Sigma_j^{\mathrm{IPW}(2)}(t) \leq \Sigma_j^{\mathrm{IPW}(1)}(t)$, which means that we gain precision from estimating the treatment propensity weights. The gain is largest when the strata-specific cumulative hazard functions are notably different than the marginal cumulative hazard functions.

For variance estimation following Theorem 3.1, we define $d_{\mathcal{J}}^{(j)}(t) = \sum_{i\in\mathcal{I}_{\mathcal{J}}} \delta_i I(Y_i = t) I(A_i = j)\pi_j(Z_i)^{-1}$ and $n_{\mathcal{J}}^{(j)}(t) = \sum_{i\in\mathcal{I}_{\mathcal{J}}} I(Y_i \geq t) I(A_i = j)\pi_j(Z_i)^{-1}$ as the weighted event and at-risk counts at time s, based on known weights. Let $\widehat{d}_{\mathcal{J}}^{(j)}(t)$ and $\widehat{n}_{\mathcal{J}}^{(j)}(t)$ be their counterparts based on estimated propensities $\widehat{\pi}_j(H_i)$. A consistent variance estimator with known weights is:

$$\widehat{\Sigma}_{(1)}^{\mathrm{IPW}}(t) = \left[\widehat{S}_j^{\mathrm{IPW}(1)}(t)\right]^2 O_{n,j}(t)$$

$$O_{n,j}(t) = \sum_{s\in\mathcal{T}_{\mathcal{J}}^{(j)}: s\leq t} \left[\frac{d_{\mathcal{J}}^{(j)}(s)}{n_{\mathcal{J}}^{(j)}(s)\left\{n_{\mathcal{J}}^{(j)}(s) - d_{\mathcal{J}}^{(j)}(s)\right\}} \cdot \frac{\sum_{i\in\mathcal{I}_{\mathcal{J}}} \pi_j(Z_i)^{-2} I(A_i = j) I(Y_i \geq s)}{n_{\mathcal{J}}^{(j)}(s)}\right]$$

where $\mathcal{T}_{\mathcal{J}}^{(j)} = \{Y_i : \delta_i = 1, A_i = j\}_{i\in\mathcal{I}_{\mathcal{J}}}$ is the set of unique failure times for those within the ECE sample and assigned to treatment j. If the ECE population were the entire trial,

this variance estimator is equivalent to the variance estimator for the AKME with known weights proposed in Xie and Liu (2005). Let $d_{\mathcal{J}}^{(j,h)}(t) = \sum_{i \in \mathcal{I}_{\mathcal{J}}^{(h)}} \delta_i I(Y_i = t) I(A_i = j)$ and $n_{\mathcal{J}}^{(j,h)}(t) = \sum_{i \in \mathcal{I}_{\mathcal{J}}^{(h)}} I(Y_i \geq t) I(A_i = j)$ be the strata-specific event and at-risk sizes. Then a consistent variance estimator with estimated weights is:

$$\widehat{\Sigma}_{(2)}^{\text{IPW}}(t) = \left[\widehat{S}_j^{\text{IPW}(2)}(t)\right]^2 \left(\widehat{O}_{n,j}(t) - \sum_{h=1}^{L} \left\{ \widehat{\pi}_j(h)^{-1} \left[1 - \widehat{\pi}_j(h)\right] \widehat{p}(h) \left[n(\mathcal{I}_{\mathcal{J}}^{(j,h)}) \right]^{-2} \left[\widehat{D}_{n,j,h}(t) \right]^2 \right\} \right)$$

$$\widehat{D}_{n,j,h}(t) = \sum_{s \in \mathcal{T}_{\mathcal{J}}^{(j)} : s \leq t} \left(\frac{d_{\mathcal{J}}^{(j,h)}(s)}{n_{\mathcal{J}}^{(j,h)}(s)} - \frac{\widehat{d}_{\mathcal{J}}^{(j)}(s)}{\widehat{n}_{\mathcal{J}}^{(j)}(s)} \right) \left(\frac{n_{\mathcal{J}}^{(j,h)}(s)}{\widehat{n}_{\mathcal{J}}^{(j)}(s) - \widehat{d}_{\mathcal{J}}^{(j)}(s)} \right)$$

where $\widehat{O}_{n,h}(t)$ is $O_{n,h}(t)$ with $\pi_j(Z_i)$ replaced by $\widehat{\pi}_j(H_i)$, and $n(\mathcal{I}_{\mathcal{J}}^{(j,h)}) = \sum_{i \in \mathcal{I}_{\mathcal{J}}^{(h)}} I(A_i = j)$. This variance estimator is different than the variance estimator proposed by Xie and Liu (2005) with unknown weights. In their derivation, they treat the estimated $\widehat{\pi}_j$ as fixed. In our derivation, we apply a more traditional approach to estimating the variance of an inverse probability weighted estimator. As a result, we have one variance term for known weights (the $\widehat{O}_{n,j}(t)$) and subtract a term that is related to the estimation of $\widehat{\pi}_j$. Variance estimation for $\widehat{S}_j^{\text{IPW}(3)}(t)$ and $\widehat{S}_j^{\text{IPW}(4)}(t)$ is substantially more involved. When $V^P = \emptyset$, i.e., there are no proportional hazards covariates, then the model in Assumption 3.3 is non-parametric, and the probability of censoring can be estimated based on the Nelson-Aalen estimator within each strata and treatment group combination. This results in a much simplified influence function. We can then use the estimated influence function to calculate the variance of $\widehat{S}_j^{\text{IPW}(3)}(t)$ and $\widehat{S}_j^{\text{IPW}(4)}(t)$ (see Supplemental Materials for details). When V^P is non-empty, then the bootstrap procedure can be used to obtain variance estimates for $\widehat{S}_j^{\text{IPW}(3)}(t)$ and $\widehat{S}_j^{\text{IPW}(4)}(t)$, as is also done in other double-weighted Kaplan-Meier literature such as Schaubel and Wei (2011).

We also consider *post-stratification* (PS) estimators which estimate the survival function within the unique levels of a stratification variable H, given in Definition 3.3. Importantly, this H also needs to confer a random censoring model, which we specify in Assumption 3.4

below.

Assumption 3.4. *The random variable H described in Definition 3.3 also satisfies Assumption 3.2.*

Let $\mathcal{I}_{\mathcal{J}}^{(h)}$ be the set of observations within each strata level from $h = 1, \ldots, L$. Then the post-stratification estimator for the survival function is given by:

$$\widehat{S}_j^{\text{PSS}}(t) = \frac{1}{n\left(\mathcal{I}_{\mathcal{J}}\right)} \sum_{h=1}^{L} n\left(\mathcal{I}_{\mathcal{J}}^{(h)}\right) \left(\prod_{t_k \leq t} \left[1 - \frac{\sum_{i \in \mathcal{I}_{\mathcal{J}}^{(h)}: Y_i = t_k} I(A_i = j)\delta_i}{\sum_{i \in \mathcal{I}_{\mathcal{J}}^{(h)}: Y_i \geq t_k} I(A_i = j)} \right] \right) \tag{3.6}$$

The estimator in (3.6) is a weighted average of Kaplan-Meier estimates using individuals eligible for randomization to the comparison set $\mathcal{J}$, within unique levels of the treatment allocation vector, where the weighted average targets the ECE population denoted by $\mathcal{J}$. This is similar to G-computation on the scale of the predicted survival. Theorem 3.2 provides the asymptotic results for post-stratification estimators.

Theorem 3.2 (Asymptotic Behavior for Post-Stratification Estimators)**.** *Under Assumptions 3.1, and for some H that satisfies Assumption 3.4, $\widehat{S}_j^{\text{PSS}}(t)$ is uniformly consistent for $S_j(t)$, for $t \in [0, \tau(H)]$, and*

$$\sqrt{n(\mathcal{I}_{\mathcal{J}})} \left(\widehat{S}_j^{\text{PSS}} - S_j \right) \rightsquigarrow \mathbb{G}\left(\Sigma_j^{\text{PSS}} \right), \tag{3.7}$$

where $\mathbb{G}$ is a mean-zero Gaussian process with variance function at t given by:

$$\Sigma_j^{\text{PSS}}(t,t) = \sum_{h=1}^{L} S_{j,h}(t) P_{\mathcal{J}}(H = h) \left[\frac{S_{j,h}(t)}{\pi_j(h)} \int_0^t \frac{d\Lambda_{j,h}(s)}{E_{\mathcal{J}}(R_j(s)|H = h)\left\{1 - \Delta\Lambda_{j,h}(s)\right\}} + S_{j,h}(t) - S_j(t) \right].$$

Let $\mathcal{T}_{\mathcal{J}}^{(j,h)} = \{Y_i : \delta_i = 1, A_i = j\}_{i \in \mathcal{I}_{\mathcal{J}}^{(h)}}$ be the set of unique failure times for those within the ECE sample and assigned to treatment j and strata h. Define the strata-specific failure and risk set counts as $d_{\mathcal{J}}^{(j,h)}(t) = \sum_{i \in \mathcal{I}_{\mathcal{J}}^{(h)}} \delta_i I(Y_i = t) I(A_i = j)$ and $n_{\mathcal{J}}^{(j,h)}(t) = \sum_{i \in \mathcal{I}_{\mathcal{J}}^{(h)}} I(Y_i \geq t) I(A_i = j)$, respectively. Then a consistent variance estimator for the post-stratification

estimator in Theorem 3.2 is given by:

$$\widehat{\Sigma}^{\text{PSS}}(t) = n(\mathcal{I}_{\mathcal{J}})^{-1} \sum_{h=1}^{L} \widehat{p}_{\mathcal{J}}(h) \widehat{S}_{j,h}(t) \left[\widehat{S}_{j,h}(t) - \widehat{S}_{j}^{\text{PSS}}(t) \right] + \sum_{h=1}^{L} \frac{\widehat{S}_{j,h}(t)^2 \widehat{p}_{\mathcal{J}}(h)}{\widehat{\pi}_j(h)} \sum_{s \in \mathcal{T}_{\mathcal{J}}^{(j,h)}: s \leq t} \frac{d_{\mathcal{J}}^{(j,h)}(s)}{n_{\mathcal{J}}^{(j,h)}(s) \left[n_{\mathcal{J}}^{(j,h)}(s) - d_{\mathcal{J}}^{(j,h)}(s) \right]},$$

where $\widehat{S}_{j,h}(t)$ is the Kaplan-Meier estimator in (3.2) using only individuals within the strata h $(H_i = h)$, $\widehat{p}_{\mathcal{J}}(h) = n\left(\mathcal{I}_{\mathcal{J}}^{(h)}\right) n(\mathcal{I}_{\mathcal{J}})^{-1}$, $\widehat{\pi}_j(h) = \sum_{i=1}^{n} I(A_i = j) I(H_i = h) / \sum_{i=1}^{n} I(H_i = h)$.

3.5 Simulations

We perform a simulation study to empirically evaluate the finite-sample behavior of the methods that we presented in Section 3.4. We simulate data from a platform trial with the structure shown in Figure 3. Our estimands of interest are the difference in potential survival probability at 23 months of follow-up if an individual eligible for both placebo and treatment 2 receives either placebo or treatment 2. The design of the trial is as follows. Initially only two arms are open for enrollment. After one year, two additional treatment arms are added, and one of them is only available to individuals with a prognostic biomarker $R = 1$. Each treatment arm has two years of enrollment, and a minimum of one year of follow-up. All four trial arms are administratively censored at year 4, and each individual has a maximum of 3 years of follow-up. We consider two Randomization Regimes (a) and (b) for the treatment propensities, which are given in Table 3.2. In Regime (a), each individual has equal probability of randomization to all treatment arms that are available and that they are eligible for, at the time they enroll. Eligibility is based off of the variable R. Therefore, $Z = (E, R)$. In Regime (b), the allocation ratios for $R = 0$ are the same as Regime (a). But for $R = 1$, treatment 3 is favored over 2 and 1, since it is specifically available for those with prognostic variable $R = 1$. The window that someone enrolls in is a function of observed marker R and an unobserved binary prognostic marker U. We consider two different cases:

The design of the trial is as follows. Initially only two arms are open for enrollment. After one year, two additional treatment arms are added, and one of them is only available to individuals with a prognostic biomarker $R = 1$. Each treatment arm has two years of enrollment, and a minimum of one year of follow-up. All four trial arms are administratively censored at year 4, and each individual has a maximum of 3 years of follow-up. We consider two Randomization Regimes (a) and (b) for the treatment propensities, which are given in Table 3.2. In Regime (a), each individual has equal probability of randomization to all treatment arms that are available and that they are eligible for, at the time they enroll. Eligibility is based off of the variable R. Therefore, $Z = (E, R)$. In Regime (b), the allocation ratios for $R = 0$ are the same as Regime (a). But for $R = 1$, treatment 3 is favored over 2 and 1, since it is specifically available for those with prognostic variable $R = 1$. The window that someone enrolls in is a function of observed and unobserved prognostic variables, which are described in the following cases:

(1) Risk Profile Constant: The enrollment windows each have the same distribution of individuals with binary prognostic biomarkers R and U. The percentage of individuals with $R = 1$ in the whole trial is 50%, and the percent of individuals with $U = 1$ in the whole trial is also 50%.

(2) Higher Risk Over Time: More individuals with $R = 1$, and fewer individuals with $U = 1$, enroll over time. This leads to the trial population being higher risk in the latter enrollment windows. The percent of individuals with $R = 1$ among those enrolled in windows 1, 2, and 3, are approximately 15%, 55%, and 70%, respectively. The percent of individuals with $U = 1$ among those enrolled in windows 1 and 2 is approximately 70%, and among those enrolled in window 3 is 10%.

For Case 1, the sufficient adjustment set for censoring based on Assumption 3.2 is $V = \emptyset$; for case 2, it is $V = (R, E)$. It does not require a proportional hazards censoring model in Assumption 3.3 because the V in each case only consists of categorical variables. Therefore, adjustment for censoring can be done completely nonparametrically. We simulate survival

Figure 3.3: Structure of the platform trial for simulations. At time 0, the placebo arm and treatment arm 1 open for enrollment. At time 1, treatment arm 2 open for enrollment. At time 2, treatment arm 1 ceases enrollment, and at time 3, treatment arm 2 ceases enrollment. The trial arms are all administratively censored at time 4 so that all participants have at least 1 year of follow-up, and individuals are censored at 3 years of follow-up.

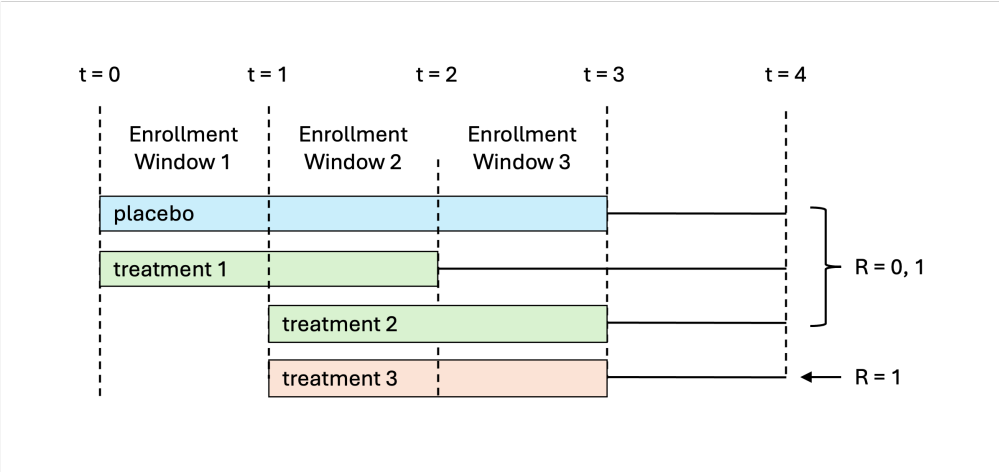


Table 3.2: Placebo and treatment group assignment probabilities by prognostic marker R and enrollment window, in Randomization Regimes (a) and (b).

	Enrollment Window	Placebo	Treatment 1	Treatment 2	Treatment 3
$R = 0$	1	1/2	1/2	0	0
	2	1/3	1/3	1/3	0
	3	1/2	0	1/2	0
$R = 1$ in Regime (a)	1	1/2	1/2	0	0
	2	1/4	1/4	1/4	1/4
	3	1/3	0	1/3	1/3
$R = 1$ in Regime (b)	1	1/2	1/2	0	0
	2	4/10	1/10	1/10	4/10
	3	2/5	0	1/5	2/5

time as an exponential random variable: un-treated individuals with $R = 0$ and $R = 1$ have a median survival time of 4 years and 6 months, respectively, so that R is a negatively prognostic biomarker. Regardless of their value of R, individuals with $U = 1$ have 75% lower hazard than individuals with $U = 0$, so U is a highly beneficial marker. Within a specific

enrollment window, enrollment time is uniform.

We consider the following estimators in our simulation: the Kaplan-Meier estimator in (3.2) (denoted as KM), the Kaplan-Meier estimator using only ECE samples in (3.3) (denoted as KM-ECE), the proposed Kaplan-Meier IPW estimator with four different options. We consider the Kaplan-Meier IPW estimator with known or estimated treatment propensity weights for Case (1). For Case (2), we add inverse probability of censoring weights (IPCW) using $V = (E, R)$. In both cases, we also consider the post-stratification estimator in (3.6), with H as the unique levels of E and R.

Figure 3 shows the results of our simulation study across 5000 simulated datasets of studies each having $n = 5000$ (for the whole platform), for Cases 1 and 2, and with the allocation ratios corresponding to Randomization Regimes (a) and (b). We add the estimand in the ECE population to the plot along with the estimand in the ACE population (ECE + as-treated) for a comparison.

We summarize the findings from our simulations below, based on the results in Figure 4:

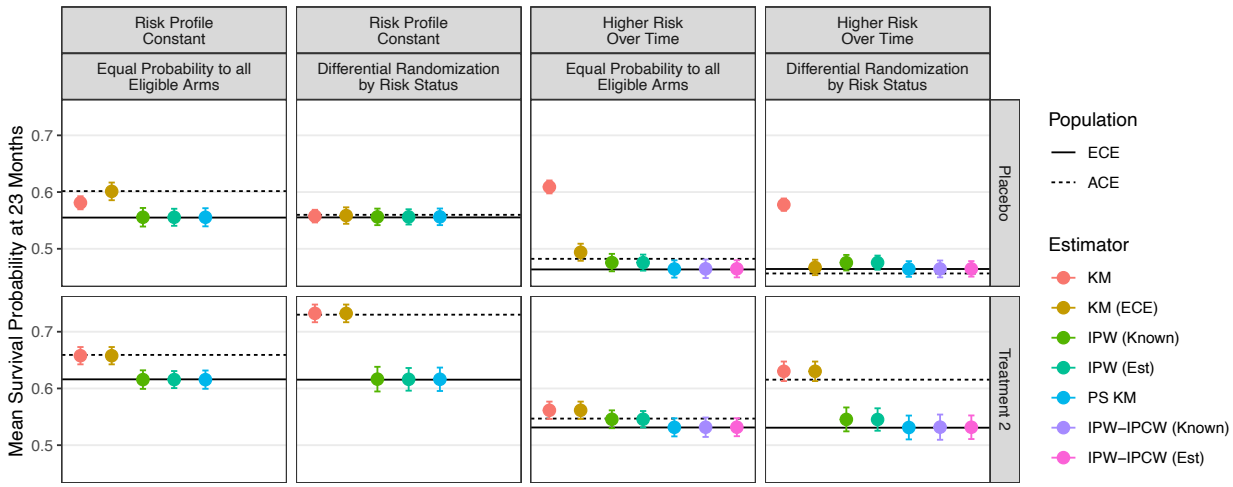
- The inclusion of Randomization Regimes (a) and (b) shows how changing the allocation ratios changes the ACE population estimand, but does not change the ECE population estimand. That is, the ACE population is a function of the allocation ratios, whereas the ECE population is not. This shows the benefits of defining the estimand as a population summary with respect to the ECE population, rather than the ACE population, because it is unaffected by these design considerations.

- The naive Kaplan-Meier estimator using all individuals as they were assigned to either treatment 1 or placebo results in biased estimates of ECE treatment effect. This is because even in Case 1, the randomization ratios change over time, and they change differently for those with $R = 0$ and $R = 1$ since those with $R = 0$ are not eligible to treatment arm 2. We have even more bias in Cases 2, when the distribution of $R = 1$ in the trial changes by enrollment window.

- Using the Kaplan-Meier ACE estimator from (3.3) accurately estimates the ACE treatment effect under a constant risk profile by restricting the analysis set to individuals who were eligible for randomization to treatment 1 or placebo. Specifically, this excludes individuals enrolled in window 3, since those individuals were not eligible for randomization to treatment 1. However, this approach is biased for the ECE treatment effect. It also does not account for dependent censoring.

- Using any of our proposed estimators eliminates the bias as long as dependent censoring, when present, is accounted for. As expected, estimating the treatment propensity weights π_j results in slightly lower variability of the effect estimates compared to using known weights.

Figure 3.4: Simulation results for Cases 1 and 2.



3.6 Discussion

A clear and unambiguous population of interest for platform trials is the *entire concurrent and eligible* population. We introduced a meaningful estimand for this population for clinical trials that have a time-to-event outcome, and enumerated the assumptions that are required

for its identification. Building on the landmark work of Qian et al. (2025), we proposed several new features for time-to-event analysis in the ECE population, including valid follow-up times based on sufficient adjustment sets for dependent censoring. We showed how to define sufficient adjustment sets for dependent censoring in platform trials in both stylistic and real case examples. We have proposed several estimators, including inverse probability-weighted and post-stratification estimators, and justify their use based on a simulation study. Stratified estimators have previously been proposed for platform trials when there are temporal shifts in the distribution of prognostic variables among those enrolling in the trial. Stratification addresses the confounding that can result if randomization probabilities also change over time (Bofill Roig et al., 2024). However, in these cases the strata-specific estimates are re-combined to match the characteristics of the population that was randomized to the treatments of interest, rather than to the population that was eligible for the treatments of interest. In our work, we are focused on the ECE population. We provide numerous examples of how our framework can be applied in stylistic and real examples, to clearly define an estimand, and develop relevant estimators.

The nuances that we discuss in Section 3.4.1 about the risk set composition highlight an important issue when using the sub-study design. In our stylistic example, we assumed that the experience of the placebo group is identical across sub-studies. This may be a reasonable assumption if the study is using a so-called "multiple-dummy design," where the participants are totally blinded to both their sub-study and treatment assignments (U.S. Food and Drug Administration, 2023), or if the trial is open-label and all participants assigned to placebo are receiving the same standard of care regardless of sub-study. However, with "partial blinding" strategies, where participants are aware of their sub-study assignment, but blinded to whether they are receiving placebo or experimental therapy, the placebo treatment received by placebo participants in sub-study A may be completely different than sub-study B. Ideally, placebos are inert and have no affect on the outcome, but there may be cases where in order to closely resemble an active therapy, the placebo treatment has unintended effects on the outcome (Demasi and Jefferson, 2021). Or there may be different routes of administration

for placebos, which could impact patient outcomes (Webb and Derde, 2025). Additionally, participants may change their behavior if they know their sub-study assignment. In these cases, combining placebos across sub-studies may not be advisable.

There are variations on platform trials that our work does not address, and which should be the subject of future research. Some master protocols are "adaptive" in that they use information about currently observed responses on different treatment arms to inform future allocation ratios (Angus et al., 2019). A notable example is I-SPY 2 trial of novel therapeutics in combination with neoadjuvant chemotherapy for the treatment of locally advanced breast cancer (Barker et al., 2009). Adaptive platform trials do not fit into our framework, since the baseline covariate vector Z no longer confers conditional independence between treatment assignment and the potential outcomes (treatment assignments are now intrinsically related to potential outcomes). Given that these trials have the complex design of a master protocol and also incorporate the complications of response-adaptive randomization, it is essential that future work carefully defines estimands in this setting as well.

Some design variations can easily be incorporated into our framework. For example, some trials allow for re-randomization of participants to another arm at, say, progression. In Lung-MAP, almost 5% of patients were re-assigned to a new sub-study after progression on a prior sub-study (Redman et al., 2020). Since we assume independent and identically distributed observations, this type of study does not directly fit into our framework. However, we can still analyze data from these studies if we restrict our attention to only sub-studies that do collectively have independent observations in our definition of the comparison set $\mathcal{J}$. Handling re-enrollment data explicitly will be the subject of future work.

There is substantial room for methodological innovation in the framework that we have laid out, beyond the proposed estimators in this current paper. In the adjustment for censoring, we only allowed for baseline covariates, but it may be desirable to also allow for time-varying covariates in the Cox proportional hazards model for censoring. Incorporating time-varying covariates in the Cox model is not new (Robins and Rotnitzky, 1992), but care must be taken to work this into our framework, specifically with regards to Assumption

3.2. Another area for innovation is to allow for different estimands that are relevant in the context of competing risks, for example in estimating the cumulative incidence function, or cause-specific hazard functions (Kalbfleisch and Prentice, 2002; Deng and Wang, 2025). Finally, covariate adjustment is a technique that can be used to reduce the variability of estimates of treatment effect, and when used properly, it does not change the estimand. Future work could expand the estimators that we have proposed to allow for augmentation with covariates to improve efficiency (Zhang et al., 2008).

BIBLIOGRAPHY

Angus, D. C., Alexander, B. M., Berry, S., Buxton, M., Lewis, R., Paoloni, M., Webb, S. A. R., Arnold, S., Barker, A., Berry, D. A., Bonten, M. J. M., Brophy, M., Butler, C., Cloughesy, T. F., Derde, L. P. G., Esserman, L. J., Ferguson, R., Fiore, L., Gaffey, S. C., Gaziano, J. M., Giusti, K., Goossens, H., Heritier, S., Hyman, B., Krams, M., Larholt, K., LaVange, L. M., Lavori, P., Lo, A. W., London, A. J., Manax, V., McArthur, C., O'Neill, G., Parmigiani, G., Perlmutter, J., Petzold, E. A., Ritchie, C., Rowan, K. M., Seymour, C. W., Shapiro, N. I., Simeone, D. M., Smith, B., Spellberg, B., Stern, A. D., Trippa, L., Trusheim, M., Viele, K., Wen, P. Y., Woodcock, J., and The Adaptive Platform Trials Coalition (2019). Adaptive platform trials: Definition, design, conduct and reporting considerations. *Nature Reviews Drug Discovery*, 18(10):797–807.

Antognini, A. B. and Zagoraiou, M. (2015). On the almost sure convergence of adaptive allocation procedures. *Bernoulli*, 21(2):881–908.

Armitage, P., McPherson, C. K., and Rowe, B. C. (1969). Repeated Significance Tests on Accumulating Data. *Journal of the Royal Statistical Society. Series A (General)*, 132(2):235–244.

Barker, A., Sigman, C., Kelloff, G., Hylton, N., Berry, D., and Esserman, L. (2009). I-SPY 2: An Adaptive Breast Cancer Trial Design in the Setting of Neoadjuvant Chemotherapy. *Clinical Pharmacology & Therapeutics*, 86(1):97–100.

Bateman, R. J., Benzinger, T. L., Berry, S., Clifford, D. B., Duggan, C., Fagan, A. M., Fanning, K., Farlow, M. R., Hassenstab, J., McDade, E. M., Mills, S., Paumier, K., Quintana, M., Salloway, S. P., Santacruz, A., Schneider, L. S., Wang, G., Xiong, C., and the, D.-T. P. C. f. t. D. I. A. N. (2017). The DIAN-TU Next Generation Alzheimer's

prevention trial: Adaptive design and disease progression model. *Alzheimer's & Dementia*, 13(1):8–19.

Berry, S. M., Connor, J. T., and Lewis, R. J. (2015). The Platform Trial: An Efficient Strategy for Evaluating Multiple Treatments. *JAMA*, 313(16):1619–1620.

Bofill Roig, M., Glimm, E., Mielke, T., and Posch, M. (2024). Optimal allocation strategies in platform trials with continuous endpoints. *Statistical Methods in Medical Research*, 33(5):858–874.

Bozic, M., Goss, C. H., Tirouvanziam, R. M., Baines, A., Kloster, M., Antoine, L., Borowitz, D., Schwarzenberg, S. J., and study Group, f. t. G. (2020). Oral Glutathione and Growth in Cystic Fibrosis: A Multicenter, Randomized, Placebo-controlled, Double-blind Trial. *Journal of Pediatric Gastroenterology and Nutrition*, 71(6):771–777.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017). Double/Debiased/Neyman Machine Learning of Treatment Effects. *The American Economic Review*, 107(5):261–265.

Cohen, P. L. and Fogarty, C. B. (2024). No-harm calibration for generalized Oaxaca–Blinder estimators. *Biometrika*, 111(1):331–338.

Colantuoni, E. and Rosenblum, M. (2015). Leveraging prognostic baseline variables to gain precision in randomized trials. *Statistics in Medicine*, 34(18):2602–2617.

Cole, S. R. and Hernán, M. A. (2004). Adjusted survival curves with inverse probability weights. *Computer Methods and Programs in Biomedicine*, 75(1):45–49.

Davies, M. J., Russell-Jones, D., Selam, J.-L., Bailey, T. S., Kerényi, Z., Luo, J., Bue-Valleskey, J., Iványi, T., Hartman, M. L., Jacobson, J. G., Jacober, S. J., and IMAGINE 2 Study Investigators (2016). Basal insulin peglispro versus insulin glargine in insulin-naïve type 2 diabetes: IMAGINE 2 randomized trial. *Diabetes, Obesity & Metabolism*, 18(11):1055–1064.

Demasi, M. and Jefferson, T. (2021). Placebo—The Unknown Variable in a Controlled Trial. *JAMA Internal Medicine*, 181(5):577–578.

Deng, Y. and Wang, R. (2025). Adjusted Nelson–Aalen estimators by inverse treatment probability weighting with an estimated propensity score.

Elm, JordanJ., Palesch, YukoY., Koch, GaryG., Hinson, V., Ravina, B., and Zhao, W. (2012). Flexible Analytical Methods for Adding a Treatment Arm Mid-Study to an Ongoing Clinical Trial. *Journal of Biopharmaceutical Statistics*, 22(4):758–772.

EMA (2015). Guideline on adjustment for baseline covariates in clinical trials.

Emerson, S. S., Kittelson, J. M., and Gillen, D. L. (2007). Bayesian evaluation of group sequential clinical trial designs. *Statistics in Medicine*, 26(7):1431–1449.

Firth, D. and Bennett, K. E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):3–21.

Fleming, T. R. and Harrington, D. P. (2011). *Counting Processes and Survival Analysis*. John Wiley & Sons.

Freedman, D. A. (2008). Randomization Does Not Justify Logistic Regression. *Statistical Science*, 23(2):237–249.

Gernot, W. and Brannath, W. (2016). *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Springer.

Glynn, A. N. and Quinn, K. M. (2010). An Introduction to the Augmented Inverse Propensity Weighted Estimator. *Political Analysis*, 18(1):36–56.

Guo, K. and Basse, G. (2021). The Generalized Oaxaca-Blinder Estimator. *Journal of the American Statistical Association*, 118(541):524–536.

Hu, F., Ye, X., and Zhang, L.-X. (2023). Multi-arm covariate-adaptive randomization. *Science China. Mathematics*, 66(1):163–190.

Huang, T.-J., Luedtke, A., and THE AMP INVESTIGATOR GROUP (2023). Improved efficiency for cross-arm comparisons via platform designs. *Biostatistics*, 24(4):1106–1124.

ICH (2019). E9(R1) Statistical Principles for Clinical Trials: Addendum: Estimands and Sensitivity Analysis in Clinical Trials. Technical report, FDA.

Kaizer, A. M., Hobbs, B. P., and Koopmeiners, J. S. (2018). A multi-source adaptive platform design for testing sequential combinatorial therapeutic strategies. *Biometrics*, 74(3):1082–1094.

Kalbfleisch, J. D. and Prentice, R. L. (2002). Competing Risks and Multistate Models. In *The Statistical Analysis of Failure Time Data*, chapter 8, pages 247–277. John Wiley & Sons, Ltd.

Kim, E. S., Herbst, R. S., Wistuba, I. I., Lee, J. J., Blumenschein, G. R., Tsao, A., Stewart, D. J., Hicks, M. E., Erasmus, J., Gupta, S., Alden, C. M., Liu, S., Tang, X., Khuri, F. R., Tran, H. T., Johnson, B. E., Heymach, J. V., Mao, L., Fossella, F., Kies, M. S., Papadimitrakopoulou, V., Davis, S. E., Lippman, S. M., and Hong, W. K. (2011). The BATTLE trial: Personalizing therapy for lung cancer. *Cancer Discovery*, 1(1):44–53.

Kim, K. (1989). Point Estimation Following Group Sequential Tests. *Biometrics*, 45(2):613–617.

Kim, K. and DeMets, D. L. (1987). Confidence Intervals Following Group Sequential Tests in Clinical Trials. *Biometrics*, 43(4):857–864.

Kim, K. and Tsiatis, A. A. (2020). Independent increments in group sequential tests: A review. *SORT-Statistics and Operations Research Transactions*, pages 223–264.

Koenig, F., Spiertz, C., Millar, D., Rodríguez-Navarro, S., Machín, N., Van Dessel, A., Genescà, J., Pericàs, J. M., and Posch, M. (2023). Current state-of-the-art and gaps in platform trials: 10 things you should know, insights from EU-PEARL. *eClinicalMedicine*, 67:102384.

Lan, K. K. G. and DeMets, D. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663.

Lee, K. M. and Wason, J. (2020). Including non-concurrent control patients in the analysis of platform trials: Is it worth it? *BMC Medical Research Methodology*, 20(1):NA–NA.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318.

Liu, H. and Yang, Y. (2020). Regression-adjusted average treatment effect estimates in stratified randomized experiments. *Biometrika*, 107(4):935–948.

Moore, K. L. and van der Laan, M. J. (2009). Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Statistics in medicine*, 28(1):39–64.

Nichols, D. P., Singh, P. K., Baines, A., Caverly, L. J., Chmiel, J. F., GIbson, R. L., Lascano, J., Morgan, S. J., Retsch-Bogart, G., Saiman, L., Sadeghi, H., Billings, J. L., Heltshe, S. L., Kirby, S., Kong, A., Nick, J. A., and Mayer-Hamblett, N. (2021). Testing the effects of combining azithromycin with inhaled tobramycin for *P. aeruginosa* in cystic fibrosis: A randomised, controlled clinical trial. *Thorax*, pages thoraxjnl–2021–217782.

O'Brien, P. C. and Fleming, T. R. (1979). A Multiple Testing Procedure for Clinical Trials. *Biometrics*, 35(3):549–556.

Paganoni, S., Berry, J. D., Quintana, M., Macklin, E., Saville, B. R., Detry, M. A., Chase, M., Sherman, A. V., Yu, H., Drake, K., Andrews, J., Shefner, J., Chibnik, L. B., Vestrucci,

M., Cudkowicz, M. E., and Group, f. t. H. A. P. T. S. (2022). Adaptive Platform Trials to Transform Amyotrophic Lateral Sclerosis Therapy Development. *Annals of Neurology*, 91(2):165–175.

Parker, C., Sydes, M., Mason, M., Clarke, N., Aebersold, D., De Bono, J., Dearnaley, D., Ritchie, A., Russell, J., Thalmann, G., Parmar, M., and James, N. (2013). Prostate Radiotherapy for Men with Metastatic Disease: A New Comparison in the STAMPEDE Trial. *Clinical Oncology*, 25(5):318–320.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199.

Pocock, S. J. and Simon, R. (1975). Sequential Treatment Assignment with Balancing for Prognostic Factors in the Controlled Clinical Trial. *Biometrics*, 31(1):103–115.

Qian, Y., Yi, Y., Shao, J., Yi, Y., Levin, G., Mayer-Hamblett, N., Heagerty, P. J., and Ye, T. (2025). From Estimands to Robust Inference of Treatment Effects in Platform Trials.

Quintana, M., Saville, B. R., Vestrucci, M., Detry, M. A., Chibnik, L., Shefner, J., Berry, J. D., Chase, M., Andrews, J., Sherman, A. V., Yu, H., Drake, K., Cudkowicz, M., Paganoni, S., Macklin, E. A., and Group, f. t. H. A. P. T. S. (2023). Design and Statistical Innovations in a Platform Trial for Amyotrophic Lateral Sclerosis. *Annals of Neurology*, 94(3):547–560.

Rafi, A. (2023). Efficient Semiparametric Estimation of Average Treatment Effects Under Covariate Adaptive Randomization.

Redman, M. W., Papadimitrakopoulou, V. A., Minichiello, K., Hirsch, F. R., Mack, P. C., Schwartz, L. H., Vokes, E., Ramalingam, S., Leighl, N., Bradley, J., Miao, J., Moon, J., Highleyman, L., Miwa, C., LeBlanc, M. L., Malik, S., Miller, V. A., Sigal, E. V., Adam, S., Wholley, D., Sigman, C., Smolich, B., Blanke, C. D., Kelly, K., Gandara, D. R., and

Herbst, R. S. (2020). Biomarker-driven therapies for previously treated squamous non-small-cell lung cancer (Lung-MAP SWOG S1400): A biomarker-driven master protocol. *The Lancet Oncology*, 21(12):1589–1601.

Robins, J. M. and Finkelstein, D. M. (2000). Correcting for Noncompliance and Dependent Censoring in an AIDS Clinical Trial with Inverse Probability of Censoring Weighted (IPCW) Log-Rank Tests. *Biometrics*, 56(3):779–788.

Robins, J. M. and Rotnitzky, A. (1992). Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers. In Jewell, N. P., Dietz, K., and Farewell, V. T., editors, *AIDS Epidemiology: Methodological Issues*, pages 297–331. Birkhäuser, Boston, MA.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*, 89(427):846–866.

Robinson, L. D. and Jewell, N. P. (1991). Some Surprising Results about Covariate Adjustment in Logistic Regression Models. *International Statistical Review / Revue Internationale de Statistique*, 59(2):227–240.

Rosenblum, M. and Steingrimsson, J. A. (2016). Matching the Efficiency Gains of the Logistic Regression Estimator While Avoiding its Interpretability Problems, in Randomized Trials.

Rosenblum, M. and van der Laan, M. J. (2010). Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. *The International Journal of Biostatistics*, 6(1):Article 13.

Santacatterina, M., Giron, F. M., Zhang, X., and Díaz, I. (2025). Identification and Estimation of Causal Effects Using Non-Concurrent Controls in Platform Trials. *Statistics in Medicine*, 44(6):e70017.

Schaubel, D. E. and Wei, G. (2011). Double Inverse-Weighted Estimation of Cumulative Treatment Effects under Nonproportional Hazards and Dependent Censoring. *Biometrics*, 67(1):29–38.

Schiavone, F., Bathia, R., Letchemanan, K., Masters, L., Amos, C., Bara, A., Brown, L., Gilson, C., Pugh, C., Atako, N., Hudson, F., Parmar, M., Langley, R., Kaplan, R. S., Parker, C., Attard, G., Clarke, N. W., Gillessen, S., James, N. D., Maughan, T., Sydes, M. R., and On behalf of past and present members of the STAMPEDE and FOCUS4 Trial Management Group (2019). This is a platform alteration: A trial management perspective on the operational aspects of adaptive and platform and umbrella protocols. *Trials*, 20(1):264.

Siegmund, D. (1978). Estimation Following Sequential Tests. *Biometrika*, 65(2):341–349.

Steingrimsson, J. A., Hanley, D. F., and Rosenblum, M. (2017). Improving precision by adjusting for prognostic baseline variables in randomized trials with binary outcomes, without regression model assumptions. *Contemporary Clinical Trials*, 54:18–24.

Sydes, M. R., Parmar, M. K., James, N. D., Clarke, N. W., Dearnaley, D. P., Mason, M. D., Morgan, R. C., Sanders, K., and Royston, P. (2009). Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: The MRC STAMPEDE trial. *Trials*, 10(1):39.

Sydes, M. R., Parmar, M. K. B., Mason, M. D., Clarke, N. W., Amos, C., Anderson, J., de Bono, J., Dearnaley, D. P., Dwyer, J., Green, C., Jovic, G., Ritchie, A. W. S., Russell, J. M., Sanders, K., Thalmann, G., and James, N. D. (2012). Flexible trial design in practice - stopping arms for lack-of-benefit and adding research arms mid-trial in STAMPEDE: A multi-arm multi-stage randomized controlled trial. *Trials*, 13:168.

Taves, D. R. (1974). Minimization: A new method of assigning patients to treatment and control groups. *Clinical Pharmacology and Therapeutics*, 15(5):443–453.

The Accelerating Covid-19 Therapeutic Interventions and Vaccines (ACTIV)-6 Study Group (2023). ACTIV-6: Operationalizing a decentralized, outpatient randomized platform trial to evaluate efficacy of repurposed medicines for COVID-19. *Journal of Clinical and Translational Science*, 7(1):e221.

Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine*, 27(23):4658–4677.

U.S. Food and Drug Administration (2023). Master Protocols for Drug and Biological Product Development. *Product Development*.

U.S. Food and Drug Administration (2024). Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adjusting-covariates-randomized-clinical-trials-drugs-and-biological-products.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

van der Vaart, AW. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media.

Van Lancker, K., Betz, J., and Rosenblum, M. (2022). Combining Covariate Adjustment with Group Sequential and Information Adaptive Designs to Improve Randomized Trial Efficiency.

Van Lancker, K., Vandebosch, A., and Vansteelandt, S. (2019). Improving interim decisions in randomized trials by exploiting information on short-term outcomes and prognostic baseline covariates.

Ventz, S., Cellamare, M., Parmigiani, G., and Trippa, L. (2018). Adding experimental arms to platform clinical trials: Randomization procedures and interim analyses. *Biostatistics*, 19(2):199–215.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Wang, B., Ogburn, E. L., and Rosenblum, M. (2019). Analysis of covariance in randomized trials: More precision and valid confidence intervals, without model assumptions. *Biometrics*, 75(4):1391–1400.

Wang, B., Susukida, R., Mojtabai, R., Amin-Esmaeili, M., and Rosenblum, M. (2021). Model-Robust Inference for Clinical Trials that Improve Precision by Stratified Randomization and Covariate Adjustment. *Journal of the American Statistical Association*, 0(0):1–12.

Webb, S. A. and Derde, L. P. G. (2025). Platform Clinical Trials for the Efficient Evaluation of Multiple Treatments. *JAMA*, 333(13):1166–1167.

Woodcock, J. and LaVange, L. M. (2017). Master Protocols to Study Multiple Therapies, Multiple Diseases, or Both. *New England Journal of Medicine*, 377(1):62–70.

Xie, J. and Liu, C. (2005). Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in Medicine*, 24(20):3089–3110.

Yang, L. and Tsiatis, A. A. (2001). Efficiency Study of Estimators for a Treatment Effect in a Pretest-Posttest Trial. *The American Statistician*, 55(4):314–321.

Ye, T., Shao, J., Yi, Y., and Zhao, Q. (2023). Toward Better Practice of Covariate Adjustment in Analyzing Randomized Clinical Trials. *Journal of the American Statistical Association*, 118(544):2370–2382.

Yuan, Y., Guo, B., Munsell, M., Lu, K., and Jazaeri, A. (2016). MIDAS: A practical Bayesian design for platform trials with molecularly targeted agents. *Statistics in Medicine*, 35(22):3892–3906.

Zelen, M. (1974). The randomization and stratification of patients to clinical trials. *Journal of Chronic Diseases*, 27(7):365–375.

Zhang, H. and Ding, F. (2013). On the Kronecker Products and Their Applications. *Journal of Applied Mathematics*, 2013:e296185.

Zhang, M. and Schaubel, D. E. (2012). Contrasting treatment-specific survival using double-robust estimators. *Statistics in Medicine*, 31(30):4255–4268.

Zhang, M., Tsiatis, A. A., and Davidian, M. (2008). Improving Efficiency of Inferences in Randomized Clinical Trials Using Auxiliary Covariates. *Biometrics*, 64(3):707–715.

Appendix A

SUPPLEMENTAL MATERIAL FOR CHAPTER 1

A.1 Review of Covariate Adjustment in RCTs

In this section, we re-prove well-known results about ANCOVA in randomized trials with simple randomization and equal treatment allocation Yang and Tsiatis (2001); Colantuoni and Rosenblum (2015); Wang et al. (2019). The lemmas in this section form the proof of Lemma 1.1 for identification and estimation. Since ANOVA and ANCOVA can be written in the same way as the ordinary least squares (OLS) estimator, we first prove a results about OLS, and then apply them to the ANOVA and ANCOVA setting. We do this first for parameter identification, and then for parameter estimation and asymptotic behavior.

A.1.1 Parameter Identification

Lemma A.1 (Parameter Identification for OLS)**.** *Consider* $(Z, Y) \sim P$ *where* P *is the population joint distribution of* Z *and* Y*, and* $Z \in \mathbb{R}^d$*. We define*

$$\beta^* = \arg\min_{\beta \in \mathbb{R}} E_P[(Y - Z^T\beta)^2]$$

Assume that $\mathrm{Var}(Y - Z^T\beta) < \infty$*. Then* $\beta^* = E_P[ZZ^T]^{-1}E_P[ZY]$*.*

Proof. Note that this loss function is convex in β, so we can find a unique minimizer β^*. Thus, β^* is the solution to:

$$\begin{aligned}
\frac{\partial}{\partial\beta}\int (y - z^T\beta)^2 \quad dP(z,y) &= 0 \\
\int 2x(y - z^T\beta) \quad dP(z,y) &= 0 \\
\int 2zy \quad dP(z,y) &= \int 2xz^T\beta \quad dP(z,y) \\
E_P[ZY] &= E_P[ZZ^T]\beta \\
\beta^* &= E_P[ZZ^T]^{-1}E_P[ZY].
\end{aligned}$$

□

Lemma A.2 (Identification of Average Treatment Effect: Lemma 1.1(i))**.** *Under Assumptions 1.2 and 1.3, the population minimizer for* (1.1) *is* $(\Delta^*, \theta^*) = E_P[(A,1)(A,1)^T]E_P[(A,1)^T Y]$*, and population minimizer for* (1.2) *is*

$$(\Delta^{*C}, \theta^{*C}, [\gamma^{*C}]^T)^T = E_P[(A,1,X^T)^T(A,1,X^T)]^{-1}E_p[(A,1,X^T)Y].$$

In both cases, because of the independence of A and X from Assumption 1.2,

$$2\Delta^* = 2\Delta^{*C} = E_P[Y|A=1] - E_P[Y|A=-1]. \tag{A.1}$$

Proof. ANOVA and ANCOVA are special cases of the least squares formulation that we have presented in Lemma A.1. Consider the two sets of predictors,

$$\begin{aligned} \text{ANOVA:} \quad & \beta := (\Delta, \theta) && Z := (A,1) \\ \text{ANCOVA:} \quad & \beta^C := (\Delta^C, \theta^C, \gamma^C) && Z^C := (A,1,X) \end{aligned}$$

where $X \in \mathbb{R}^p$ are covariates, and $A \in \{-1,1\}$ is a treatment group indicator. In the ANCOVA model, we assume that $X \perp A$, i.e. covariates are independent of treatment assignment in the population. Additionally, assume that we have equal randomization probabilities for the two treatment groups,

$$P[A=a] = \begin{cases} 0.5 & \text{if} \quad a=1 \\ 0.5 & \text{if} \quad a=-1. \end{cases}$$

Using the form of Ω as defined from Theorem A.1, let $\Omega = E_P[ZZ^T]$, and $\Omega^C = E_P[Z^C[Z^C]^T]$. Note that $E[A]=0$, and $E[A^2]=1$ since $A^2=1$ with probability 1, and that

$E[AX_j] = 0$ since the covariates are independent of the treatment assignment. Therefore,

$$\Omega = E_P[(A,1)(A,1)^T] = \begin{pmatrix} E_P[A^2] & E_P[A] \\ E_P[A] & E_P[1^2] \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\Omega^C = E_P[(A,1,X^T)^T(A,1,X^T)] = \begin{pmatrix} \Omega_{1,1} & \Omega_{1,2} & E_P[AX^T] \\ \Omega_{2,1} & \Omega_{2,2} & E_P[X^T] \\ E_P[AX] & E_P[X] & E_P[XX^T] \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0_p^T \\ 0 & 1 & E_P[X^T] \\ 0_p & E_P[X] & E_P[XX^T] \end{pmatrix}.$$

In each case, Ω, and Ω^C are block-diagonal, so the inverse of the matrix is the inverse of each block, $\Omega^{-1} = I_2$, and $[\Omega^C]^{-1} = \text{diag}[1, E_P[(1,X^T)^T(1,X^T)]^{-1}]$. Recall from Lemma A.1 that $\beta^* = \Omega^{-1}E_P[ZY]$. Therefore, for the ANOVA and ANCOVA problem,

$$\beta^* \equiv (\Delta^*, \theta^*) = \Omega^{-1}E_P[(A,1)^TY] = \Big\{E_P[AY], E_P[Y]\Big\}^T$$

$$\beta^{*C} \equiv (\Delta^{*C}, \theta^{*C}, \gamma^{*C}) = [\Omega^C]^{-1}E_P[(A,1,X)^TY]$$

$$= \Big\{E_P[AY], E_P[(1,X^T)^T(1,X^T)]^{-1}E_P[(1,X^T)Y]\Big\}^T.$$

Thus, $\Delta^* = \Delta^{*C}$, and

$$\begin{aligned} \Delta^* &= E_P[AY] \\ &= E_P[AY|A=1]P[A=1] + E_P[AY|A=-1]P[A=-1] \\ &= E_P[Y|A=1]P[A=1] + E_P[-Y|A=-1]P[A=-1] \\ &= E_P[Y|A=1]P[A=1] - E_P[Y|A=-1]P[A=-1] \\ &= \frac{1}{2}\Big\{E_P[Y|A=1] - E_P[Y|A=-1]\Big\}. \end{aligned}$$

□

A.1.2 Asymptotic Behavior of ANOVA and ANCOVA

In this section, we derive the influence functions for ANOVA and ANCOVA to show that they are consistent and asymptotically normal estimators. Furthermore, we use the asymptotic forms to derive their variance. In these lemmas, we will use n as our sample size. In the main text, we instead use n_k^*, to be consistent with the group sequential trials notation. By Assumption 1.1, $n_k^* \to \infty$ as $n \to \infty$, so these results are equivalent.

Definition A.1 (Asymptotically Linear Estimator)**.** *An estimator $\hat{\psi}$ is an asymptotically linear estimator if it can be written as*

$$\hat{\psi}_n - \psi = \frac{1}{n}\sum_{i=1}^{n}\phi(O_i) + o_p\Big(n^{-\frac{1}{2}}\Big) \quad \text{where} \quad E_P[\phi] = 0.$$

We call ϕ the influence function for the estimator $\hat{\psi}$, that takes as an argument all of the data for a single observation O_i.

Lemma A.3 (Influence Function for OLS Estimator)**.** *Let Z_i be a vector of d covariates for the ith observation, $\tilde{Z} = (Z_1, ..., Z_n)^T$ be the design matrix of covariates and $\tilde{Y} = (Y_1, ..., Y_n)$ be the outcome vector. The multivariate influence function for the least squares estimator $\hat{\beta}$ of β^* is $\phi(O_i) = \Omega^{-1}Z_i(Y_i - Z_i^T\beta^*)$ where Ω is $E[ZZ^T]$. Furthermore, $\hat{\beta}$ is $\sqrt{n}$-consistent for β^*, defined in Theorem A.1.*

Proof. The least squares estimator for β^* is $\hat{\beta} = (\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}\tilde{Y}$, since it minimizes:

$$\hat{\beta} = \arg\min_{\beta\in\mathbb{R}^p}\sum_{i=1}^{n}(Y_i - Z_i^T\beta^*)^2.$$

We can see this by rewriting this in matrix form and taking the gradient with respect to β:

$$\frac{\partial}{\partial\beta}||\tilde{Y}-\tilde{Z}\beta||^2 = 2\tilde{Z}^T(\tilde{Y}-\tilde{Z}\beta) = 0$$
$$\tilde{Z}^T\tilde{Z}\beta = \tilde{Z}^T\tilde{Y}$$
$$\hat{\beta} = (\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}^T\tilde{Y}.$$

Define $E[ZZ^T] = \Omega$. Our goal is to find the multivariate influence function for $\hat{\beta}-\beta^*$. Define $\epsilon = Y - Z^T\beta$, or equivalently $Y = Z\beta + \epsilon$. Let $\tilde{\epsilon} = (\epsilon_1, ..., \epsilon_n)$ where $\epsilon_i = Y_i - Z_i^T\beta$. Using the form of $\hat{\beta}$, we have

$$\begin{aligned}
\hat{\beta} &= (\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}^T y \\
&= (\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}^T(\tilde{Z}\beta^* + \epsilon) \\
&= (\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}^T\tilde{Z}\beta^* + (\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}^T\epsilon \\
&= \beta^* + (\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}^T\tilde{\epsilon} \\
\implies \hat{\beta} - \beta^* &= (\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}^T\tilde{\epsilon}.
\end{aligned}$$

To derive the influence function, we want to find the right-hand side in terms of a sum of independent observations plus some second order term(s). We notice that $\tilde{Z}^T\tilde{Z}$ is the empirical second moment of Z, multiplied by n. We can rewrite the above in this form, and then add and subtract the true second moment to get it in terms of the population

parameters:

$$\begin{aligned}
\hat{\beta} - \beta^* &= \frac{1}{n}\left(\frac{\tilde{Z}^T\tilde{Z}}{n}\right)^{-1}\tilde{Z}^T\tilde{\epsilon} \\
&= \frac{1}{n}\left(\frac{\tilde{Z}^T\tilde{Z}}{n}\right)^{-1}\sum_{i=1}^{n}\tilde{Z}_i\epsilon_i \\
&= \frac{1}{n}\left(\frac{\tilde{Z}^T\tilde{Z}}{n}\right)^{-1}\sum_{i=1}^{n}Z_i\epsilon_i + \frac{1}{n}\Omega^{-1}\sum_{i=1}^{n}Z_i\epsilon_i - \frac{1}{n}\Omega^{-1}\sum_{i=1}^{n}Z_i\epsilon_i \\
&= \frac{1}{n}\Omega^{-1}\sum_{i=1}^{n}Z_i\epsilon_i + \left\{\left(\frac{\tilde{Z}^T\tilde{Z}}{n}\right)^{-1} - \Omega^{-1}\right\}\frac{1}{n}\sum_{i=1}^{n}Z_i\epsilon_i.
\end{aligned}$$

Focusing on the right-most term, we will show that

$$\left\{\left(\frac{\tilde{Z}^T\tilde{Z}}{n}\right)^{-1} - \Omega^{-1}\right\}\frac{1}{n}\sum_{i=1}^{n}Z_i\epsilon_i = o_p(n^{-\frac{1}{2}}). \tag{A.2}$$

By the WLLN, $\frac{1}{n}\sum_{i=1}^{n}Z_i\epsilon_i \to E_P[Z\epsilon]$ in probability. However,

$$\begin{aligned}
E_P[Z\epsilon] &= E_P[ZY - ZZ^T\beta^*] \\
&= E_P[ZY] - E_P[ZZ^T]\beta^* \\
&= E_P[ZY] - E_P[ZZ^T]E_P[ZZ^T]^{-1}E_P[ZY] \\
&= E_P[ZY] - E_P[ZY] = 0.
\end{aligned}$$

Thus, by the central limit theorem, $\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}Z_i\epsilon_i = O_P(1)$, which implies that $\frac{1}{n}\sum_{i=1}^{n}Z_i\epsilon_i = O_P(n^{-\frac{1}{2}})$. By the WLLN, $\frac{\tilde{Z}^T\tilde{Z}}{n} = \frac{1}{n}\sum_{i=1}^{n}Z_iZ_i^T \to E[ZZ^T] = \Omega$, which again by the central limit theorem means $(\frac{\tilde{Z}^T\tilde{Z}}{n} - \Omega) = O_p(n^{-\frac{1}{2}})$. Applying the continuous mapping theorem by noting that the inverse of a matrix is a continuous function, $((\tilde{Z}^T\tilde{Z}/n)^{-1} - \Omega^{-1}) = O_P(n^{-\frac{1}{2}})$. Using these two parts, (A.2) is $O_p(n^{-\frac{1}{2}})O_p(n^{-\frac{1}{2}}) = O_p(n^{-1}) = o_P(n^{-\frac{1}{2}})$, which means (A.2) is a second-order term, and can be ignored asymptotically. Further, note that $E[\phi] = 0$, since $E[\phi] = \Omega^{-1}E[Z\epsilon]$, which we have shown above to be 0. Therefore, the

multivariate influence function for $\hat{\beta}$ is

$$\phi(O_i) = \Omega^{-1} Z_i \epsilon_i = \Omega^{-1} Z_i (Y_i - Z_i^T \beta^*). \tag{A.3}$$

Since $\hat{\beta} - \beta^* = \frac{1}{n} \sum_{i=1}^n \phi(O_i) + o_P(n^{-\frac{1}{2}})$, we have $\hat{\beta} \to \beta^*$ in probability (by the WLLN since $E[\phi] = 0$, and applying Slutsky's Theorem). □

Lemma A.4 (Consistency of ANOVA and ANCOVA Estimators: Lemma 1.1(ii))**.** *Denote* $X = (X_1, ..., X_n)$, $A = (A_1, \ldots, A_n)^T$, *and* $Y = (Y_1, \ldots, Y_n)^T$. *Define* $\epsilon_i = Y_i - \theta^{*C} - \Delta^{*C} A_i - X_i^T \gamma^{*C}$ *and* $\tilde{\epsilon}_i = Y_i - \theta^* - \Delta^* A_i$ *as the sample versions of* ϵ *and* $\tilde{\epsilon}$. *Under Assumptions 1.2 and 1.3, the sample minimizers of* (1.3) *and* (1.4) *are*

$$(\hat{\Delta}, \hat{\theta}) = ((A,1)^T (A,1))^{-1} (A,1)^T Y$$
$$(\hat{\Delta}^C, \hat{\theta}^C, \hat{\gamma}^C) = ((A,1,X)^T (A,1,X))^{-1} (A,1,X)^T Y,$$

$\hat{\Delta}$ *and* $\hat{\Delta}^C$ *are asymptotically linear estimators with influence functions* $\phi(O_i) := A_i(\epsilon_i + \theta^{*C} - \theta^* + X_i^T \gamma^{*C})$ *and* $\phi^C(O_i) := A_i \epsilon_i$, *and* $\hat{\Delta}$ *and* $\hat{\Delta}^C$ *are* $\sqrt{n}$*-consistent estimators for* Δ^*.

Proof. We directly apply the results from Lemmas A.2 and A.3. Using the definitions of the ANOVA and ANCOVA problem and the definition of the OLS estimator from Lemma A.3, the sample minimizers in the ANOVA and ANCOVA problems are given by

$$(\hat{\Delta}, \hat{\theta}) = ((A,1)^T (A,1))^{-1} (A,1)^T Y$$
$$(\hat{\Delta}^C, \hat{\theta}^C, \hat{\gamma}^C) = ((A,1,X)^T (A,1,X))^{-1} (A,1,X)^T Y,$$

By Lemma A.3, $\phi(O_i) = \Omega^{-1} Z_i \epsilon_i$ where ϵ_i is the error term for the generic least squares problem. In the ANCOVA problem, we have $\epsilon_i = Y_i - \theta^{*C} - \Delta^{*C} A_i - X_i^T \gamma^{*C}$. In the ANOVA problem, we can rewrite the error term $\tilde{\epsilon}_i$:

$$\tilde{\epsilon}_i - \epsilon_i = (Y_i - Y_i) + (\theta^{*C} - \theta^*) + (\Delta^{*C} - \Delta^*) A_i + X_i^T \gamma^{*C}.$$

However, by Lemma A.2, we know that $\Delta^{*C} = \Delta^*$. So we have $\tilde{\epsilon}_i = \epsilon_i + (\theta^{*C} - \theta^*) + X_i^T \gamma^{*C}$. Then the multivariate influence functions have the form:

$$\phi(O_i) = \Omega^{-1}(A_i, 1)^T \tilde{\epsilon}_i = \Omega^{-1}(A_i, 1)^T (\epsilon_i + \theta^{*C} - \theta^* + X_i^T \gamma^{*C})$$
$$\phi^C(O_i) = [\Omega^C]^{-1}(A_i, 1, X_i^T)^T \epsilon_i$$

and the influence functions for just the Δ^* component of the regression parameters, corresponding to the $\Omega^{-1}_{1,1} = 1$ component, are $\phi(O_i) = A_i(\epsilon_i + \theta^{*C} - \theta^* + X_i^T \gamma^{*C})$ and $\phi^C(O_i) = A_i \epsilon_i$. Finally, by Lemma A.3, $\hat{\Delta}^C$ and $\hat{\Delta}$ are both $\sqrt{n}$-consistent for Δ^*. □

Lemma A.5 (Asymptotic Variance of ANOVA and ANCOVA Estimators: Lemma 1.1(iii))**.** *Define* $\sigma^2 = \text{Var}[\epsilon]$ *where* $\epsilon = Y - \theta^{*C} - \Delta^{*C} A - X^T \gamma^{*C}$ *as in Assumption 1.3. Then under Assumptions 1.2 and 1.3,* $\text{Var}[(\phi, \phi^C)^T] = \left(\begin{smallmatrix} \tilde{\sigma}^2 & \sigma^2 \\ \sigma^2 & \sigma^2 \end{smallmatrix}\right)$ *where* $\tilde{\sigma}^2 = \sigma^2 + \text{Var}[(X^T \gamma^{*C})^2]$.

Proof. We calculate the asymptotic variance of the ANOVA and ANCOVA estimators based on their influence functions from Lemma A.4. As in Lemma A.4, define the random variables $\epsilon = Y - \theta^{*C} - A\Delta^{*C} - X^T \gamma^{*C}$ and $\tilde{\epsilon} = Y - \theta^* - A\Delta^*$. Then by the same logic as Lemma A.4, we have $\tilde{\epsilon} - \epsilon = (\theta^{*C} - \theta^*) + X^T \gamma^{*C}$. By definition of the population minimizer, $(\theta^{*C}, \Delta^{*C}, \gamma^{*C})$ and (θ^*, Δ^*) solve the population score equations for ANCOVA and ANOVA, respectively. The score equations have expectation zero, hence, $E[(1, A, X^T)^T (Y - \theta^{*C} - A\Delta^{*C} - X^T \gamma^{*C})] =$

0 and $E[(1,A)^T(Y-\theta^*-A\Delta^*)]=0$. Let $\mathrm{Var}[\epsilon]=\sigma^2$. Then,

$$
\begin{aligned}
&\mathrm{Var}[\phi^C] \\
&= \mathrm{Var}[A\epsilon] \\
&= E[A^2\epsilon^2] - E[A(Y-\theta^{*C}-\Delta^{*C}A-X^T\gamma^{*C})]^2 \\
&= E[\epsilon^2] = \mathrm{Var}[\epsilon] = \sigma^2 \\
&\mathrm{Var}[\phi] \\
&= \mathrm{Var}[A(\theta^{*C}-\theta^*+X^T\gamma^{*C}+\epsilon)] \\
&= \mathrm{Var}[A\epsilon] + \mathrm{Var}[A(\theta^{*C}-\theta^*+X^T\gamma^{*C})] + 2\mathrm{Cov}[A\epsilon, A(\theta^{*C}-\theta^*+X^T\gamma^{*C})] \\
&= \sigma^2 + \mathrm{Var}[E(A(\theta^{*C}-\theta^*+X^T\gamma^{*C})|A)] + E[\mathrm{Var}(A(\theta^{*C}-\theta^*+X^T\gamma^{*C})|A)] \\
&\qquad + 2\mathrm{Cov}[A\epsilon, A(\theta^{*C}-\theta^*+X^T\gamma^{*C})] \\
&= \sigma^2 + \mathrm{Var}[A(E(\theta^{*C}-\theta^*+X^T\gamma^{*C})] + E[\mathrm{Var}(\theta^{*C}-\theta^*+X^T\gamma^{*C})] + 2\mathrm{Cov}[A\epsilon, A(\theta^{*C}-\theta^*+X^T\gamma^{*C})] \\
&= \sigma^2 + E[\theta^{*C}-\theta^*+X^T\gamma^{*C}]^2 + \mathrm{Var}[\theta^{*C}-\theta^*+X^T\gamma^{*C}] + 2\mathrm{Cov}[A\epsilon, A(\theta^{*C}-\theta^*+X^T\gamma^{*C})] \\
&= \sigma^2 + E[\tilde{\epsilon}-\epsilon]^2 + \mathrm{Var}[X^T\gamma^{*C}] + 2\mathrm{Cov}[A\epsilon, A(\theta^{*C}-\theta^*+X^T\gamma^{*C})] \\
&= \sigma^2 + \mathrm{Var}[X^T\gamma^{*C}] + 2\left\{E[A^2\epsilon(\theta^{*C}-\theta^*+X^T\gamma^{*C})] - E[A\epsilon]E[A(\theta^{*C}-\theta^*+X^T\gamma^{*C})]\right\} \\
&= \sigma^2 + \mathrm{Var}[X^T\gamma^{*C}] + 2E[\epsilon(\theta^{*C}-\theta^*+X^T\gamma^{*C})] \\
&= \sigma^2 + \mathrm{Var}[X^T\gamma^{*C}] + 2(\gamma^{*C})^T E[X\epsilon] \\
&= \sigma^2 + \mathrm{Var}[X^T\gamma^{*C}] \\
&\mathrm{Cov}[\phi, \phi^C] \\
&= \mathrm{Cov}[A\epsilon, A(X^T\gamma+\epsilon)] \\
&= E[A^2\epsilon(X^T\gamma+\epsilon)] - E[A\epsilon]E[A(Y-\theta^{*C}-\Delta^{*C}A)] \\
&= E[\epsilon X^T\gamma^{*C}] + E[\epsilon^2] \\
&= (\gamma^{*C})^T E[X(Y-\theta^{*C}-\Delta^{*C}A-X^T\gamma^{*C})] + \mathrm{Var}[\epsilon] \\
&= \mathrm{Var}[\epsilon] = \sigma^2
\end{aligned}
$$

□

Let $\rho^2 = \sigma^2/\tilde{\sigma}^2$, i.e., ρ^2 is the ratio of variance of ANCOVA to ANOVA. If $\rho = 1$, then $\text{Var}[(X^T\gamma^{*C})^2]$ must be 0, i.e., there is no relationship between the covariates and the outcome.

Remark A.1. *The asymptotic variance of the ANCOVA estimator is no larger than the asymptotic variance of the ANOVA estimator* $(0 < \rho \leq 1)$*, and the reduction in variance by using ANCOVA is* $\text{Var}[(X^T\gamma^{*C})^2]$*. The reduction in variance by using ANCOVA is a function of how predictive the covariates* X *are with respect to the outcome* Y*.*

A.2 The Hybrid Design

We use the results from the previous review lemmas in Section A.1 on covariate adjustment to show our desired results for the hybrid design.

A.2.1 Joint Distribution of ANOVA and ANCOVA Test Statistics in Group Sequential Design

Lemma A.6. *Let* $n_k^* = \sum_{k'=1}^{k} n_{k'}$*, the sample size accumulated up through stage* k*, with* $n = n_K^*$*, as defined in Assumption 1.1. Consider the influence functions for ANOVA and ANCOVA* ϕ *and* ϕ^C *from Lemma A.4, and let*

$$\bar{\phi}_k^* := \frac{1}{n_k^*}\sum_{i=1}^{n_k^*}\phi(O_i) \equiv \frac{1}{n_k^*}\sum_{i=1}^{n_k^*} A_i(X_i^T\gamma^{*C} + \theta^{*C} - \theta^* + \epsilon_i)$$
$$\bar{\phi}_k^{*C} := \frac{1}{n_k^*}\sum_{i=1}^{n_k^*}\phi^C(O_i) \equiv \frac{1}{n_k^*}\sum_{i=1}^{n_k^*} A_i\epsilon_i.$$

Then under Assumptions 1.2 and 1.3,

$$\text{Cov}[\sqrt{n}\bar{\phi}_k^*, \sqrt{n}\bar{\phi}_{k'}^*] = t_{\max(k,k')}^{*-1}\tilde{\sigma}^2$$
$$\text{Cov}[\sqrt{n}\bar{\phi}_k^{*C}, \sqrt{n}\bar{\phi}_{k'}^*] = \text{Cov}[\sqrt{n}\bar{\phi}_k^{*C}, \sqrt{n}\bar{\phi}_{k'}^{*C}] = t_{\max(k,k')}^{*-1}\sigma^2$$

Proof of Lemma A.6. We make use of the variance of the influence functions that we derived in Lemma A.5. Recall that $\text{Var}[\phi] = \tilde{\sigma}^2$ and $\text{Var}[\phi^C] = \sigma^2$. Then,

$$\begin{aligned}
\text{Cov}[\sqrt{n}\bar{\phi}_k^*, \sqrt{n}\bar{\phi}_k^{*C}] &= n\text{Cov}\left[\frac{1}{n_k^*}\sum_{i=1}^{n_k^*}\phi(O_i), \frac{1}{n_{k'}}\sum_{i=1}^{n_{k'}^*}\phi^C(O_i)\right] \\
&= \frac{n}{n_k^* n_{k'}^*}\sum_{i=1}^{n_k^*}\sum_{j=1}^{n_{k'}^*}\text{Cov}[\phi(O_i), \phi^C(o_j)] \\
&= \frac{n}{n_k^* n_{k'}^*}\sum_{i=1}^{\min(n_k^*, n_{k'}^*)}\text{Cov}[\phi(O_i), \phi^C(O_i)] \\
&= \frac{n}{\max(n_k^*, n_{k'}^*)}\text{Cov}[\phi(O_i), \phi^C(O_i)] = t^{*-1}_{\max(k,k')}\sigma^2.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\text{Cov}[\sqrt{n}\bar{\phi}_k^*, \sqrt{n}\bar{\phi}_k^*] &= \frac{n}{\max(n_k^*, n_{k'}^*)}\text{Cov}[\phi(O_i), \phi(O_i)] = t^{*-1}_{\max(k,k')}\tilde{\sigma}^2 \\
\text{Cov}[\sqrt{n}\bar{\phi}_k^{*C}, \sqrt{n}\bar{\phi}_k^{*C}] &= \frac{n}{\max(n_k^*, n_{k'}^*)}\text{Cov}[\phi^C(O_i), \phi^C(O_i)] = t^{*-1}_{\max(k,k')}\sigma^2.
\end{aligned}$$

□

Lemma A.7 (Asymptotic Joint Distribution of Empirical Averages of Influence Functions)**.** *Consider $\bar{\phi}_k^*$ and $\bar{\phi}_k^{*C}$ from Lemma A.6. Define $\bar{\phi}^* := (\bar{\phi}_1^*, \bar{\phi}_1^{*C}..., \bar{\phi}_K^*, \bar{\phi}_K^{*C})^T \in \mathbb{R}^{2K}$. Then under Assumptions 1.1-1.3, $\sqrt{n}\bar{\phi}^*$ converges in distribution to a multivariate normal, i.e.,*

$$\sqrt{n}\bar{\phi}^* \Rightarrow N(0, \Sigma_{\bar{\phi}^*}) \qquad \Sigma_{\bar{\phi}^*} = T \otimes \text{Var}[(\phi, \phi^C)^T]$$

where $T_{k,k'} = t^{*-1}_{\max(k,k')}$.

Proof. Define $\bar{\phi} = (\bar{\phi}_1, \bar{\phi}_1^C, ..., \bar{\phi}_K, \bar{\phi}_K^C) \in \mathbb{R}^{2K}$. This is the vector of the empirical averages of the influence functions for only the observations obtained in each stage (as opposed to the cumulative vector $\bar{\phi}^*$). Let $n_0 = 0$. We will start by showing that (i) $\sqrt{n}\bar{\phi} \Rightarrow N(0, \Sigma_{\bar{\phi}})$, and (ii) then will multiply it by a matrix in order to get the cumulative vector $\sqrt{n}\bar{\phi}^*$.

Starting with (i), for each element of $\bar{\phi}$,

$$\begin{aligned}\sqrt{n}\bar{\phi}_k &= \sqrt{n}\frac{1}{n_k}\sum_{i=n_{(k-1)}+1}^{n_k}\phi(O_i)\\ &= \frac{1}{\sqrt{n_k}}\sum_{i=n_{(k-1)}+1}^{n_k}t_k^{-1/2}\phi(O_i).\end{aligned}$$

Similarly, $\sqrt{n}\bar{\phi}_k^C = \frac{1}{\sqrt{n_k}}\sum_{i=n_{(k-1)}+1}^{n_k} t_k^{-1/2}\phi^C(O_i)$. By the CLT, we know that

$$\begin{aligned}\sqrt{n}(\phi_k,\phi_k^C) &\Rightarrow N\left(0,\begin{pmatrix}\mathrm{Var}(t_k^{-1/2}\phi) & \mathrm{Cov}(t_k^{-1/2}\phi, t_k^{-1/2}\phi^C)\\ \mathrm{Cov}(t_k^{-1/2}\phi, t_k^{-1/2}\phi^C) & \mathrm{Var}(t_k^{-1/2}\phi^C)\end{pmatrix}\right)\\ &= N\left(0,\begin{pmatrix}t_k^{-1}\tilde{\sigma}^2 & t_k^{-1}\sigma^2\\ t_k^{-1}\sigma^2 & t_k^{-1}\sigma^2\end{pmatrix}\right)\\ &= N\Big(0, t_k^{-1}\mathrm{Var}[(\phi,\phi^C)^T]\Big).\end{aligned}$$

By the Cramer-Wold Theorem, $\sqrt{n}\bar{\phi} \Rightarrow N(0,\Sigma_{\bar{\phi}})$ if for all $a\in\mathbb{R}^{2K}$, $a^T\sqrt{n}\bar{\phi} \Rightarrow N(0, a^T\Sigma_{\bar{\phi}}a)$. Let $b,c\in\mathbb{R}^K$, and $a = (a_1^T,...,a_K^T)^T = (b_1,c_1,...,b_K,c_K)^T$ with $a_k = (b_k,c_k)^T$,. Since each of the $(\phi_k,\phi_k^C) \perp (\phi_{k'},\phi_{k'}^C)$ when $k\neq k'$ (since they share no observations), the covariance matrix $\Sigma_{\bar{\phi}}$ is

$$\Sigma_{\bar{\phi}} = \mathrm{diag}t_1^{-1},...,t_K^{-1}\otimes\mathrm{Var}[(\phi,\phi^C)^T],$$

i.e., $\Sigma_{\bar{\phi}}$ is a block-diagonal matrix with the k^{th} 2×2 diagonal block equal to $t_k^{-1}\mathrm{Var}[(\phi,\phi^C)^T]$. Our goal is to show that $a^T\sqrt{n}\bar{\phi}$ converges in distribution to a univariate normal distribution

with covariance matrix $a^T\Sigma_{\bar{\phi}}a$, which is

$$\begin{aligned}
a^T\Sigma_{\bar{\phi}}a &= (a_1^T, ..., a_K^T)\text{diag}t_1^{-1}\text{Var}[(\phi, \phi^C)^T], ..., t_K^{-1}\text{Var}[(\phi, \phi^C)^T](a_1^T, ..., a_K^T)^T \\
&= \sum_{k=1}^{K} t_k^{-1} a_k^T \text{Var}[(\phi, \phi^C)^T] a_k \\
&= \sum_{k=1}^{K} t_k^{-1}\big(b_k^2\tilde{\sigma}^2 + c_k^2\sigma^2 + 2b_kc_k\sigma^2\big).
\end{aligned}$$

To show that $a^T\sqrt{n}\bar{\phi}$ converges in distribution to a univariate normal distribution, we note that

$$\begin{aligned}
a^T\sqrt{n}\bar{\phi} &= a_1^T\sqrt{n}(\bar{\phi}_1, \bar{\phi}_1^C)^T, ..., a_K^T\sqrt{n}(\bar{\phi}_K, \bar{\phi}_K^C)^T \\
&= \sum_{k=1}^{K}\Big\{b_k\frac{1}{\sqrt{n_k}}\sum_{i=n_{(k-1)}+1}^{n_k} t_k^{-1/2}\phi(O_i) + c_k\frac{1}{\sqrt{n_k}}\sum_{i=n_{(k-1)}+1}^{n_k} t_k^{-1/2}\phi^C(O_i)\Big\} \\
&= \sum_{k=1}^{K}\Big\{\frac{1}{\sqrt{n_k}}\sum_{i=n_{(k-1)}+1}^{n_k} t_k^{-1/2}\big\{b_k\phi(O_i) + c_k\phi^C(O_i)\big\}\Big\}.
\end{aligned}$$

Define $S_{k,i} = t_k^{-1/2}\big\{b_k\phi(O_i) + c_k\phi^C(O_i)\big\}$ and $S_k = \frac{1}{\sqrt{n_k}}\sum_{i=n_{(k-1)}+1}^{n_k} S_{k,i}$. Now, it is clear that $a^T\sqrt{n}\bar{\phi}$ can be written as a sum of K independent random variables, i.e., $a^T\sqrt{n}\bar{\phi} = \sum_{k=1}^{K} S_k$. Note that $S_{k,i}$, $i = n_{(k-1)} - 1, ..., n_k$ are i.i.d. terms, thus, by the Central Limit Theorem,

$$\begin{aligned}
S_k &= \frac{1}{\sqrt{n_k}}\sum_{i=n_{(k-1)}+1}^{n_k} t_k^{-1/2}\big\{b_k\phi(O_i) + c_k\phi^C(O_i)\big\} \\
&\Rightarrow N(0, \text{Var}[S_{k,i}]) \\
&= N\Big(0, t_k^{-1}\big\{b_k^2\text{Var}[\phi] + c_k^2\text{Var}[\phi^C] + 2b_kc_k\text{Cov}[\phi, \phi^C]\big\}\Big) \\
&= N\Big(0, t_k^{-1}\big(b_k^2\tilde{\sigma}^2 + c_k^2\sigma^2 + 2b_kc_k\sigma^2\big)\Big).
\end{aligned}$$

By Levy's Continuity Theorem (applying characteristic functions), if $S_1, ..., S_K$ are independent random variables, and each $S_k \Rightarrow S_k' \sim N(0, \text{Var}[S_k])$, then $S_1 + \cdots + S_K \Rightarrow S' :=$

$S'_1 + \cdots + S'_K$ with $S' \sim N(0, \sum_{k=1}^K \text{Var}[S_k])$. Putting this together with the above,

$$\begin{aligned} a^T\sqrt{n}\bar{\phi} &= \sum_{k=1}^K S_k \\ &\Rightarrow N\Big(0, \sum_{k=1}^K \text{Var}[S_k]\Big) \\ &= N\Big(0, \sum_{k=1}^K t_k^{-1}\big(b_k^2\tilde{\sigma}^2 + c_k^2\sigma^2 + 2b_kc_k\sigma^2\big)\Big) \\ &= N(0, a^T\Sigma_{\bar{\phi}}a). \end{aligned}$$

Thus, by the Cramer-Wold Theorem, $\sqrt{n}\bar{\phi} \Rightarrow N(0, \Sigma_{\bar{\phi}})$, and this completes (i).

Now for (ii), we apply a linear transformation to $\sqrt{n}\bar{\phi}$ in order to obtain $\sqrt{n}\bar{\phi}^*$. Notice that since $\sqrt{n}\bar{\phi}_k = \frac{\sqrt{n}}{n_k}\sum_{i=n_{(k-1)}+1}^{n_k}\phi(O_i)$ (and similarly for $\bar{\phi}_k^C$),

$$\begin{aligned} \sqrt{n}\bar{\phi}_k^* &= \frac{n_1}{n_k^*}\sqrt{n}\bar{\phi}_1 + \cdots + \frac{n_k}{n_k^*}\sqrt{n}\bar{\phi}_k \\ \sqrt{n}\bar{\phi}_k^{*C} &= \frac{n_1}{n_k^*}\sqrt{n}\bar{\phi}_1^C + \cdots + \frac{n_k}{n_k^*}\sqrt{n}\bar{\phi}_k^C. \end{aligned}$$

Therefore, we can obtain $\sqrt{n}\bar{\phi}^*$ with the linear transformation $\sqrt{n}\bar{\phi}^* = (L \otimes I_2)\sqrt{n}\bar{\phi}$, where L is a lower triangular matrix with entries

$$L = \begin{pmatrix} \frac{n_1}{n_1^*} & 0 & \dots & 0 \\ \frac{n_1}{n_2^*} & \frac{n_2}{n_2^*} & \dots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ \frac{n_1}{n_K^*} & \cdots & \cdots & \frac{n_K}{n_K^*} \end{pmatrix}$$

We need the Kronecker product with I_2 (the 2×2 identity matrix) here because we only want to select the ANOVA entries $\bar{\phi}_k$ for the cumulative $\bar{\phi}_k^*$, and likewise only the ANCOVA entries $\bar{\phi}_k^C$ for the cumulative $\bar{\phi}_k^{*C}$. It will be easier to work with information fractions, so we can write the $(k, k')^{th}$ entry of L (with $k \leq k'$ since it is lower triangular) as $t_k/t_{k'}^*$. By the

continuous mapping theorem, we know that

$$\sqrt{n}\bar{\phi}^* = (L \otimes I_2)\sqrt{n}\bar{\phi} \Rightarrow N\big(0, (L \otimes I_2)\Sigma_{\bar{\phi}}(L \otimes I_2)^T\big).$$

Define $\Sigma_{\bar{\phi}^*} := (L \otimes I_2)\Sigma_{\bar{\phi}}(L \otimes I_2)^T$. Then, using the matrix product property of the Kronecker product (see Theorem 3 of Zhang and Ding (2013)),

$$\begin{aligned}
\Sigma_{\bar{\phi}^*} &= (L \otimes I_2)\Sigma_{\bar{\phi}}(L \otimes I_2)\big) \\
&= (L \otimes I_2)(\text{diag}t_1^{-1}, ..., t_K^{-1} \otimes \text{Var}[(\phi, \phi^C)^T])(L \otimes I_2)^T \\
&= (L\text{diag}t_1^{-1}, ..., t_K^{-1}L^T) \otimes (I_2\text{Var}[(\phi, \phi^C)^T]I_2) \\
&= (L\text{diag}t_1^{-1}, ..., t_K^{-1}L^T) \otimes \text{Var}[(\phi, \phi^C)^T].
\end{aligned}$$

We will show that $T = (L\text{diag}t_1^{-1}, ..., t_K^{-1}L^T)$ where $T_{k,k'} = t^{*-1}_{\max(k,k')}$.

$$
\begin{aligned}
L\text{diag}t_1^{-1}, ..., t_K^{-1}L^T &= \begin{pmatrix} \frac{n_1}{n_1^*} & 0 & \dots & 0 \\ \frac{n_1}{n_2^*} & \frac{n_2}{n_2^*} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ \frac{n_1}{n_K^*} & \dots & \dots & \frac{n_K}{n_K^*} \end{pmatrix} \text{diag}t_1^{-1}, ..., t_K^{-1} \begin{pmatrix} \frac{n_1}{n_1^*} & \frac{n_1}{n_2^*} & \dots & \frac{n_1}{n_K^*} \\ 0 & \frac{n_2}{n_2^*} & \dots & \dots \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{n_K}{n_K^*} \end{pmatrix} \\
&= \begin{pmatrix} \frac{n_1}{n_1^*}t_1^{-1} & 0 & \dots & 0 \\ \frac{n_1}{n_2^*}t_1^{-1} & \frac{n_2}{n_2^*}t_2^{-1} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ \frac{n_1}{n_K^*}t_1^{-1} & \dots & \dots & \frac{n_K}{n_K^*}t_K^{-1} \end{pmatrix} \begin{pmatrix} \frac{n_1}{n_1^*} & \frac{n_1}{n_2^*} & \dots & \frac{n_1}{n_K^*} \\ 0 & \frac{n_2}{n_2^*} & \dots & \dots \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{n_K}{n_K^*} \end{pmatrix} \\
&= \begin{pmatrix} \frac{n_1^2}{n_1^{*2}}t_1^{-1} & \frac{n_1^2}{n_1^*n_2^*}t_1^{-1} & \dots & \frac{n_1^2}{n_1^*n_K^*}t_1^{-1} \\ \frac{n_1^2}{n_1^*n_2^*}t_1^{-1} & \frac{n_1^2t_1^{-1}+n_2^2t_2^{-1}}{n_2^{*2}} & \dots & \frac{n_1^2t_1^{-1}+n_2^2t_2^{-1}}{n_2^*n_K^*} \\ \dots & \dots & \dots & \dots \\ \frac{n_1^2}{n_1^*n_K^*}t_1^{-1} & \frac{n_1^2t_1^{-1}+n_2^2t_2^{-1}}{n_2^*n_K^*} & \dots & \frac{\sum_{k=1}^K n_k^{*2}t_k^{-1}}{n_K^{*2}} \end{pmatrix} \\
&= \begin{pmatrix} \frac{n}{n_1^*} & \frac{n}{n_2^*} & \dots & \frac{n}{n_K^*} \\ \frac{n}{n_2^*} & \frac{n}{n_2^*} & \dots & \frac{n}{n_K^*} \\ \dots & \dots & \dots & \dots \\ \frac{n}{n_K^*} & \frac{n}{n_K^*} & \dots & \frac{n}{n_K^*} \end{pmatrix} = \begin{pmatrix} t_1^{*-1} & t_2^{*-1} & \dots & t_K^{*-1} \\ t_2^{*-1} & t_2^{*-1} & \dots & t_K^{*-1} \\ \dots & \dots & \dots & \dots \\ t_K^{*-1} & t_K^{*-1} & \dots & t_K^{*-1} \end{pmatrix} = T,
\end{aligned}
$$

where we have used the fact that for $k' \leq k''$, $\frac{\sum_{k=1}^{k'} n_k^{*2}t_k^{-1}}{n_{k'}^*n_{k''}^*} = \frac{\sum_{k=1}^{k'} n \cdot n_k}{n_{k'}^*n_{k''}^*} = \frac{nn_{k'}^*}{n_{k'}^*n_{k''}^*} = \frac{n}{n_{k''}^*}$. □

Proof of Theorem 1.1. We now prove convergence in distribution for the full influence function, rather than just the first-order approximation. Define $\bar{\phi}^* := (\bar{\phi}_1^*, \bar{\phi}_1^{*C} ..., \bar{\phi}_K^*, \bar{\phi}_K^{*C}.)^T$ as in Lemma A.7. Let $r := \left(o_p(n^{-\frac{1}{2}}), ..., o_p(n^{-\frac{1}{2}})\right)^T \in \mathbb{R}^{2K}$.

By the Cramr-Wold theorem, $\sqrt{n}(\hat{\Delta} - \Delta^*) \Rightarrow N\big(0, \Sigma_{\bar{\phi}^*}\big)$ if and only if $\forall a \in \mathbb{R}^{2K}$, $a^T\sqrt{n}(\hat{\Delta} - \Delta^*) \Rightarrow N\big(0, a^T\Sigma_{\bar{\phi}^*}a\big)$, i.e., we can prove convergence in distribution by proving that all linear combinations of the vector converge in distribution to a univariate Gaussian distribution. By the definition of asymptotic linearity, $\sqrt{n}(\bar{\phi}^* + r) = \sqrt{n}(\hat{\Delta} - \Delta^*)$. We want

to show that all linear combinations of $\sqrt{n}(\bar{\phi}^* + r)$ converge in distribution, because that implies that $\sqrt{n}(\hat{\Delta} - \Delta^*)$ converges in distribution.

Take any $a \in \mathbb{R}^k$. Then $a^T(\bar{\phi}^* + r) = a^T\bar{\phi}^* + a^T r$. From Lemma A.7, $\sqrt{n}\bar{\phi}^* \Rightarrow N(0, \Sigma_{\bar{\phi}^*})$, and by the properties of the multivariate Gaussian distribution, $a^T\sqrt{n}(\bar{\phi}^* - \phi) \Rightarrow N(0, a^T\Sigma_{\bar{\phi}^*}a)$. $a^T r = \sum_{i=1}^{2K} a_i o_p(1)$. Since each element of the summation is $o_p(1)$, and since the summation has a finite number of elements, then $a^T r = o_p(1)$, i.e., $a^T r \to 0$. By Slutsky's theorem, $a^T\bar{\phi}^* + a^T r \Rightarrow N(0, a^T\Sigma_{\bar{\phi}^*}a)$, which, since this holds $\forall a$, by the Cramr-Wold theorem, $\sqrt{n}(\hat{\Delta} - \Delta^*) \Rightarrow N\Big(0, \Sigma_{\bar{\phi}^*}\Big)$.

□

Proof of Corollary 1.1. Define $M := \text{diag}\sqrt{t_1^*}, ..., \sqrt{t_k^*}$ and $B := \text{diag}\tilde{\sigma}^{-1}, \sigma^{-1}$, and $D := M \otimes B$. Recognize that D is a symmetric matrix, and that

$$D = \text{diag}\tilde{\sigma}^{-1}\sqrt{t_1^*}, \sigma^{-1}\sqrt{t_1^*}, ..., \tilde{\sigma}^{-1}\sqrt{t_k^*}, \sigma^{-1}\sqrt{t_k^*}.$$

Let $\hat{D}$ be the same as D but with kth diagonal entry as $\hat{\sigma}_k^{-1}\sqrt{t_k^*}$ and the $(k+1)$th diagonal entry as $(\hat{\sigma}_k^C)^{-1}\sqrt{t_1^*}$. That is, $\hat{D}$ is D where the standard error estimates have replaced the asymptotic standard error. By Wang et al. (2019), $\hat{D} \to D$ in probability since $\hat{\sigma}_k$ and $\hat{\sigma}_k^C$ are consistent estimators of their population counterparts. Note that the asymptotic variance of $\sqrt{n}\hat{\Delta}_k$ is $\tilde{\sigma}^2 t_k^{*-1}$, and the asymptotic variance of $\sqrt{n}\hat{\Delta}_k^C$ is $\sigma^2 t_k^{*-1}$, so multiplying by D will standardize the random variable $\hat{\Delta}$. In particular, for an ANOVA estimate at stage k, $Z_k^* = \sqrt{n_k}(\hat{\Delta}_k - \Delta^*)/\hat{\sigma}_k$, and for an ANCOVA estimate at stage k, $Z_k^{*C} = \sqrt{n_k}(\hat{\Delta}_k - \Delta^*)/\hat{\sigma}_k^C$. Since $\sqrt{t_k^*} = \sqrt{n_k^*}/\sqrt{n}$, multiplying $\sqrt{n}(\hat{\Delta} - \Delta^*)$ by $\hat{D}$ will get us a vector of suitably standardized test statistics Z^*.

To find the distribution of Z^*, we have by Slutsky's theorem with the multivariate normal

$$Z^* = \hat{D}\sqrt{n}(\hat{\Delta} - \Delta^*) \Rightarrow N(0, D\Sigma_{\hat{\Delta}}D^T).$$

Z^* is a vector of suitably standardized test statistics under the null hypothesis that $\Delta^* = 0$.

We now derive the form of the variance of Z^*, i.e., $\Sigma_{Z^*} := D\Sigma_{\hat{\Delta}}D^T$.

Recall that $\Sigma_{\hat{\Delta}} = T \otimes \text{Var}[(\phi, \phi^C)^T]$. Define $V := \text{Var}[(\phi, \phi^C)^T]$. Then, using properties of the Kronecker product (see Theorem 3 of Zhang and Ding (2013)),

$$\begin{aligned} D\Sigma_{\hat{\Delta}}D^T &= (M \otimes B)(T \otimes V)(M \otimes B) \\ &= ((MT) \otimes (BV))(M \otimes B) \\ &= (MTM) \otimes (BVB). \end{aligned}$$

We now compute (MTM) and (BVB). We have

$$\begin{aligned} MTM &= \text{diag}\sqrt{t_1^*}, ..., \sqrt{t_K^*} \begin{pmatrix} t_1^{*,-1} & t_2^{*,-1} & \dots & t_K^{*,-1} \\ t_2^{*,-1} & t_2^{*,-1} & \dots & t_K^{*,-1} \\ \dots & \dots & \dots & \dots \\ t_K^{*,-1} & t_K^{*,-1} & \dots & t_K^{*,-1} \end{pmatrix} \text{diag}\sqrt{t_1^*}, \dots, \sqrt{t_K^*} \\ &= \begin{pmatrix} 1 & \sqrt{t_1^*}/\sqrt{t_2^*} & \sqrt{t_1^*}/\sqrt{t_3^*} & \dots & \sqrt{t_1^*}/\sqrt{t_K^*} \\ \sqrt{t_1^*}/\sqrt{t_2^*} & 1 & \sqrt{t_2^*}/\sqrt{t_3^*} & \dots & \dots \\ \sqrt{t_1^*}/\sqrt{t_3^*} & \sqrt{t_2^*}/\sqrt{t_3^*} & 1 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \sqrt{t_1^*}/\sqrt{t_K^*} & \dots & \dots & 1 & \end{pmatrix} \\ BVB &= \begin{pmatrix} \tilde{\sigma}^{-1} & 0 \\ 0 & \sigma^{-1} \end{pmatrix} \begin{pmatrix} \tilde{\sigma}^2 & \sigma^2 \\ \sigma^2 & \sigma^2 \end{pmatrix} \begin{pmatrix} \tilde{\sigma}^{-1} & 0 \\ 0 & \sigma^{-1} \end{pmatrix} \\ &= \begin{pmatrix} 1 & \sigma/\tilde{\sigma} \\ \sigma/\tilde{\sigma} & 1 \end{pmatrix} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}. \end{aligned}$$

Putting them together, we have $(MTM) \otimes (BVB) = T' \otimes \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$, i.e.,

$$Z^* \Rightarrow N\Big(0, T' \otimes \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)\Big)$$

where T' has entries $T'_{k,k'} = \sqrt{\frac{t^*_{\min(k,k')}}{t^*_{\max(k,k')}}}$. □

Note: Under an alternative hypothesis, $\Delta^* \neq 0$. We can instead have Δ^* be the treatment effect under the alternative hypothesis we are considering, for example, to construct adjusted point estimates and confidence intervals as in Section 1.3.3. Instead of comparing the test statistic to the mean-zero multivariate Gaussian above, we compare our test statistic minus $\sqrt{n}\hat{D}\Delta^*$ to the mean-zero multivariate Gaussian above (or equivalently to a multivariate Gaussian with the same variance as above but with mean $\sqrt{n}\hat{D}\Delta^*$).

A.2.2 Interim Monitoring in the Hybrid Design

For background details on interim monitoring in a group sequential design, see Gernot and Brannath (2016). Briefly, monitoring boundaries for common forms like Pocock (Pocock, 1977) or O'Brien-Fleming (O'Brien and Fleming, 1979) can be derived using the joint distribution of test statistics in the group sequential design. Specifically, the constant c is chosen such that

$$\lim_{n\to\infty} P(|Z_1^*| \geq c \text{ or } ... \text{ or } |Z_K^*| \geq c) = \alpha$$

for the Pocock design, and

$$\lim_{n\to\infty} P(|Z_1^*| \geq c \text{ or } ... \text{ or } |Z_K^*| \geq c/\sqrt{K}) = \alpha$$

for the O'Brien-Fleming design. In the standard group sequential (non-hybrid) design, under the null hypothesis, the joint distribution of $(Z_1^*, ..., Z_K^*)$ converges to a mean-zero multivariate Gaussian distribution with a covariance matrix determined by the information fractions. In Theorem 1.1 and Corollary 1.1, we showed that we can similarly characterize the joint distribution of the test statistics under the null hypothesis for the hybrid design proposed in 1.1. Therefore, to solve for the constants in Corollaries 1.1 and 1.2, we are able to use the same type of strategy, but where we have modified the joint distribution of the test statistics to reflect the hybrid design.

Proof of Examples 1.1 and 1.2. In Corollary 1.1, we characterized the asymptotic joint distribution of the vector $Z^* = (Z_1^*, Z_1^{*C}, ..., Z_K^*, Z_K^{*C})^T$. Since sub-vectors of multivariate Gaussian vectors are also multivariate Gaussian, we can simply select the covariance elements corresponding to a sequence of un-adjusted and covariate-adjusted estimators in line with the hybrid design propsed in 1.1. Specifically, the test statistic sequence for the hybrid design $(Z_1^*, ..., Z_{K-1}^*, Z_K^{*C})$ converges in distribution to the multivariate Gaussian distribution characterized by U^ρ defined in Example 1.1, under the null hypothesis. Therefore, by definition, since the constant c satisfies equation (1.7), then $\{c, c/\sqrt{2}, ..., c/\sqrt{K}\}$ satisfies Definition 1.4 of asymptotically acceptable critical values. The same proof by definition applies for Example 1.2. In this case, we first select the constant c based on the joint distribution of U^1, which is what the sub-vector $(Z_1^*, ..., Z_{K-1}^*, Z_K^{*C})$ converges to under the null hypothesis with no covariate adjustment (i.e., $\rho = 1$, these are the standard OBF boundaries), and then modify only the last critical value to account for ρ using U^ρ. □

A Note on Computation: A trick that is done to compute the Gaussian probabilities above in non-hybrid designs is called *recursive integration* (Gernot and Brannath, 2016; Armitage et al., 1969). Recursive integration allows for quick computation of the probabilities based on the fact that there are independent increments of sizes $n_1, ..., n_K$, similar to how we derived the joint distribution of $\sqrt{n}\bar{\boldsymbol{\phi}}^*$ by partitioning it into a vector with elements that sum only observations within stage k. In the hybrid ANOVA-ANCOVA case, we can no longer construct independent increments because of the correlation between ANOVA and ANCOVA within the same stage. Instead, we use software that allows us to compute probabilities of Gaussian polytopes numerically. Using standard statistical software, this integration is feasible as the vast majority of trials have several stages, rather than tens of stages.

Alpha-Spending: Alpha-spending can be incorporated into our approach (Lan and DeMets, 1983). There are several scenarios where one might want to employ alpha-spending related to our paper. The first is when the information fractions are not known ahead of

time, and where the hybrid design in Procedure 1.1 is used. This is a standard application of alpha-spending. Rather than solving for the critical values as in Corollaries 1.1 and 1.2, one would specify an alpha-spending function, and sequentially solve for the critical values as the trial progresses.

Specifically, the critical values are calculated on the fly based on a pre-specified amount of type I error (alpha) to be spent at each stage. Let $\alpha^*(t)$ be a strictly increasing function with $\alpha^*(0) = 0$ and $\alpha^*(1) = \alpha$. Then, for the hybrid design in Procedure 1.1, we would solve for the following critical values $\{u_1, ..., u_K\}$ sequentially, at each stage based on the observed information fractions that define the joint distribution of the test statistics in Corollary 1.1:

$$\begin{aligned} P(|Z_1^*| \geq u_1) &= \alpha^*(t_1^*) \\ P(|Z_2^*| \geq u_2, |Z_1^*| < u_1) &= \alpha^*(t_2^*) - \alpha^*(t_1^*) \\ &... \\ P(|Z_K^{*C}| \geq u_K, |Z_{K-1}^*| < u_{K-1}, ..., |Z_2^*| < u_2, |Z_1^*| < u_1) &= \alpha^*(t_K^*) - \alpha^*(t_{K-1}^*). \end{aligned}$$

If we used the O'Brien-Fleming alpha-spending function, we would end up with the boundary shape of Corollary 1.2, since all of the inflation would occur at the last stage. Therefore, we could use this approach with known, or estimated, ρ.

A second application of alpha-spending would be to choose a different alpha-spending function that spends relatively more type I error in earlier stages. This may be desirable if one wanted boundaries like that in Corollary 1.1, but did not know ρ initially. A third application of alpha-spending is to extend our methods to an arbitrary sequence – possibly unknown ahead of time – of un-adjusted and covariate-adjusted estimators throughout the stages of a group sequential trial, and where ρ could be re-estimated throughout the course of the trial.

A.2.3 Adjusted Estimation and Inference in the Hybrid Design

In a standard group sequential design, calculating adjusted p-values, confidence intervals, and point estimates comes down to being able to define and calculate $P((Z^*_{k'}, k') \succeq (z^*_k, k))$. That is, when is the observed test statistic $Z^*_{k'}$ and stopping stage k "more extreme" than some arbitrary values (z^*_k, k)? Calculating $P((Z^*_{k'}, k') \succeq (z^*_k, k))$ requires three main things: a joint distribution of test statistics, critical values for each stage that define the stopping boundaries, and a sample space ordering that defines "extreme". Our contribution is to modify these three elements for the hybrid design, which we describe below. We provide more details about sample space orderings in Section A.2.3. Finally, we review how computing $P((Z^*_{k'}, k') \succeq (z^*_k, k))$ can be used for adjusted estimation and inference that takes into account the group sequential design.

Modifications for Estimation and Inference in the Hybrid Design

1. The joint distribution of test statistics across stages. In particular, we need to know the covariance matrix between the test statistics across stages and when performing ANOVA and ANCOVA at each stage. In Corollary 1.1, we derived this covariance matrix. In practice, we would select the subset of the covariance matrix that represents the tests that are needed for a particular ordering, e.g., $(Z^*_1, Z^*_2, Z^*_3, Z^{*C}_4)$.

2. Critical values (l_k, u_k) at each stage k. The pairing of these critical values with the joint distribution of test statistics in Corollary 1.1 allows us to characterize the distribution of observed test statistics. The distribution of the observed test statistics differs from the joint distribution in Corollary 1.1 in that only some values of the test statistics will be observed at particular stages, based on whether or not the null hypothesis was rejected at that stage. For example, if $(l_1, u_1) = (-4, 4)$, then the probability of observing $(z^*_1 = 3, k = 1)$ is exactly 0 because the null hypothesis cannot be rejected with a test statistic of 3 in the first stage. These critical values should be chosen so that they control type I error in the hybrid design, discussed in Section A.2.2.

3. An sample space ordering for the two-dimensional vector (z_k^*, k). The orderings tell us whether $(z_{k'}^*, k') \succeq (z_k^*, k)$. It is then used in conjunction with (1) and (2) to perform computations with the observed test statistic, i.e., $P((Z_{k'}^*, k') \succeq (z_k^*, k))$. We provide more information about sample space orderings and how we modify them for the hybrid design in Section A.2.3.

Sample Space Orderings and their Modifications for the Hybrid Design

We briefly describe four commonly used sample space orderings and discuss two in detail: stage-wise ordering and sample mean ordering. Our contribution is to modify the computations for what is considered "as or more extreme" in the context of the hybrid ANOVA-ANCOVA analysis, which is then used to construct p-values, confidence intervals, and point estimates.

Four sample space orderings are used in practice. Consider two test statistics (z_k^*, k) and $(z_{k'}^*, k')$, and we want to say whether $(z_k^*, k) \succeq (z_{k'}^*, k')$. Additional details of these common sample space orderings can be found elsewhere (Gernot and Brannath, 2016).

- Stage-wise ordering is based on the value of the z_k's alone if $k = k'$. $(z_{k'}^*, k') \succeq (z_k^*, k)$ if $k' < k$ and $z_{k'}^* \geq u_{k'}$ (crossing the upper boundary at an earlier stage), and $(z_{k'}^*, k') \preceq (z_k^*, k)$ if $k' < k$ and $z_{k'}^* \leq u_{k'}$ (crossing the lower boundary at an earlier stage).

- Likelihood ratio ordering has $(z_{k'}^*, k') \succeq (z_k^*, k)$ if $z_{k'}^* > z_k^*$.

- Sample mean ordering ignores stage and sample size altogether and just compares the values of the standardized sample means, rather than the test statistics. If each stage has an equal sample size, this reduces to considering $(z_{k'}^*, k') \succeq (z_k^*, k)$ if $\frac{z_{k'}^*}{\sqrt{k'}} > \frac{z_k^*}{\sqrt{k}}$.

- Score test ordering orders test statistics based on the score statistic. If each stage has an equal sample size, this reduces to considering $(z_{k'}^*, k') \succeq (z_k^*, k)$ if $z_{k'}^* \sqrt{k'} > z_k^* \sqrt{k}$.

For the hybrid ANOVA-ANCOVA design that we consider, the ordering is based on the ANCOVA test at the last stage, whether that be from stopping early or from continuing through the last stage. The key difference between performing standard ANOVA or ANCOVA analyses at each stage and a hybrid ANOVA-ANCOVA analysis is that since the ordering is based off of ANCOVA at the last stage, in the hybrid setting, there may be two tests performed at the same stage (i.e., an ANOVA that crosses an interim boundary followed by an ANCOVA).

Now we detail the computations of $P[(Z^*_{k'}, k') \succeq (z^*_k, k)]$ for two of these orderings and describe the modifications that we need to make in order to account for hybrid ANOVA-ANCOVA analyses. In both settings, we focus on the setting where the trial is monitored with an ANOVA test statistic, Z^*_k, and estimation and inference is desired based on a final ANCOVA test Z^{*C}_k. Let C_j be the event of not rejecting through stage j using an ANOVA statistic, and R_j be the event of rejecting at stage j using an ANOVA statistic. We can obtain the probabilities below using the covariance of Z^* from Corollary 1.1, which is a function of both the information fraction (and sometimes the joint distribution between ANOVA and ANCOVA).

Stage-Wise Ordering

When using the stage-wise ordering, rejecting at a later stage is considered a *smaller* value, so we don't consider it in the upper tail probabilities. We note that for the non-hybrid and hybrid designs, we only need the distribution of Z^*_k and Z^{*C}_k values at stages $1, ..., k$. We do not need the distribution for stages $k+1, ..., K$. Furthermore, the hybrid designs require the joint distribution of Z^*_k and Z^{*C}_k.

In the non-hybrid design, we perform ANOVA tests in the interim and we wish to perform estimation and inference using ANOVA. The following shows how to compute $P[(Z^*_{k'}, k') \succeq (z^*_k, k)]$ in this setting. It is written for ANOVA, but also covers ANCOVA just by replacing Z^* with Z^{*C} since the covariance is the same as long as we're not switching between ANOVA

and ANCOVA.

$$
\begin{aligned}
P[(Z^*_{k'}, k') \succeq (z^*_k, k)] &= \sum_{j=1}^{k-1} \Big\{ P[Z^*_j \geq u_j | C_{j-1}, R_j] P[C_{j-1}, R_j] \Big\} + \\
&\qquad P[Z^*_k \geq z^*_k | C_{k-1}] P[C_{k-1}] \\
&= \sum_{j=1}^{k-1} \Big\{ P[Z^*_j \geq u_j, (Z^*_j \geq u_j \text{ or } Z^*_j \leq \ell_j), \ell_{j-1} < Z^*_{j-1} < u_{j-1}, ..., \ell_1 < Z^*_1 < u_1] \Big\} + \\
&\qquad P[Z^*_k \geq z^*_k, \ell_{k-1} < Z^*_{k-1} < u_{k-1}, ..., \ell_1 < Z^*_1 < u_1] \\
&= \sum_{j=1}^{k-1} \Big\{ P[Z^*_j \geq u_j, \ell_{j-1} < Z^*_{j-1} < u_{j-1}, ..., \ell_1 < Z^*_1 < u_1] \Big\} + \\
&\qquad P[Z^*_k \geq z^*_k, \ell_{k-1} < Z^*_{k-1} < u_{k-1}, ..., \ell_1 < Z^*_1 < u_1].
\end{aligned}
$$

In the hybrid design, there are two options. The first scenario is that the trial is stopped early after crossing a boundary using an ANOVA statistic, and an ANCOVA analysis is performed at that same stage. The second scenario is that the trial continues until the last stage (i.e., no boundaries are crossed during interim monitoring), and only an ANCOVA analysis is performed. The computation for $P[(Z^*_{k'}, k') \succeq (z^*_k, k)]$ is different between these cases:

Early Stopping

$$
\begin{aligned}
P[(Z^{*C}_{k'}, k') \succeq (z^{*C}_k, k)] &= \sum_{j=1}^{k-1} \Big\{ P[Z^*_j \geq u_j | C_{j-1}, R_j] P[C_{j-1}, R_j] \Big\} + \\
&\qquad P[Z^{*C}_k \geq z^{*C}_k | R_k, C_{k-1}] P[R_k, C_{k-1}] \\
&= \sum_{j=1}^{k-1} \Big\{ P[Z^*_j \geq u_j, \ell_{j-1} < Z^*_{j-1} < u_{j-1}, ..., \ell_1 < Z^*_1 < u_1] \Big\} + \\
&\qquad P[Z^{*C}_k \geq z^{*C}_k, (Z^*_k \geq u_k \text{ or } Z^*_k \leq \ell_k), ..., \ell_1 < Z^*_1 < u_1].
\end{aligned}
$$

No Early Stopping

$$P[(Z_{k'}^{*C}, k') \succeq (z_k^{*C}, k)] = \sum_{j=1}^{k-1} \Big\{ P[Z_j^* \geq u_j | C_{j-1}, R_j] P[C_{j-1}, R_j] \Big\} +$$
$$P[Z_k^{*C} \geq z_k^{*C} | C_{k-1}] P[C_{k-1}]$$
$$= \sum_{j=1}^{k-1} \Big\{ P[Z_j^* \geq u_j, \ell_{j-1} < Z_{j-1}^* < u_{j-1}, ..., \ell_1 < Z_1^* < u_1] \Big\} +$$
$$P[Z_k^{*C} \geq z_k^{*C}, \ell_{k-1} < Z_{k-1}^* < u_{k-1}, ..., \ell_1 < Z_1^* < u_1].$$

Sample Mean Ordering

Let X_k^* be the cumulative sample mean (i.e., the parameter estimate), where $X_k^* = Z_k^*/\sqrt{n_k^*}$. Let X_k^{*C} be the sample mean using ANCOVA. Let (x_k^*, k) be the observed bivariate test statistic for stopping stage and sample mean. Sample mean ordering is not directly a function of the stopping stage, so we ignore the observed stopping stage in the calculations, other than to note that we are at the end of the trial. Note that rather than using Σ_{Z^*} as the covariance, we use the covariance matrix for the sample means in Theorem 1.1, since we can translate the rejection region to the sample mean scale rather than the likelihood ratio scale.

This is written for ANOVA, but also covers ANCOVA just by replacing Z^* with Z^{*C} and X^* with X^{*C} since the covariance is the same as long as we're not switching between ANOVA and ANCOVA.

$$P[X_{k'}^* \succeq x_k^*] = \sum_{j=1}^{K-1} \Big\{ P[X_j^* \geq x_j^* | R_j, C_{j-1}] P[R_j, C_{j-1}] \Big\} +$$
$$P[X_K^* \geq x_k^* | C_{K-1}] P[C_{K-1}]$$
$$= \sum_{j=1}^{K-1} \Big\{ P[X_j^* \geq x_k^*, (Z_j^* \geq u_j \text{ or } Z_j^* \leq \ell_j), \ell_{j-1} < Z_{j-1}^* < u_{j-1}, ..., \ell_1 < Z_1^* < u_1] \Big\} +$$
$$P[X_K^* \geq x_k^*, \ell_{K-1} < Z_{K-1}^* < u_{K-1}, ..., \ell_1 < Z_1^* < u_1].$$

Since the sample mean ordering is not directly a function of the stopping stage, the computation for scenarios where there is early stopping versus not can be grouped together.

$$\begin{aligned} P[X_{k'}^{*C} \succeq x_k^{*C}] &= \sum_{j=1}^{K-1} \Big\{ P[X_j^{*C} \geq x_k^{*C} | R_j, C_{j-1}] P[R_j, C_{j-1}] \Big\} + \\ &\qquad P[X_K^{*C} \geq x_k^{*C} | C_{K-1}] P[C_{K-1}] \\ &= \sum_{j=1}^{K-1} \Big\{ P[X_j^{*C} \geq x_k^{*C}, (Z_j^* \geq u_j \text{ or } Z_j^* \leq \ell_j), \ell_{j-1} < Z_{j-1}^* < u_{j-1}, ..., \ell_1 < Z_1^* < u_1] \Big\} + \\ &\qquad P[X_K^{*C} \geq x_k^{*C}, \ell_{K-1} < Z_{K-1}^* < u_{K-1}, ..., \ell_1 < Z_1^* < u_1]. \end{aligned}$$

Review of Estimation and Inference in Group Sequential Designs

Up until now, we have modified the sample space ordering and related computation for $P((Z_{k'}^*, k') \succeq (z_k^*, k))$. To implement these probabilities, we use the results from Theorem 1.1 and Corollary 1.1. Now, we can calculate p-values, the probability of observing (z_k^*, k) or something *more extreme* under the null hypothesis P_0 using a standard group sequential approach (Gernot and Brannath, 2016):

$$P_0((Z_{k'}^*, k') \succeq (z_k^*, k)). \tag{A.4}$$

A two-sided p-value can be computed similarly, using both the above and $P_0((Z_{k'}^*, k') \preceq (z_k^*, k))$. We can use these probabilities to compute a $(1-\alpha) * 100\%$ confidence interval, we find Δ_l and Δ_u effect sizes for P such that

$$P_{\Delta_l}((Z_{k'}^*, k') \succeq (z_k^*, k)) = P_{\Delta_u}((Z_{k'}^*, k') \preceq (z_k^*, k)) = \alpha/2.$$

Then (Δ_l, Δ_u) is a $(1-\alpha) * 100\%$ confidence interval (Gernot and Brannath, 2016). In a similar fashion, we can derive median unbiased point estimates. We define $\tilde{\Delta}$ as the median

unbiased point estimate (Gernot and Brannath, 2016), which is the value that satisfies

$$P_{\tilde{\Delta}}((Z^*_{k'}, k') \succeq (z^*_k, k)) = P_{\tilde{\Delta}}((Z^*_{k'}, k') \succeq (z^*_k, k)) = 0.5.$$

In all of these cases, we use the covariance matrix of Z^* described in Corollary 1.1 to compute these probabilities. When working with P_0, we use the distribution exactly as described in Corollary 1.1; when working with P_Δ for some non-zero effect size, we suitably center and scale the test statistics first based on the alternative hypothesis of effect size Δ, then use Corollary 1.1.

A.3 Supplemental Simulation Results

Figure A.1: Type I error under scenario (B)(ii) with misspecified ρ at the design stage across 25,000 simulated trials. Stars indicate where the anticipated reduction and the true reduction in variance from using ANCOVA match up (correct specification), which represents the orange line in Figure 4. The dark purple line at the top of each subplot represents anticipating that there is no reduction in variance from using ANCOVA, equivalent to not modifying the boundaries, which represents the green line in Figure 4.

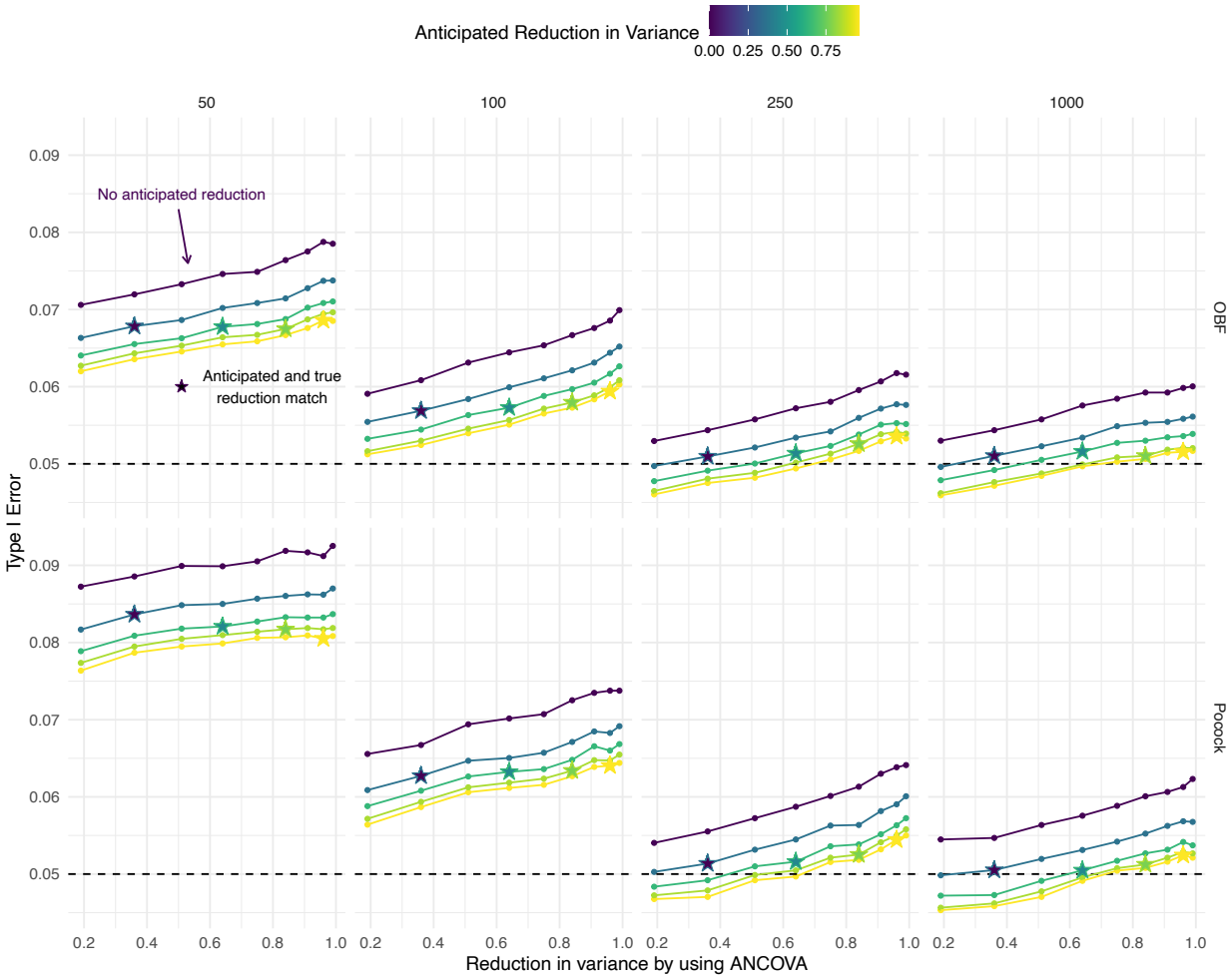


Table A.1: Bias (×100) and coverage results based on 10,000 simulated trials for each row, end-inflated OBF boundaries for 3-stage trial. Bias is shown as median bias and mean bias in parentheses. Δ is the true effect size, ρ is the true ratio of standard error between the ANCOVA and ANOVA estimators, and n is the total sample size for the trial.

			Scaled Bias: Median (Mean)			Coverage		
Δ	ρ	n	Simple	GS	GS + Adjust	Simple	GS	GS + Adjust
0.00	0.25	50	-0.03 (-0.02)	-0.11 (-0.18)	-0.03 (-0.10)	0.93	0.92	0.93
0.00	0.25	100	0.01 (-0.00)	-0.04 (-0.11)	0.02 (-0.01)	0.94	0.93	0.94
0.00	0.25	250	0.02 (0.01)	-0.01 (-0.07)	0.02 (-0.00)	0.95	0.94	0.94
0.00	0.25	1000	-0.00 (-0.00)	-0.01 (-0.03)	0.00 (-0.01)	0.95	0.94	0.95
0.00	0.50	50	-0.04 (-0.02)	-0.16 (-0.29)	-0.04 (-0.06)	0.93	0.92	0.93
0.00	0.50	100	0.03 (0.00)	-0.06 (-0.18)	0.03 (-0.01)	0.94	0.93	0.94
0.00	0.50	250	0.04 (0.01)	-0.01 (-0.10)	0.04 (0.02)	0.95	0.94	0.95
0.00	0.50	1000	-0.00 (-0.01)	-0.03 (-0.06)	-0.01 (-0.02)	0.95	0.94	0.95
0.10	0.25	50	-0.01 (0.04)	-0.93 (-1.54)	0.00 (0.92)	0.93	0.95	0.94
0.10	0.25	100	0.03 (0.03)	-2.60 (-2.64)	0.06 (0.98)	0.94	0.96	0.94
0.10	0.25	250	0.03 (0.04)	-5.14 (-4.65)	0.03 (0.87)	0.95	0.11	0.95
0.10	0.25	1000	0.01 (0.04)	-5.07 (-5.95)	-0.00 (0.28)	0.95	0.03	0.95
0.10	0.50	50	0.00 (0.19)	-0.30 (-0.81)	-0.01 (0.79)	0.93	0.94	0.94
0.10	0.50	100	0.07 (0.12)	-0.24 (-1.03)	0.07 (0.71)	0.94	0.94	0.94
0.10	0.50	250	0.09 (0.15)	-1.06 (-1.54)	0.06 (0.68)	0.95	0.96	0.95
0.10	0.50	1000	0.07 (0.16)	-2.13 (-2.60)	-0.02 (0.30)	0.95	0.52	0.95
0.20	0.25	50	0.03 (0.11)	-9.10 (-8.21)	0.10 (2.12)	0.93	0.38	0.93
0.20	0.25	100	0.03 (0.09)	-12.19 (-10.57)	0.10 (1.78)	0.94	0.20	0.94
0.20	0.25	250	0.07 (0.09)	-10.68 (-11.94)	-0.21 (0.54)	0.95	0.03	0.95
0.20	0.25	1000	0.03 (0.07)	-14.68 (-7.98)	-0.08 (0.21)	0.95	0.43	0.95
0.20	0.50	50	0.13 (0.42)	-1.79 (-2.74)	0.11 (1.67)	0.93	0.94	0.93
0.20	0.50	100	0.17 (0.35)	-4.98 (-4.01)	0.15 (1.39)	0.94	0.95	0.94
0.20	0.50	250	0.23 (0.36)	-4.55 (-5.26)	-0.16 (0.57)	0.94	0.52	0.94
0.20	0.50	1000	0.14 (0.27)	-9.36 (-4.79)	-0.12 (0.28)	0.95	0.43	0.95
0.50	0.25	50	0.14 (0.32)	-29.23 (-28.71)	-0.17 (0.89)	0.93	0.08	0.94
0.50	0.25	100	0.12 (0.25)	-33.36 (-26.01)	0.21 (1.10)	0.94	0.22	0.94
0.50	0.25	250	0.06 (0.11)	-0.78 (-8.80)	-0.05 (-1.20)	0.95	0.73	0.95
0.50	0.25	1000	-0.02 (-0.01)	-0.02 (-0.01)	-0.02 (-0.01)	0.95	0.95	0.95
0.50	0.50	50	0.58 (1.26)	-10.30 (-12.75)	-0.05 (1.33)	0.92	0.61	0.93
0.50	0.50	100	0.42 (0.89)	-16.76 (-13.11)	0.40 (1.29)	0.94	0.50	0.94
0.50	0.50	250	0.23 (0.40)	-1.02 (-5.87)	-0.10 (-0.61)	0.95	0.74	0.95
0.50	0.50	1000	-0.04 (-0.02)	-0.04 (-0.02)	-0.04 (-0.02)	0.95	0.95	0.95

Table A.2: Bias (×100) and coverage results based on 10,000 simulated trials for each row, end-inflated Pocock boundaries, 2-stage trial. Bias is shown as median bias and mean bias in parentheses. Δ is the true effect size, ρ is the true ratio of standard error between the ANCOVA and ANOVA estimators, and n is the total sample size for the trial.

			Scaled Bias: Median (Mean)			Coverage		
Δ	ρ	n	Simple	GS	GS + Adjust	Simple	GS	GS + Adjust
0.00	0.25	50	0.02 (0.03)	0.02 (0.03)	0.03 (0.13)	0.94	0.96	0.93
0.00	0.25	100	0.02 (0.01)	0.02 (0.01)	0.03 (0.07)	0.94	0.97	0.94
0.00	0.25	250	0.02 (0.01)	0.02 (0.01)	0.02 (-0.05)	0.95	0.97	0.95
0.00	0.25	1000	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.95	0.97	0.95
0.00	0.50	50	0.04 (0.07)	0.04 (0.07)	0.05 (0.19)	0.93	0.96	0.93
0.00	0.50	100	0.05 (0.01)	0.05 (0.01)	0.05 (-0.01)	0.94	0.96	0.94
0.00	0.50	250	0.04 (0.03)	0.04 (0.03)	0.04 (-0.03)	0.94	0.97	0.94
0.00	0.50	1000	0.02 (0.02)	0.02 (0.02)	0.02 (0.01)	0.95	0.97	0.95
0.10	0.25	50	0.06 (0.12)	-1.12 (-1.36)	0.09 (1.39)	0.94	0.97	0.94
0.10	0.25	100	0.05 (0.07)	-2.70 (-2.52)	0.04 (1.23)	0.94	0.96	0.95
0.10	0.25	250	0.05 (0.09)	-5.13 (-4.25)	0.04 (1.07)	0.95	0.16	0.95
0.10	0.25	1000	0.03 (0.06)	-1.46 (-3.44)	0.01 (0.07)	0.95	0.50	0.95
0.10	0.50	50	0.15 (0.36)	-0.01 (-0.15)	0.14 (1.21)	0.94	0.96	0.94
0.10	0.50	100	0.15 (0.28)	-0.20 (-0.47)	0.06 (1.01)	0.94	0.96	0.94
0.10	0.50	250	0.16 (0.31)	-1.25 (-1.15)	0.08 (0.90)	0.94	0.96	0.95
0.10	0.50	1000	0.13 (0.24)	-2.19 (-1.97)	0.04 (0.19)	0.95	0.51	0.95
0.20	0.25	50	0.08 (0.19)	-9.09 (-7.63)	0.11 (2.32)	0.94	0.58	0.94
0.20	0.25	100	0.08 (0.16)	-12.17 (-9.11)	0.12 (1.75)	0.94	0.23	0.95
0.20	0.25	250	0.06 (0.13)	-2.99 (-6.84)	-0.06 (0.13)	0.95	0.50	0.95
0.20	0.25	1000	0.03 (0.02)	0.01 (-0.17)	0.02 (-0.08)	0.95	0.94	0.95
0.20	0.50	50	0.30 (0.67)	-2.06 (-1.84)	0.22 (1.92)	0.93	0.95	0.94
0.20	0.50	100	0.29 (0.59)	-5.13 (-3.38)	0.13 (1.53)	0.94	0.95	0.94
0.20	0.50	250	0.27 (0.48)	-4.68 (-3.92)	-0.09 (0.36)	0.95	0.52	0.95
0.20	0.50	1000	0.05 (0.05)	0.03 (-0.10)	0.04 (-0.04)	0.95	0.94	0.95
0.50	0.25	50	0.10 (0.21)	-3.37 (-13.93)	-0.48 (-0.89)	0.93	0.59	0.94
0.50	0.25	100	-0.01 (0.05)	-0.36 (-3.94)	-0.07 (-1.40)	0.94	0.85	0.94
0.50	0.25	250	0.00 (0.01)	0.00 (-0.00)	0.00 (0.00)	0.95	0.94	0.94
0.50	0.25	1000	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)	0.95	0.95	0.95
0.50	0.50	50	0.40 (0.88)	-4.76 (-8.60)	-0.58 (-0.11)	0.93	0.60	0.94
0.50	0.50	100	0.03 (0.24)	-0.41 (-2.74)	-0.09 (-0.86)	0.94	0.86	0.94
0.50	0.50	250	0.01 (0.03)	0.00 (0.01)	0.00 (0.02)	0.95	0.94	0.94
0.50	0.50	1000	0.04 (0.03)	0.04 (0.03)	0.04 (0.03)	0.95	0.95	0.95

Table A.3: Bias (×100) and coverage results based on 10,000 simulated trials for each row, end-inflated Pocock boundaries, 3-stage trial, 2 covariates. Bias is shown as median bias and mean bias in parentheses. Δ is the true effect size, ρ is the true ratio of standard error between the ANCOVA and ANOVA estimators, and n is the total sample size for the trial.

			Scaled Bias: Median (Mean)			Coverage		
Δ	ρ	n	Simple	GS	GS + Adjust	Simple	GS	GS + Adjust
0.00	0.25	50	-0.04 (-0.01)	-0.14 (-0.18)	-0.07 (-0.29)	0.93	0.94	0.91
0.00	0.25	100	-0.00 (-0.01)	-0.05 (-0.12)	0.01 (-0.02)	0.94	0.96	0.94
0.00	0.25	250	-0.02 (-0.01)	-0.05 (-0.08)	-0.02 (-0.03)	0.94	0.96	0.94
0.00	0.25	1000	-0.01 (0.00)	-0.02 (-0.03)	-0.01 (-0.01)	0.95	0.96	0.95
0.00	0.50	50	-0.09 (-0.00)	-0.28 (-0.33)	-0.09 (-0.14)	0.92	0.93	0.91
0.00	0.50	100	-0.01 (-0.00)	-0.10 (-0.21)	0.00 (-0.02)	0.94	0.95	0.94
0.00	0.50	250	-0.04 (-0.04)	-0.11 (-0.17)	-0.04 (-0.04)	0.94	0.95	0.94
0.00	0.50	1000	-0.02 (0.01)	-0.05 (-0.05)	-0.02 (0.01)	0.94	0.96	0.95
0.10	0.25	50	0.00 (0.12)	-1.63 (-1.92)	0.02 (1.79)	0.93	0.95	0.93
0.10	0.25	100	0.03 (0.11)	-3.37 (-3.13)	0.02 (1.76)	0.94	0.96	0.95
0.10	0.25	250	0.03 (0.09)	-5.68 (-4.86)	0.04 (1.71)	0.94	0.19	0.94
0.10	0.25	1000	0.03 (0.09)	-6.87 (-4.89)	-0.02 (0.36)	0.95	0.30	0.95
0.10	0.50	50	0.06 (0.49)	-0.36 (-0.76)	0.00 (1.57)	0.92	0.93	0.93
0.10	0.50	100	0.13 (0.44)	-0.41 (-0.98)	0.05 (1.60)	0.94	0.94	0.94
0.10	0.50	250	0.12 (0.39)	-1.87 (-1.74)	0.05 (1.40)	0.94	0.94	0.94
0.10	0.50	1000	0.16 (0.36)	-3.78 (-2.75)	-0.04 (0.48)	0.95	0.61	0.95
0.20	0.25	50	0.03 (0.25)	-10.27 (-8.81)	0.07 (3.28)	0.92	0.28	0.94
0.20	0.25	100	0.07 (0.22)	-12.98 (-10.56)	0.10 (2.89)	0.94	0.16	0.95
0.20	0.25	250	0.06 (0.18)	-13.75 (-9.64)	0.05 (0.73)	0.94	0.31	0.95
0.20	0.25	1000	0.03 (0.04)	-0.09 (-1.45)	0.01 (-0.50)	0.95	0.86	0.95
0.20	0.50	50	0.23 (0.93)	-3.07 (-3.00)	0.11 (2.82)	0.92	0.93	0.93
0.20	0.50	100	0.34 (0.87)	-6.37 (-4.50)	0.15 (2.46)	0.94	0.94	0.95
0.20	0.50	250	0.33 (0.72)	-7.66 (-5.52)	-0.01 (0.91)	0.94	0.61	0.95
0.20	0.50	1000	0.10 (0.14)	-0.08 (-1.00)	0.03 (-0.29)	0.95	0.86	0.95
0.50	0.25	50	0.18 (0.43)	-35.04 (-21.86)	-0.59 (-0.12)	0.92	0.36	0.93
0.50	0.25	100	0.07 (0.21)	-1.75 (-10.78)	-0.10 (-1.70)	0.94	0.68	0.94
0.50	0.25	250	-0.03 (-0.01)	-0.05 (-0.54)	-0.02 (-0.27)	0.94	0.93	0.94
0.50	0.25	1000	0.02 (0.01)	0.01 (0.01)	0.01 (0.01)	0.94	0.94	0.94
0.50	0.50	50	0.89 (1.68)	-20.33 (-12.93)	-0.74 (1.00)	0.92	0.55	0.93
0.50	0.50	100	0.35 (0.78)	-2.33 (-7.00)	-0.10 (-0.62)	0.94	0.68	0.94
0.50	0.50	250	-0.04 (0.01)	-0.07 (-0.42)	-0.04 (-0.20)	0.94	0.93	0.94
0.50	0.50	1000	0.03 (0.03)	0.03 (0.03)	0.03 (0.03)	0.94	0.94	0.94

Appendix B

SUPPLEMENTAL MATERIAL FOR CHAPTER 2

B.1 Illustration of G-Computation and Prediction Bias

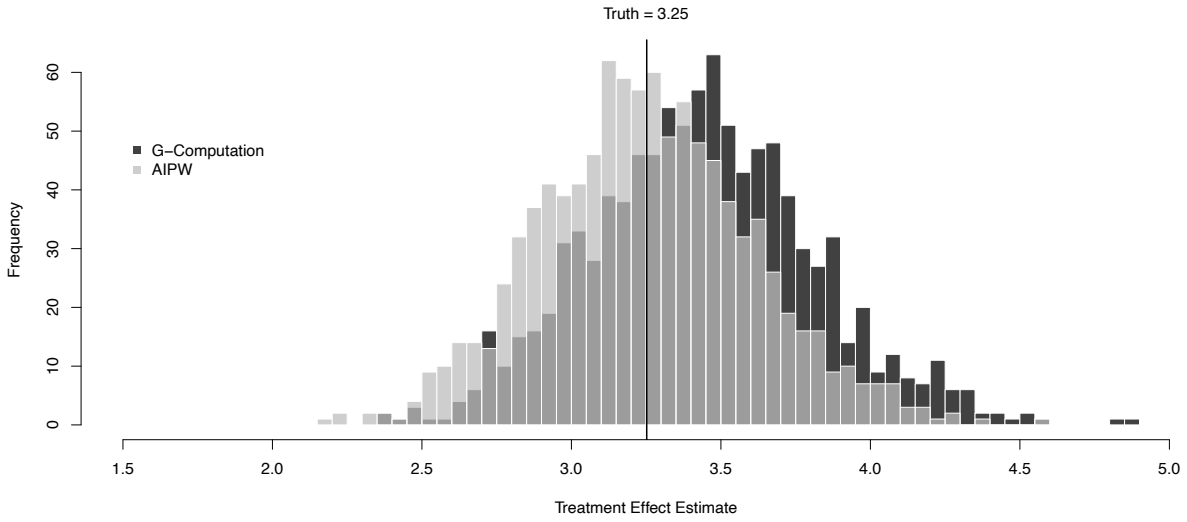
In order to make minimal assumptions, many proposals for model-assisted covariate adjustment – including those listed above – fall under the umbrella of *g-computation* (**?**). G-computation estimates the unconditional mean response by first estimating the conditional mean response as a function of the prognostic baseline covariates and averaging that function over the distribution of covariates in the trial. The g-computation estimator can theoretically be computed using any working model for the mean response as a function of the covariates. However, one fact that may often be ignored is not every model will yield an estimate that is robust to misspecification of this functional form (Guo and Basse, 2021). For example, Poisson and negative binomial regressions are commonly used to model count variables. While the g-computation estimator using a Poisson working model is always asymptotically unbiased, the one using a negative binomial working model can be biased (see Figure B.1). To address this limitation, a better alternative is the augmented inverse propensity weighted (AIPW) estimator (Robins et al., 1994; Glynn and Quinn, 2010; Guo and Basse, 2021). The AIPW estimator can be viewed as a de-biased g-computation estimator (Firth and Bennett, 1998) and is a general form of covariate adjustment (Tsiatis et al., 2008). The AIPW estimator also provides a rigorous inference framework when flexible machine learning methods are used to estimate the conditional mean response. This motivates our focus on the AIPW estimator.

B.2 Asymptotic Linearity (AL) of AIPW Estimators under CAR

B.2.1 Preliminaries

Throughout, let P_n denote the empirical measure, i.e., $P_n(\cdot) = \frac{1}{n}\sum_{i=1}^{n}(\cdot)$, let P_0 denote the population measure, i.e., $P_0(\cdot) = E(\cdot)$, and let $P_0^n(\cdot) = \frac{1}{n}\sum_{i=1}^{n} P_0(\cdot|\mathcal{A}_n, \mathcal{Z}_n)$ denote the conditional population measure where the conditioning is on the set of all strata and treatment variables, $\mathcal{A}_n, \mathcal{Z}_n$, i.e., $\mathcal{A}_n = \{A_1, \ldots, A_n\}$ and $\mathcal{Z}_n = \{Z_1, \ldots, Z_n\}$. Let I_a be the indicator function of $A = a$, and $I_{a,i}$ be the ith realization.

Figure B.1: Treatment effect estimates for 1000 simulated two-arm trials with $n = 500$ based on a Poisson data generating process, under simple randomization with $\pi_1 = 1/3$. The response means conditional on $X = (X_c, X_b)^T$ are $E(Y_1 \mid X) = \log(1 + X_c + 6X_c^3 + X_b)$ and $EY(Y_2 \mid X) = X_c + 10X_c^2 + X_b$, where X_c is uniformly distributed on the interval (-5, 5), X_b is binary with $P(X_b = 1) = 0.5$, and X_c and X_b are independent. The working model is negative binomial in covariate X including treatment-by-covariate interactions, with an unknown dispersion parameter.



Let $W_i = (Y_{a,i}, A_i, X_i, Z_i)$, and define the following function:

$$\psi_a(W_i; \pi, \mu) = \frac{I_{a,i}}{\pi} \{Y_{a,i} - \mu(X) + P_0(\mu) - \theta_a\} + \mu(X_i) - P_0(\mu). \tag{B.1}$$

Then we have the following relationship for the AIPW estimator in equation (1): $\hat{\theta}_a - \theta_a \equiv P_n \psi_a(\cdot; \hat{\pi}_a, \hat{\mu}_a)$. This facilitates a convenient asymptotic expansion. We will show that, regardless of whether we use Assumptions 2.1–2.3, *or* we use Assumptions 2.1 and 2.4 (i.e., we use cross-fitting), the influence function for $\hat{\theta}_a$ is given by $\psi_a(W_i; \pi_a, \mu_a)$.

Let $\psi_{a,i}(\pi, \mu) = \psi_a(W_i; \pi, \mu)$. Then by definition, $\hat{\theta}_a - \theta_a = P_n \psi_a(\hat{\pi}_a, \hat{\mu}_a)$, and

$P_0\psi_a(\pi_a, \mu_a) = 0$. Then we can form the following expansion:

$$\hat{\theta}_a - \theta_a \equiv P_n\psi_a(\hat{\pi}_a, \hat{\mu}_a) = \underbrace{P_n\psi_a(\hat{\pi}_a, \hat{\mu}_a) - P_n\psi_a(\hat{\pi}_a, \mu_a)}_{(I)} + \underbrace{P_n\psi_a(\hat{\pi}_a, \mu_a) - P_n\psi_a(\pi_a, \mu_a)}_{(II)} + \underbrace{P_n\psi_a(\pi_a, \mu_a)}_{(III)}. \tag{B.2}$$

Term (III) is the linear term that we want that gives us the desired influence function. Therefore, it remains to show that terms (I) and (II) are asymptotically negligible, $o_P(1/\sqrt{n})$. Term (II) does not depend on $\hat{\mu}_a$, so it is unaffected by estimation of that function, and all assumptions on $\hat{\mu}_a$. For term (II), we have:

$$(II) = \left(\frac{1}{\hat{\pi}_a} - \frac{1}{\pi_a}\right)\frac{1}{n}\sum_{i=1}^{n} I_{a,i}\Big\{Y_{a,i} - \mu_a(X_i) - (\theta_a - P_0\mu_a(X_i))\Big\}.$$

For this to hold, by an application of Slutsky's theorem, since $\hat{\pi}_a \to \pi_a$ in probability, it suffices to show that $\frac{1}{n}\sum_{i=1}^{n} I_{a,i}\left\{Y_{a,i} - \mu_a(X_i) - (\theta_a - P_0\mu_a(X_i))\right\} = O_p(1/\sqrt{n})$. In Section B.3.1, we break up the influence function that we derive in this section into five parts. Parts 1, 2, and 3 can be added together $(\psi_a^1 + \psi_a^2 + \psi_a^3)$ to obtain (II). Using the results of Section B.3 about the convergence in distribution of $P_n(\psi_a^1)$ and $P_n(\psi_a^3)$, together with Remark B.3 that shows convergence in distribution of $P_n(\psi_a^2)$, we see that (II) is $O_P(1/\sqrt{n})$.

Term (I) is the key term that we need to deal with by using either Donsker conditions (Assumptions 2.2 and 2.3), or with cross-fitting (Assumption 2.4). In Section B.2.2, we prove that term (I) is asymptotically negligible under Donsker conditions, which completes the proof of Theorem 2.1 (i). In Section B.2.4, we prove that the same result but under cross-fitting conditions, which completes the asymptotic linearity portion of the proof of Theorem 2.2. Once we have established the asymptotic linearity, the asymptotic normality under cross-fitting follows from Theorem 2.1.

B.2.2 Proof of Theorem 2.1 (i): AL under Donsker Conditions

Showing that term (I) is $o_P(1/\sqrt{n})$ under Assumptions 2.2 and 2.3 requires non-standard empirical process results because of the dependent data from CAR. In Section B.2.3, we develop several lemmas that allow us to show asymptotic negligibility for this term. We apply these lemmas throughout this section and direct the reader to that section B.2.3 for the technical details. For term (I), we have

$$\begin{aligned}
(I) &= \frac{1}{n}\sum_{i=1}^{n}\Big\{I_{a,i}\mu_a(X_i)\cdot\hat{\pi}_a^{-1} - \mu_a(X_i)\Big\} - \frac{1}{n}\sum_{i=1}^{n}\Big\{I_{a,i}\hat{\mu}_a(X_i)\cdot\hat{\pi}_a^{-1} - \hat{\mu}_a(X_i)\Big\} \\
&= \frac{1}{n}\sum_{i=1}^{n} I_{a,i}\Big\{\mu_a(X_i) - P_0\mu_a\Big\}\cdot\hat{\pi}_a^{-1} - (P_n - P_0)\mu_a \\
&\quad - \frac{1}{n}\sum_{i=1}^{n} I_{a,i}\Big\{\hat{\mu}_a(X_i) - P_0\hat{\mu}_a\Big\}\cdot\hat{\pi}_a^{-1} + (P_n - P_0)\hat{\mu}_a \\
&= \hat{\pi}_a^{-1}\frac{1}{n}\sum_{i=1}^{n} I_{a,i}\Big\{\mu_a(X_i) - P_0\mu_a - \hat{\mu}_a(X_i) + P_0\hat{\mu}_a\Big\} + (P_n - P_0)(\hat{\mu}_a - \mu_a) \\
&= (\pi_a^{-1} + o_p(1))\cdot\underbrace{\frac{1}{n}\sum_{i=1}^{n} I_{a,i}\Big\{\mu_a(X_i) - P_0\mu_a - \hat{\mu}_a(X_i) + P_0\hat{\mu}_a\Big\}}_{\mathcal{M}} + o_P(1\sqrt{n})
\end{aligned}$$

where $(P_n - P_0)(\hat{\mu}_a - \mu_a)$ is $o_P(1/\sqrt{n})$ by Remark B.1 combined with 2.3 (iii). Since $\hat{\pi}_a \to \pi_a$ in probability, by an application of the continuous mapping theorem and Slutsky's theorem, it suffices to show that $\mathcal{M} = o_P(1/\sqrt{n})$. $\mathcal{M}$ can be broken into three parts. We represent a sequence of random variables $T_1, .., T_n$ with the notation T_1^n. Throughout, we use the fact that for any function f, $P_0 f(X_i)|\mathcal{A}_n, \mathcal{Z}_n = P_0 f(X_i)|Z_i$, and that $(X_i, Z_i), i = 1, \ldots, n$ are

i.i.d. random variables.

$$\begin{aligned}
\mathcal{M} &= \frac{1}{n}\sum_{i=1}^{n} I_{a,i}\Big\{\mu_a(X_i) - P_0\mu_a - \hat{\mu}_a(X_i) + P_0\hat{\mu}_a\Big\} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}[I_{a,i} - \pi_a]\big(P_0\mu_a(X_i)|Z_i - P_0\hat{\mu}_a(X_i)|Z_i\big) \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}[I_{a,i} - \pi_a]\big(P_0\mu_a(X_i)|Z_i - P_0\hat{\mu}_a(X_i)|Z_i\big) \\
&\quad + \pi_a\big(P_0\hat{\mu}_a - P_0\mu_a\big) - \pi_a\big(P_0\hat{\mu}_a - P_0\mu_a\big) \\
&= \underbrace{\frac{1}{n}\sum_{i=1}^{n} I_{a,i}\big(\mu_a(X_i) - P_0\mu_a(X_i)|Z_i - \hat{\mu}_a(X_i) + P_0\hat{\mu}_a(X_i)|Z_i\big)}_{\mathcal{M}_1} \\
&\quad + \underbrace{\frac{1}{n}\sum_{i=1}^{n}[I_{a,i} - \pi_a]\big(P_0\mu_a(X_i)|Z_i - P_0\hat{\mu}_a(X_i)|Z_i - P_0\mu_a + P_0\hat{\mu}_a\big)}_{\mathcal{M}_2} \\
&\quad + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\pi_a\big(P_0\mu_a(X_i)|Z_i - P_0\hat{\mu}_a(X_i)|Z_i + P_0\hat{\mu}_a - P_0\mu_a\big)}_{\mathcal{M}_3}.
\end{aligned}$$

For term $\mathcal{M}_1$, note that for a generic $f(X_i)$,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} I_{a,i}P_0 f(X_i)|Z_i &= \frac{1}{n}\sum_{i=1}^{n} I_{a,i}P_0 f(X_i)|\mathcal{A}_n, \mathcal{Z}_n \\
&= \frac{1}{n}\sum_{i=1}^{n} P_0 I_{a,i} f(X_i)|\mathcal{A}_n, \mathcal{Z}_n.
\end{aligned}$$

Furthermore, $\frac{1}{n}\sum_{i=1}^{n} P_0(f(X_i)|\mathcal{A}_n, \mathcal{Z}_n) = \frac{1}{n}\sum_{i=1}^{n} P_0(f(X_i)|Z_i)$.

$$\begin{aligned}
\mathcal{M}_1 &= \frac{1}{n}\sum_{i=1}^{n} I_{a,i}\big(\mu_a(X_i) - P_0\mu_a(X_i)|Z_i - \hat{\mu}_a(X_i) + P_0\hat{\mu}_a(X_i)|Z_i\big) \\
&= -(P_n - P_0^n)(1_a\hat{\mu}_a - 1_a\mu_a).
\end{aligned}$$

By Lemma B.11, we know that this term is $o_P(1/\sqrt{n})$. For the second term $\mathcal{M}_2$, we will exploit the fact that we have assumed that there are finitely many strata levels $l \in \{1, ..., L\}$, $L < \infty$. We have:

$$\begin{aligned}\mathcal{M}_2 &= \frac{1}{n}\sum_{i=1}^{n}[I_{a,i} - \pi_a]\big(P_0\mu_a(X_i)|Z_i - P_0\hat{\mu}_a(X_i)|Z_i - P_0\mu_a + P_0\hat{\mu}_a\big) \\ &= \frac{1}{n}\sum_{l=1}^{L}\sum_{i:Z_i=l}[I_{a,i} - \pi_a]\big((P_0\mu_a(X_i)|Z_i = l) - (P_0\hat{\mu}_a(X_i)|Z_i = l) - P_0\mu_a + P_0\hat{\mu}_a\big).\end{aligned}$$

Note that because we have broken this up by strata level l, we no longer have dependence of $P_0\mu_a(X_i)|Z_i$ and $P_0\hat{\mu}_a(X_i)|Z_i$ on the index i. Therefore, it can be written as:

$$\mathcal{M}_2 = \bigg(\sum_{l=1}^{L}\sum_{i:Z_i=l}\frac{1}{n}[I_{a,i} - \pi_a]\bigg)\big(P_0(\hat{\mu}_a - \mu_a) - P_0(\hat{\mu}_a(X_i) - \mu_a(X_i))|Z_i = l\big).$$

By assumption, we have $P_0(\hat{\mu}_a - \mu_a)^2 \to 0$ in probability. Applying $L(P)$ norm inequalities, $P_0(\hat{\mu}_a - \mu_a) \leq \|\hat{\mu}_a - \mu_a\|_{L^1(P_0)} \leq \|\hat{\mu}_a - \mu_a\|_{L^2(P_0)} \to 0$ in probability. We have a similar term conditional on $Z_i = l$. Here, we note that using iterated expectation, we have

$$P_0(\hat{\mu}_a - \mu_a)^2 = \sum_{l=1}^{L} P(Z_i = l)(P_0(\hat{\mu}_a - \mu_a)^2|Z_i = l).$$

Since the LHS is $o_P(1)$, and since $P(Z_i = l) > 0$ for all l, and $(P_0(\hat{\mu}_a - \mu_a)^2|Z_i = l)$ is a non-negative term, we know that within each strata level l, that $P_0(\hat{\mu}_a - \mu_a)^2|Z_i = l$ must also be $o_P(1)$. Otherwise we would contradict the LHS begin $o_P(1)$. Therefore, we have

$$\begin{aligned}\mathcal{M}_2 &= o_P(1)\bigg(\sum_{l=1}^{L}\sum_{i:Z_i=l}\frac{1}{n}[I_{a,i} - \pi_a]\bigg) \\ &= o_P(1)\bigg(\sum_{l=1}^{L}[n_a(l)/n - \pi_a(n_l/n)]\bigg). \qquad \text{(B.3)}\end{aligned}$$

For each strata level l, we have assumed that $n_a(l)/n_l - \pi_a = O_p(1/\sqrt{n})$. Also, $n_l/n \to P(Z = l)$, in probability. Since there are finitely many L, we have $\mathcal{M}_2 = o_P(1)LO_p(1/\sqrt{n}) = o_P(1/\sqrt{n})$.

For $\mathcal{M}_3$, we have

$$\mathcal{M}_3 = \pi_a \frac{1}{n} \sum_{i=1}^{n} \left\{ P_0\mu_a(X_i)|Z_i - P_0\hat{\mu}_a(X_i)|Z_i \right\} + \pi_a P_0(\hat{\mu}_a - \mu_a).$$

Note that in the sum above, we have only i.i.d. terms, because (X_i, Z_i) are i.i.d. random variables. Define the new random variable $\hat{g}_a(Z_i) = \int \hat{\mu}_a(X_i)dP_0(X_i|Z_i)$, and $g_a(Z_i) = \int \mu_a(X_i)dP_0(X_i|Z_i)$, as in Lemma B.10. Using iterated expectation, we have that $P_0\hat{g}_a = P_0\hat{\mu}_a$, and $P_0 g_a = P_0\mu_a$. Then $\mathcal{M}_3$ can be written as $\pi_a(P_n - P_0)(\hat{g}_a - g_a)$, which is $o_P(1/\sqrt{n})$ by Lemma B.10.

Putting terms $\mathcal{M}_1, \mathcal{M}_2$ and $\mathcal{M}_3$ together we see that $\mathcal{M} = o_P(1/\sqrt{n})$, which means $(I) = o_P(1/\sqrt{n})$.

B.2.3 Tools for Proof of Theorem 2.1 (i)

Here, we prove several lemmas that we use throughout the proof of Theorem 2.1 (i). Throughout this section, we use the following notation: $\rho_P := \|\cdot\|_{L^2(P)}$, and $1_a : A \to 1(A = a)$. Furthermore, we define the following function classes, for $\mathcal{F}_a$ defined as in 2.3:

$$\begin{aligned}
\mathcal{H}_a &:= \{(a', x) \to 1_a(a')\mu_a(X) : \mu_a \in \mathcal{F}_a\} \\
\mathcal{G}_a &:= \{z \to \int \mu_a(X)dP_0(X|Z) : \mu_a \in \mathcal{F}_a\} \\
\mathcal{D}_{\mathcal{F}} &:= \{f_1 - f_2 : f_1, f_2 \in \mathcal{F}\} \quad \text{where } \mathcal{F} \text{ is any function class} \\
\mathcal{D}_{\mathcal{F}}(\delta) &:= \{f_1 - f_2 : f_1, f_2 \in \mathcal{F}, \rho_{P_0}(f_1 - f_2) < \delta\} \quad \text{where } \mathcal{F} \text{ is any function class.}
\end{aligned}$$

Lemma B.1. *For any f, $\int 1_a(A_i)f(X_i)dP(A_i, X_i) = \pi_a \int f(X)dP(X)$.*

Proof. Note that Assumption 2.1 (i) and (ii) imply that A_i is independent of

$\{(Y^{()}i, X_i, Z_i), i = 1, ..., n\}$, and thus $A_i \perp X_i$. This directly implies the result. □

Lemma B.2. *The function class $\mathcal{G}_a$ is P_0-Donsker.*

Proof. First, let $\mathcal{S} := \{z \to s_\theta(z) \equiv \sum_{\ell=1}^{L} \theta_\ell 1(z = \ell) : \theta \in \Theta \subset \mathbb{R}^L\}$. Second, for every strata z, we know that $\int \mu_a(X) dP_0(X|Z) < \infty$ since we have assumed that $\int \mu_a(X) dP_0(X) < \infty$ for all $\mu_a \in \mathcal{F}_a$, and each strata z has positive probability. Therefore, $\mathcal{G}_a \subseteq \mathcal{S}$. Therefore, it suffices to show that $\mathcal{S}$ is P_0-Donsker. This follows by van der Vaart (1998) Example 19.7, because each function $s_\theta(z) \in \mathcal{S}$ is Lipschitz in its L-dimensional indexing parameter θ, i.e.,

$$|s_{\theta_1}(z) - s_{\theta_2}(z)| \leq \sup_\ell |\theta_{1,\ell} - \theta_{2,\ell}| \leq \|\theta_1 - \theta_2\|_2 .$$

□

Lemma B.3. *Define the functions $\hat{g}_a : z \to \int \hat{\mu}_a(X) dP_0(X|Z)$, and $g_a : z \to \int \mu_a(X) dP_0(X|Z)$. Then $P_0(\hat{g}_a - g_a)^2 \leq P_0(\hat{\mu}_a - \mu_a)^2$.*

Proof. Using only properties of iterated expectation and the law of total variance:

$$\begin{aligned}
P_0(\hat{g}_a - g_a)^2 &\equiv \left\| \int \hat{\mu}_a(X) dP_0(X|Z) - \int \mu_a(X) dP_0(X|Z) \right\|^2_{L^2(P_0)} \\
&= \left\| \int \left\{ \hat{\mu}_a(X) - \mu_a(X) \right\} dP_0(X|Z) \right\|^2_{L^2(P_0)} \\
&= P_0 \left(\left[P_0(\mu_a(X) - \mu'_a(X) \big| Z) \right]^2 \right) \\
&\leq P_0 \left[P_0 \left\{ (\mu_a(X) - \mu'_a(X))^2 \big| Z \right\} \right] \\
&= P_0(\hat{\mu}_a - \mu_a)^2,
\end{aligned}$$

where the inequality is from $\left[P_0(\mu_a(X) - \mu'_a(X) \big| Z) \right]^2 \leq P_0 \left\{ (\mu_a(X) - \mu'_a(X))^2 \big| Z \right\}$ due to a conditional version of Jensen's inequality. □

Translating Between Covering Numbers

Lemma B.4. *The covering number for the function class $\mathcal{H}_a$ with respect to ρ_{P_0} is no more than that of the function class $\mathcal{F}_a$ with respect to ρ_{P_0}.*

$$N(\epsilon, \mathcal{H}_a, \rho_{P_0}) \leq N(\epsilon, \mathcal{F}_a, \rho_{P_0}).$$

Proof. By definition, all functions $h_a \in \mathcal{H}_a$ can be written as $1_a\mu_a$ for some $\mu_a \in \mathcal{F}_a$. Let T_ϵ be a minimal ϵ-cover of $\mathcal{F}_a$. Take any $h_a \equiv 1_a\mu_a$. We know that for all μ_a, there exists a $\mu'_a \in T_\epsilon$ such that $\rho_{P_0}(\mu_a - \mu'_a) < \epsilon$. taking $h'_a = 1_a\mu'_a$, we have that

$$\begin{aligned}\rho_{P_0}\left(h_a - h'_a\right)^2 \equiv \int 1_a(\mu_a - \mu'_a)^2 dP_0 \leq \int (\mu_a - \mu'_a)^2 dP_0 \equiv \rho_{P_0}\left(\mu_a - \mu'_a\right)^2 < \epsilon^2 \\ \implies \rho_{P_0}\left(h_a - h'_a\right) < \epsilon.\end{aligned}$$

Thus, the covering number for $\mathcal{H}_a$ is no more than the covering number for $\mathcal{F}_a$. □

Lemma B.5. *With $\delta > 0$, and $\delta' = \delta/\sqrt{\pi_a}$,*

$$N(\epsilon, \mathcal{D}_{\mathcal{H}_a}(\delta), \rho_{P_n}) \leq N(\epsilon, \mathcal{D}_{\mathcal{F}_a}(\delta'), \rho_{P_n}).$$

Proof. This is very similar to Lemma B.4, but requires more care because the size of the function class is dependent on δ, so it is not guaranteed that an $h \in \mathcal{D}_{\mathcal{H}_a}(\delta)$ can be written as $1_a f$ for some $f \in \mathcal{D}_{\mathcal{F}_a}(\delta)$. We need to modify δ for the $\mathcal{D}_{\mathcal{F}_a}$ class.

Let T_ϵ be a minimal ϵ-cover for $\mathcal{D}_{\mathcal{F}_a}(\delta')$. Then $\forall f \in \mathcal{D}_{\mathcal{F}_a}(\delta')$, $\exists f' \in T_\epsilon$ such that $\|f - f'\|_{L^2(P_n)} < \epsilon$. Take an arbitrary $h \in \mathcal{D}_{\mathcal{H}_a}(\delta)$. Then we can write h as $1_a f$ for some $f \in \mathcal{D}_{\mathcal{F}_a}(\delta')$, because (recalling Lemma B.1):

$$\begin{aligned}\rho^2_{P_0}(h) = \rho^2_{P_0}(1_a f) = P_0 1_a f^2 = \pi_a P_0 f^2 = \pi_a \rho^2_{P_0}(f) < \pi_a(\delta')^2 = \delta^2 \\ \implies \rho_{P_0}(h) < \delta.\end{aligned}$$

Furthermore, we have that, using $h' = 1_a f'$ for the f' that covers f,

$$\rho^2_{P_n}(h - h') = \rho^2_{P_n}(1_a(f - f')) = \frac{1}{n}\sum_{i=1}^{n} I_{a,i}(f(X_i) - f'(X_i))^2$$
$$\leq \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - f'(X_i))^2 \leq \rho^2_{P_n}(f - f') < \epsilon^2.$$

Since $\rho_{P_n}(h - h') < \epsilon$, the covering number for $\mathcal{D}_{\mathcal{H}_a}(\delta)$ is no more than that of $\mathcal{D}_{\mathcal{F}_a}(\delta')$. □

Studying a non-standard Empirical Process

In this subsection, we prove lemmas that allow us to study the empirical process $\sqrt{n}(P_n - P_0^n)$.

Lemma B.6. *The following bound holds, for any subset $\mathcal{S} \subseteq \mathcal{D}_{\mathcal{H}_a}$ such that $\mathcal{S}$ is a symmetric function class, i..e, $\mathcal{S} = \mathcal{S} \cup -\mathcal{S}$:*

$$E(\|P_n - P_0^n\|_{\mathcal{S}} \,|\mathcal{A}_n, \mathcal{Z}_n) \leq 2E(\|R_n\|_{\mathcal{S}} \,|\mathcal{A}_n, \mathcal{Z}_n)$$

where R_n is the Rademacher process with ε_i as Rademacher random variables, i.e.,

$$R_n(s) \equiv \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i s(A_i, X_i).$$

Proof. This result is due to the symmetrization argument in van der Vaart and Wellner (1996) Lemma 2.3.1. In typical symmetrization proofs, we have a P_0 and an unconditional expectation. However, since the external and internal expectations are both conditional on $(\mathcal{A}_n, \mathcal{Z}_n)$, and because $\mathcal{X}_n|\mathcal{A}_n, \mathcal{Z}_n$ are mutually independent (but not necessarily identically distributed), the result still holds, as we show. Let X_i' be drawn from the same distribution as X_i, for $i = 1, \ldots, n$. Thus, $E(s(A_i, X_i)|\mathcal{A}_n, \mathcal{Z}_n) = E(s(A_i, X_i')|\mathcal{A}_n, \mathcal{Z}_n)$. For fixed values

$X_1, \ldots, X_n$:

$$\begin{aligned}
\|P_n - P_0^n\|_{\mathcal{S}} &= \sup_{s\in\mathcal{S}} \left| \frac{1}{n}\sum_{i=1}^{n} s(A_i, X_i) - E(s(A_i, X_i')|\mathcal{A}_n, \mathcal{Z}_n) \right| \\
&= \sup_{s\in\mathcal{S}} \frac{1}{n} \left| E\left(\sum_{i=1}^{n} s(A_i, X_i) - s(A_i, X_i') \middle| \mathcal{A}_n, \mathcal{Z}_n \right) \right| \\
&\leq E\left(\sup_{s\in\mathcal{S}} \left| \frac{1}{n}\sum_{i=1}^{n} s(A_i, X_i) - s(A_i, X_i') \right| \middle| \mathcal{A}_n, \mathcal{Z}_n \right).
\end{aligned}$$

Let ϵ_i, $1, \ldots, n$ be independent Rademacher random variables. Let E^* be the outer expectation over both $X_1, \ldots, X_n$ and $X_1', \ldots, X_n'$. Then taking the same conditional expectation over $X_1, \ldots, X_n$ as well,

$$E(\|P_n - P_0^n\|_{\mathcal{S}} |\mathcal{A}_n, \mathcal{Z}_n) \leq E^*\left(\sup_{s\in\mathcal{S}} \left| \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \left[s(A_i, X_i) - s(A_i, X_i') \right] \right| \middle| \mathcal{A}_n, \mathcal{Z}_n \right)$$

because ϵ_i just keeps a function $s(A_i, X_i)$ the same or flips it to its negative value $-s(A_i, X_i)$. Since X_i and X_i' have the same distribution, this does not change the supremum over $\mathcal{S}$. By the triangle inequality,

$$E^*\left(\sup_{s\in\mathcal{S}} \left| \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \left[s(A_i, X_i) - s(A_i, X_i') \right] \right| \middle| \mathcal{A}_n, \mathcal{Z}_n \right) \leq 2E^*\left(\sup_{s\in\mathcal{S}} \left| \frac{1}{n}\sum_{i=1}^{n} \epsilon_i s(A_i, X_i) \right| \middle| \mathcal{A}_n, \mathcal{Z}_n \right).$$

This completes the proof. □

Lemma B.7. *For any $\mathcal{S}$ as defined in Lemma B.6,*

$$E(\|R_n\|_{\mathcal{S}} |\mathcal{X}_n, \mathcal{A}_n, \mathcal{Z}_n) \leq 32 n^{-1/2} \int_0^\infty \sqrt{\log N(\epsilon, \mathcal{S}, \rho_{P_n})} d\epsilon.$$

Proof. This is a direct application of Dudley's entropy integral bound. We first note that with the definition of the Rademacher process in Lemma B.6, $R_n(s)$ can be equivalently expressed as $nR_n(s) = \sum_{i=1}^n \theta_i \varepsilon_i$, where $\theta_i \equiv s(A_i, X_i)$, and $\theta \in \mathbb{T}_n \subseteq \mathbb{R}^n$, where $\mathbb{T}_n =$

$\{(s(A_1, X_1), ..., s(A_n, X_n)) : s \in \mathcal{S}\}$. The Rademacher process is a sub-Gaussian process with respect to the Euclidean metric $\|\cdot\|_2$ and its indexing parameter $\theta \in \mathbb{R}^n$ (see Lemma 2.2.7 and discussion in van der Vaart and Wellner (1996)). Therefore, we can apply Theorem 5.22 from Wainwright (2019) to upper bound the expected supremum (where the expectation is taken over the Rademacher random variables, with the data $(\mathcal{X}_n, \mathcal{A}_n, \mathcal{Z}_n)$ fixed) of the Rademacher process by

$$E(n \|R_n\|_{\mathcal{S}} | \mathcal{X}_n, \mathcal{A}_n, \mathcal{Z}_n) \leq 32 \int_0^\infty \sqrt{\log N(u, \mathbb{T}_n, \|\cdot\|_2)} du.$$

Finally, we note that a minimal ϵ-cover for $\mathbb{T}_n$ with respect to $\|\cdot\|_2$ is an $\epsilon n^{-1/2}$-cover for $\mathcal{S}$ with respect to the pseudometric $L^2(P_n) \equiv \rho_{P_n}$. This is because for any two $\theta, \theta' \in \mathbb{T}_n$ such that $\|\theta - \theta'\|_2 < \epsilon$, we have $\sqrt{n} \|s - s'\|_{L^2(P_n)} = \|\theta - \theta'\|_2 < \epsilon$, where $s, s' \in \mathcal{S}$ are the corresponding functions to the indexing parameters θ, θ'. Therefore,

$$\begin{aligned} E(n \|R_n\|_{\mathcal{S}} | \mathcal{X}_n, \mathcal{A}_n, \mathcal{Z}_n) &\leq 32 \int_0^\infty \sqrt{\log N(un^{-1/2}, \mathcal{S}, \rho_{P_n})} du \\ &\leq 32\sqrt{n} \int_0^\infty \sqrt{\log N(\epsilon, \mathcal{S}, \rho_{P_n})} d\epsilon. \end{aligned}$$

Dividing by n on both sides completes the proof. □

Lemma B.8. *Define* $P_0^n(\cdot) = \frac{1}{n} \sum_{i=1}^n P_0(\cdot | \mathcal{A}_n, \mathcal{Z}_n)$, *and the empirical process* $\mathbb{G}'_n(\cdot) = \sqrt{n}(P_n - P_0^n)(\cdot)$. *Then* $\mathbb{G}'_n$ *is asymptotically uniformly equicontinuous on the class* $\mathcal{H}_a$ *with respect to the pseudometric* ρ_{P_0}.

Proof. To show that $\mathbb{G}'_n$ is asymptotically uniform equicontinuous, we must show that, for any deterministic sequence $\delta_n \to 0$, $\|\mathbb{G}'_n\|_{\mathcal{D}_{\mathcal{H}_a}(\delta_n)} = o_P(1)$. By Markov's inequality, it suffices to show that $E(\|\mathbb{G}'_n\|_{\mathcal{D}_{\mathcal{H}_a}(\delta_n)}) = o(1)$.

By Lemmas B.5, B.6, and B.7, we have that, for any deterministic $\delta_n \to 0$,

$$\begin{aligned}
E(\|\mathbb{G}_n'\|_{\mathcal{D}_{\mathcal{H}_a}(\delta_n)} | \mathcal{A}_n, \mathcal{Z}_n) &= \sqrt{n} \cdot E(\|P_n - P_0^n\|_{\mathcal{D}_{\mathcal{H}_a}(\delta_n)} | \mathcal{A}_n, \mathcal{Z}_n) \\
&\leq 2\sqrt{n} \cdot E(\|R_n\|_{\mathcal{D}_{\mathcal{H}_a}(\delta_n)} | \mathcal{A}_n, \mathcal{Z}_n) && \text{(Lemma B.6)} \\
&\leq 64E\left(\int_0^\infty \sqrt{\log N(\epsilon, \mathcal{D}_{\mathcal{H}_a}(\delta_n), \rho_{P_n})} d\epsilon \Big| \mathcal{A}_n, \mathcal{Z}_n\right) && \text{(Lemma B.7)} \\
&\leq 64E\left(\int_0^\infty \sqrt{\log N(\epsilon, \mathcal{D}_{\mathcal{F}_a}(\delta_n'), \rho_{P_n})} d\epsilon \Big| \mathcal{A}_n, \mathcal{Z}_n\right) && \text{(Lemma B.5).}
\end{aligned}$$

Let $\epsilon_n = \sup_{d \in \mathcal{D}_{\mathcal{F}_a}(\delta_n')} \rho_{P_n}(d)$. Then we only need to integrate to ϵ_n, the diameter of $\mathcal{D}_{\mathcal{F}_a}(\delta_n')$, because beyond that the covering number is 1. Let Q be any finitely supported probability distribution. Then,

$$\begin{aligned}
E(\|\mathbb{G}_n'\|_{\mathcal{D}_{\mathcal{H}_a}(\delta_n)} | \mathcal{A}_n, \mathcal{Z}_n) &\leq 64E\left(\int_0^{\epsilon_n} \sqrt{\log N(\epsilon, \mathcal{D}_{\mathcal{F}_a}(\delta_n'), \rho_{P_n})} d\epsilon \Big| \mathcal{A}_n, \mathcal{Z}_n\right) \\
&\leq 64E\left(\int_0^{\epsilon_n/2} \sup_Q \sqrt{\log N(\epsilon, \mathcal{F}_a, \rho_Q)} d\epsilon \Big| \mathcal{A}_n, \mathcal{Z}_n\right) \\
\implies E(\|\mathbb{G}_n'\|_{\mathcal{D}_{\mathcal{H}_a}(\delta_n)}) &\leq 64E\left(\int_0^{\epsilon_n/2} \sup_Q \sqrt{\log N(\epsilon, \mathcal{F}_a, \rho_Q)} d\epsilon\right)
\end{aligned}$$

where we have used the fact that $N(\epsilon, \mathcal{D}_{\mathcal{F}_a}(\delta_n'), \rho_Q) \leq N(\epsilon/2, \mathcal{F}_a, \rho_Q)^2$. Thus, by an application of the dominated convergence theorem and using 2.3, to show that $E(\|\mathbb{G}_n'\|_{\mathcal{D}_{\mathcal{H}_a}(\delta_n)}) = o(1)$, it suffices to show that $e_n = o_P(1)$, or equivalently that $e_n^2 = o_P(1)$.

$$\begin{aligned}
\epsilon_n^2 &= \sup_{d \in \mathcal{D}_{\mathcal{F}_a}(\delta_n')} P_n d^2 \\
&= \sup_{d \in \mathcal{D}_{\mathcal{F}_a}(\delta_n')} \left\{(P_n - P_0)d^2 + P_0 d^2\right\} \\
&\leq \sup_{d \in \mathcal{D}_{\mathcal{F}_a}(\delta_n')} (P_n - P_0)d^2 + \sup_{d \in \mathcal{D}_{\mathcal{F}_a}(\delta_n')} P_0 d^2 \\
&\leq \sup_{d \in \mathcal{D}^2_{\mathcal{F}_a}(\delta_n')} (P_n - P_0)d + (\delta_n')^2.
\end{aligned}$$

Since $\delta_n \to 0$, which implies $\delta'_n \to 0$, it suffices to show that $\sup_{d \in \mathcal{D}^2_{\mathcal{F}_a}(\delta'_n)}(P_n - P_0)d \to 0$ in probability. To show this, we cite results from Chapter 2.10 *Permanence of the Donsker Property* in van der Vaart and Wellner (1996). Since $\mathcal{F}_a$ is P_0-Donsker, this implies that $\mathcal{D}_{\mathcal{F}_a} \equiv \mathcal{F}_a - \mathcal{F}_a$ is also P_0-Donsker (Theorem 2.10.6), and $\mathcal{D}^2_{\mathcal{F}_a}$ is P_0-Glivenko Cantelli (Lemma 2.10.14). Since, $\mathcal{D}^2_{\mathcal{F}_a}(\delta'_n) \subseteq \mathcal{D}^2_{\mathcal{F}_a}$, this completes the proof. □

Asymptotic Negligibility of Two Empirical Process Terms

We now prove Lemma B.9 that allows us to use the results of previous lemmas to show that the empirical process terms $(P_n - P_0)(\hat{g}_a - g_a)$ (Lemma B.10) and $(P_n - P_0^n)(1_a\hat{\mu}_a - 1_a\mu_a)$ (Lemma B.11) are both asymptotically negligible. Lemma B.9 is a version of van der Vaart (1998) Lemma 19.24 where the estimated function falls into $\mathcal{F}$ with probability tending to 1 rather than deterministically. Note that this automatically holds for $(P_n - P_0)(\hat{\mu}_a - \mu_a)$ since $\hat{\mu}_a$ and μ_a satisfies the Donsker condition of 2.3.

Lemma B.9. *Let $(\mathcal{F}, \rho)$ be totally bounded for some pseudometric ρ such that $\rho(f)^2 \leq \rho_{P_0}(f)^2$. Let $\mathbb{X}_n$ be some empirical process that is asymptotically uniformly equicontinuous on $\mathcal{F}$ with respect to ρ. Then, if $P(\hat{f} \in \mathcal{F}) \to 1$ as $n \to \infty$ and $P_0(\hat{f} - f)^2 \to 0$ in probability, then $\mathbb{X}_n(\hat{f} - f) = o_P(1)$.*

Proof. First, note that for any sequence $\Delta_n = o_P(1)$, we can always find a deterministic sequence $\delta_n \to 0$ such that $P(\Delta_n > \delta_n) \to 0$ as $n \to \infty$. In this case, let $\Delta_n = 2P_0(\hat{f} - f)^2$, and notice that $\rho(\hat{f} - f)^2 < \Delta_n$ by assumption. Define the following events:

$$A_n = \{|\mathbb{X}_n(\hat{f} - f)| > \epsilon\}$$
$$B_n = \{\|\mathbb{X}_n\|_{\mathcal{D}_{\mathcal{F}}(\delta_n)} > \epsilon\}$$
$$C_n = \{\Delta_n \leq \delta_n\}$$
$$D_n = \{\hat{f} \in \mathcal{F}\}$$

Then we can construct the following inequality:

$$\begin{aligned} P(A_n) &\leq P(A_n \cap D_n \cap C_n) + P(C_n^c) + P(D_n^c) \\ &\leq P(B_n) + P(C_n^c) + P(D_n^c). \end{aligned}$$

$P(B_n) \to 0$ by the asymptotic uniformly equicontinuous property of $\mathbb{X}_n$ on $\mathcal{F}$, and $P(C_n^c)$ and $P(D_n^c)$ both $\to 0$ as $n \to \infty$ by assumption. Thus, $\mathbb{X}_n(\hat{f} - f) = o_P(1)$. □

Remark B.1. *Note that Lemma B.9 holds if $\mathcal{F}$ is a Donsker class with $\mathbb{X}_n \equiv \mathbb{G}_n \equiv \sqrt{n}(P_n - P_0)$, the standard empirical process, by definition of a Donsker class. This is a version of van der Vaart (1998), Lemma 19.24, where the only difference is that $\hat{f}$ falls into $\mathcal{F}$ with probability tending to 1, rather than deterministically.*

Lemma B.10. *Define the new random variables $\hat{g}_a(Z_i) = \int \hat{\mu}_a(X) dP_0(X|Z_i)$, and $g_a(Z_i) = \int \mu_a(X) dP_0(X|Z_i)$. Under the conditions in 2.3, $(P_n - P_0)(\hat{g}_a - g_a) = o_P(1/\sqrt{n})$.*

Proof. Note that Z_i are i.i.d. random variables, and that $\mathcal{G}_a$ is P_0-Donsker by Lemma B.2. Furthermore, $P_0(\hat{g}_a \in \mathcal{G}_a) \geq P_0(\hat{\mu}_a \in \mathcal{F}_a) \to 1$, by 2.3, and $P_0(\hat{g}_a - g_a)^2 \leq P_0(\hat{\mu}_a - \mu_a)^2 \to 0$, combining Lemma B.3 with 2.3. Since $\mathcal{G}_a$ is P_0-Donsker, we know that $\mathbb{X}_n := \sqrt{n}(P_n - P_0)$ is as asymptotically uniform equicontinuous on $\mathcal{G}_a$ using the standard deviation pseudometric, which is guaranteed to exist: $\rho(f) := \sqrt{P_0(f - P_0 f)^2} \leq \rho_{P_0}(f)$. Then using Remark B.1 with $\mathcal{F} := \mathcal{G}_a$, we conclude that $\mathbb{X}_n(\hat{g}_a - g_a) = o_P(1)$, which implies $(P_n - P_0)(\hat{g}_a - g_a) = o_P(1/\sqrt{n})$. □

Lemma B.11. *Under the conditions in 2.3, $(P_n - P_0^n)(1_a\hat{\mu}_a - 1_a\mu_a) = o_P(1/\sqrt{n})$.*

Proof. By 2.3, $\mathcal{F}_a$ is P_0-Donsker with respect to ρ_{P_0}, which tells us that $(\mathcal{F}_a, \rho_{P_0})$ is totally bounded. Applying Lemma B.4, we have that $(\mathcal{H}_a, \rho_{P_0})$ is also totally bounded. From Lemma B.8, we know that the empirical process $\mathbb{G}'_n := \sqrt{n}(P_n - P_0^n)$ is asymptotically uniformly equicontinuous on $\mathcal{H}_a$.

Let $h_a = 1_a\mu_a$, and $\hat{h}_a = 1_a\hat{\mu}_a$. Then, $P_0(\hat{h}_a \in \mathcal{H}_a) \geq P_0(\hat{\mu}_a \in \mathcal{F}_a) \to 1$ by 2.3, and as in Lemma B.4, $P_0(\hat{h}_a - h_a)^2 \leq P_0(\hat{\mu}_a - \mu_a)^2 \to 0$ in probability also by 2.3. Then, applying

Lemma B.9 using $\mathbb{X}_n := \mathbb{G}'_n$, $\mathcal{F} := \mathcal{H}_a$, and $\rho := \rho_{P_0}$, we have $\mathbb{G}'_n(1_a\hat{\mu}_a - 1_a\mu_a) = o_P(1)$, which implies that $(P_n - P_0^n)(1_a\hat{\mu}_a - 1_a\mu_a) = o_P(1/\sqrt{n})$. □

B.2.4 Proof of Theorem 2.2: AL under Cross-Fitting

We now show that the influence function of the cross-fitted estimator in equation (4) in the main text under Assumptions 1 and 2* is the same as equation (B.1). As a result, all of our other theoretical results hold. The result of the Theorem 3.1 in Chernozhukov et al. (2017) is that $\sqrt{n}(\hat{\theta}_{\text{CF},a} - \theta_a) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_a(W_i; \pi_a, \mu_a) + o_P(1)$. We now follow the proof outline of Theorem 3.1 in Chernozhukov et al. (2017), built off of our Lemma B.16 that has modifications to the Step 3 of their theorem that are needed to allow for covariate-adaptive randomization. *Our Lemma B.16 is the cross-fitting version of showing that term (I) in equation* (B.2) *is asymptotically negligible.*

Here it will be convenient to reparameterize the function ψ_a. In this section, we will write

$$\psi(W_i; \theta_a, \eta_a) = \frac{I_{a,i}}{\pi_a}\Big\{Y_{a,i} - \mu_a^*(X_i)\Big\} + \mu_a^*(X_i) - \theta_a \tag{B.4}$$

where the nuisance parameter $\eta_a = (\pi_a, \mu_a^*)$, and $\mu_a^*(X) = \mu_a(X) - P_0\mu_a + \theta_a$. So there is an implicit dependence between the nuisance parameter μ_a and θ_a. Note that in Section B.2.2, $\psi_a(W_i; \pi_a, \mu_a) = \psi(W_i; \theta_a, \eta_a)$. Recall that $\mathcal{I}_1, ..., \mathcal{I}_J$ is a random partition of $\{1, ..., n\}$ with a fixed integer J (and where each set is of nearly equal size), and $\mathcal{I}_j^c$ is the set containing indices in $\{1, ..., n\}$ but not in $\mathcal{I}_j$. For each $j = 1, ..., J$, we use data with indices in $\mathcal{I}_j^c$ to obtain $\hat{\mu}_{a,j}(X)$ as an estimate of $E(Y_a \mid X)$. Recall that $n_{a,j}$ is the number of indices in $\mathcal{I}_j$ with $A_i = a$, $n_j = \sum_{a=1}^{K} n_{a,j}$, and let $\hat{\pi}_{a,j} = n_{a,j}/n_j$. Let

$$\check{\theta}_{a,j} = \frac{1}{n_j}\sum_{i \in \mathcal{I}_j}\left[\frac{I_{a,i}}{\hat{\pi}_{a,j}}\{Y_{a,i} - \hat{\mu}_{a,j}(X_i)\} + \hat{\mu}_{a,j}(X_i)\right].$$

This is the AIPW estimator for partition j, but where $\hat{\mu}_{a,j}$ is estimated based on data not in partition j, and $\hat{\pi}_{a,j}$ is the treatment proportion that is partition-specific. This is the

only difference from the main text equation (1). However, we show in Remark B.2 that we can replace $\hat{\pi}_{a,j}$ with $\hat{\pi}_a$ and have the same asymptotic properties. We can write, for any partition j,

$$\begin{aligned}\sqrt{n_j}\left(\check{\theta}_{a,j} - \theta_a\right) &= \sqrt{n_j}\left(E_{n,j}[\psi(W;\theta_a,\hat{\eta}_{a,j}]\right) \\ &= \frac{1}{\sqrt{n_j}}\sum_{i\in\mathcal{I}_j}\psi(W_i;\theta_a,\eta_a) + \mathcal{R}_j\end{aligned}$$

where $\mathcal{R}_j$ is from Lemma B.16. Now we can take the average across all J partitions to get the cross-fitted estimator on the LHS, and the desired influence function on the RHS:

$$\begin{aligned}\sqrt{n}(\hat{\theta}_a - \theta_a) &= \sqrt{n}\left(\frac{1}{J}\sum_{j=1}^{J}\check{\theta}_{a,j} - \theta_a\right) \\ &= \sum_{j=1}^{J}\frac{\sqrt{n_j}}{\sqrt{J}}\left(\check{\theta}_{a,j} - \theta_a\right) \\ &= \sum_{j=1}^{J}\sum_{i\in I_j}\frac{1}{\sqrt{n}}\psi(W_i;\theta_a,\eta_a) + \frac{1}{\sqrt{J}}\sum_{k=1}^{J}\mathcal{R}_j \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(W_i;\theta_a,\eta_a) + o_P(1)\end{aligned}$$

since each $\mathcal{R}_j = o_P(1)$ (Lemma B.16), their sum (over a finite number of partitions) is also $o_P(1)$.

Remark B.2. *Theorem 2.2 holds replacing $\hat{\pi}_{a,j}$ with $\hat{\pi}_a$ in $\hat{\theta}_{\mathrm{CF,a}}$ in (4). That is,*

$$\frac{1}{J}\sum_{j=1}^{J}\frac{1}{n_j}\sum_{i\in\mathcal{I}_j}\left[\frac{I_{a,i}}{\hat{\pi}_a}\left\{Y_{a,i} - \hat{\mu}_{a,j}(X_i)\right\} + \hat{\mu}_{a,j}(X_i)\right] \tag{B.5}$$

has the same influence function given in (B.1)*. This is because in each step of the proof of Theorem 2.2, we could have equivalently used $\hat{\pi}_a$ without changing the result, asymptotically. Specifically, since we condition on the full $(\mathcal{A}_n, \mathcal{Z}_n)$ in the proof, we could have just as easily removed $\hat{\pi}_a$ from the conditional expectation as we did $\hat{\pi}_{a,j}$. Also, $O_p(n_j^{-1/2}) \equiv O_p(n^{-1/2})$,*

since n_j/n is constant as n grows.

Lemma B.12 (Equivalence of Cross-Fitted AIPW Estimators in Simulation)**.** *Let $\check{\mu}_a(X_i) = \sum_{j=1}^{J} I\{i \in \mathcal{I}_j\}\hat{\mu}_{a,j}(X_i)$ be the estimated $\mu_a(X_i)$ from AIPW using cross-fitting with J partitions. Then if the partitions are equally sized, the AIPW estimator in (1) with $\check{\mu}_a$ is equivalent to the cross-fitted AIPW estimator in (4).*

Proof. The AIPW estimator using $\check{\mu}_a(X_i) = \sum_{j=1}^{J} I\{i \in \mathcal{I}_j\}\hat{\mu}_{a,j}(X_i)$ is

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{I_{a,i}}{\hat{\pi}_a}\left\{Y_{a,i} - \check{\mu}_a(X_i)\right\} + \check{\mu}_a(X_i)\right]. \tag{B.6}$$

With the additional assumption of equally sized partitions, we have

$$\begin{aligned}
\text{(B.6)} &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{I_{a,i}}{\hat{\pi}_a}\left\{Y_{a,i} - \sum_{j=1}^{J} I\{i \in \mathcal{I}_j\}\hat{\mu}_{a,j}(X_i)\right\} + \sum_{j=1}^{J} I\{i \in \mathcal{I}_j\}\hat{\mu}_{a,j}(X_i)\right] \\
&= \frac{1}{J}\frac{1}{n_j}\sum_{j=1}^{J}\sum_{i\in\mathcal{I}_j}\left[\frac{I_{a,i}}{\hat{\pi}_a}\left\{Y_{a,i} - \sum_{j=1}^{J} I\{i \in \mathcal{I}_j\}\hat{\mu}_{a,j}(X_i)\right\} + \sum_{j=1}^{J} I\{i \in \mathcal{I}_j\}\hat{\mu}_{a,j}(X_i)\right] \\
&= \frac{1}{J}\sum_{j=1}^{J}\frac{1}{n_j}\sum_{i\in\mathcal{I}_j}\left[\frac{I_{a,i}}{\hat{\pi}_a}\left\{Y_{a,i} - \hat{\mu}_{a,j}(X_i)\right\} + \hat{\mu}_{a,j}(X_i)\right] = \text{(B.5)}.
\end{aligned}$$

□

B.2.5 Tools for Proof of Theorem 2.2

Here we include several lemmas that we use to show the asymptotic linearity of the cross-fitted AIPW estimator. These are used in Lemma B.16, which is the key technical result for cross-fitting with dependent data from CAR.

Lemma B.13. *Let $\mu_a^*(X) = \mu_a(X) - P_0\mu_a(X) + \theta_a$. Let*

$$\mathcal{F}_{a,n}^* = \{x \to \mu(X) - P_0\mu + \theta_a : \mu \in \mathcal{F}_{a,n}\}.$$

Then $\sup_{\mu \in \mathcal{F}_{a,n}} E([\mu(X) - \mu_a(X)]^2) = o_P(1)$ *implies* $\sup_{\mu^* \in \mathcal{F}^*_{a,n}} E([\mu^*(X) - \mu^*_a(X)]^2) = o_P(1)$.

Proof. Let $P_{n,a}(\cdot) = \frac{1}{n_a} \sum_{i:A_i=a}(\cdot)$. We can write, for any $\mu^* \in \mathcal{F}^*_{a,n}$,

$$\begin{aligned} E([\mu^*(X) - \mu^*_a(X)]^2) &= E([\mu(X) - \mu_a(X) + P_0(\mu_a - \mu)]^2) \\ &\leq C \cdot \left\{ E([\mu(X) - \mu_a(X)]^2) + E([P_0\mu - P_0\mu_a]^2) \right\} \\ &= C \cdot \left\{ E([\mu(X) - \mu_a(X)]^2) + [P_0(\mu - \mu_a)]^2 \right\} \\ &\leq 2C \cdot \left\{ E([\mu(X) - \mu_a(X)]^2) \right\} \end{aligned}$$

for some $\mu \in \mathcal{F}_{a,n}$ and some constant C, where the last step is due to Jensen's Inequality. Taking a supremum over $\mu^* \in \mathcal{F}^*_{a,n}$ on the LHS is equivalent to taking a supremum over $\mu \in \mathcal{F}_{a,n}$ on the RHS. Thus, to show that $\sup_{\mu^* \in \mathcal{F}^*_{a,n}} E([\mu^*(X) - \mu^*_a(X)]^2) = o_P(1)$, it suffices to show that $\sup_{\mu \in \mathcal{F}_{a,n}}$ over the RHS is $o_P(1)$, which is true by assumption. □

Lemma B.14. *The condition that* $\sup_{\mu^* \in \mathcal{F}^*_{a,n}} E([\mu^*(X) - \mu^*_a(X)]^2) = o_P(1)$ *implies that for each strata* Z*,*

$$\sup_{\mu^* \in \mathcal{F}^*_{a,n}} E([\mu^*(X) - \mu^*_a(X)]^2 | Z) = o_P(1).$$

Proof. We use proof by contradiction. Assume that $\sup_{\mu^* \in \mathcal{F}^*_{a,n}} E([\mu^*(X) - \mu^*_a(X)]^2) = o_P(1)$, and that there exists a $\delta > 0$ and a strata Z' such that for all $\epsilon > 0$,

$$\lim_{n \to \infty} P\left(\sup_{\mu^* \in \mathcal{F}^*_{a,n}} E([\mu^*(X) - \mu^*_a(X)]^2 | Z') > \epsilon \right) > \delta.$$

Furthermore, by iterated expectation, $E([\mu^*(X) - \mu^*_a(X)]^2) = \sum_{l=1}^{L} E([\mu^*(X) - \mu^*_a(X)]^2 | Z = l) > E([\mu^*(X) - \mu^*_a(X)]^2 | Z') P(Z')$. Since each strata has positive probability, there exists a $\rho > 0$ such that $E([\mu^*(X) - \mu^*_a(X)]^2) > E([\mu^*(X) - \mu^*_a(X)]^2 | Z')\rho$. Therefore, for all $\epsilon' > 0$ (where we can re-parameterize $\epsilon' = \epsilon \cdot \rho$),

$$P\left(\sup_{\mu^* \in \mathcal{F}^*_{a,n}} E([\mu^*(X) - \mu^*_a(X)]^2) > \epsilon' \right) \geq P\left(\sup_{\mu^* \in \mathcal{F}^*_{a,n}} E([\mu^*(X) - \mu^*_a(X)]^2 | Z') > \epsilon \right).$$

Taking the limit of both sides as $n \to \infty$, we have that the RHS is $> \delta$ by assumption, for all $\epsilon > 0$. Therefore, the limit of the LHS is also $> \delta$, meaning that $\sup_{\mu^* \in \mathcal{F}^*_{a,n}} E([\mu^*(X) - \mu^*_a(X)]^2)$ is not $o_P(1)$. This implies that if $\sup_{\mu^* \in \mathcal{F}^*_{a,n}} E([\mu^*(X) - \mu^*_a(X)]^2) = o_P(1)$, then each strata Z must also have $\sup_{\mu^* \in \mathcal{F}^*_{a,n}} E([\mu^*(X) - \mu^*_a(X)]^2|Z) = o_P(1)$. □

Lemma B.15. *Let $\hat{\mu}_{a,j}$ be a function such that, with probability $1 - \Delta_n$, $\|\hat{\mu}_{a,j} - \mu_a\|_{L^2(P_0)} \leq \delta_n$, where $\delta_n = o(1)$ and $\Delta_n = o(1)$. Then with probability $1 - \Delta_n$, $\hat{\mu}_{a,j} \in \mathcal{F}_{a,n}$, and*

$$\sup_{\mu \in \mathcal{F}_{a,n}} \|\hat{\mu}_{a,j} - \mu_a\|^2_{L^2(P_0)} \equiv \sup_{\mu \in \mathcal{F}_{a,n}} E([\hat{\mu}_{a,j}(X) - \mu_a(X)_{L^2(P_0)}]^2) = o(1)$$

where $\mathcal{F}_{a,n} = \{\mu : \|\mu - \mu_a\|_{L^2(P_0)} \leq \delta_n\}$.

Proof. First, by definition, note that since with probability $1 - \Delta_n$, $\|\hat{\mu}_{a,j} - \mu_a\|_{L^2(P_0)} \leq \delta_n$, with probability $1 - \Delta_n$, $\hat{\mu}_{a,j} \in \mathcal{F}_{a,n}$ since it satisfies the condition of $\mathcal{F}_{a,n}$. Furthermore, for any μ in $\mathcal{F}_{a,n}$, we know that $\|\mu - \mu_a\|_{L^2(P_0)} \leq \delta_n$. Therefore, the supremum over $\mu \in \mathcal{F}_{a,n}$ also satisfies $\sup_{\mu \in \mathcal{F}_{a,n}} \|\mu - \mu_a\|_{L^2(P_0)} \leq \delta_n$ (because it is a compact set). Finally, since $\sup_{\mu \in \mathcal{F}_{a,n}} \|\hat{\mu}_{a,j} - \mu_a\|_{L^2(P_0)} \leq \delta_n$, $\sup_{\mu \in \mathcal{F}_{a,n}} \|\mu_{a,j} - \mu_a\|^2_{L^2(P_0)} \leq \delta_n^2$. Since $\delta_n = o(1)$, $\delta_2^n = o(1)$ as well. This completes the proof. □

Lemma B.16. *Defining $\psi(W_i; \theta_a, \eta_a)$ as in* (B.4)*, we have the following:*

$$\mathcal{R}_j := \frac{1}{\sqrt{n}_j} \sum_{i \in \mathcal{I}_j} \{\psi(W_i; \theta_a, \hat{\eta}_{a,j}) - \psi(W_i; \theta_a, \eta_a)\} = o_P(1).$$

Proof. Here we show asymptotic negligibility for the difference in Averages of Estimating Function Evaluated at $\hat{\eta}_{a,j}$ versus η_a. This roughly follows the proof in Chernozhukov et al. (2017) for Step 3 of Theorem 3.1, but we define a different empirical process notation that allows us to have conditional independence in the setting of covariate adaptive randomization. Finally, in what follows, we are considering the following probabilities *on the event that* $\hat{\mu}_{a,j} \in \mathcal{F}_{a,n}$, where $\mathcal{F}_{a,n}$ is defined in Lemma B.15. Recall that by Assumption 2.4 combined with Lemma B.15, this occurs with probability $1 - \Delta_n$, with $\Delta_n \to 0$. Therefore, asymptotically, the following results hold.

We define the following empirical process term:

$$\mathbb{G}^*_{n,j}[\psi(W;\theta,\eta)] = \frac{1}{\sqrt{n}_j}\sum_{i\in\mathcal{I}_j}\Big(\psi(W_i;\theta,\eta) - E\big\{\psi(W_i;\theta,\eta)\big|(W_i)_{i\in\mathcal{I}_j^c},\mathcal{A}_n,\mathcal{Z}_n\big\}\Big) \tag{B.7}$$

The second term is similar to how P_0^n is defined in the proof of Theorem 2.1 (i), but where we restrict the sum to the jth partition, and condition on the auxiliary sample $(W_i)_{i\in\mathcal{I}_j^c}$, as well as all of the treatment indicators and strata $(\mathcal{A}_n,\mathcal{Z}_n)$. We aim to show that, for any j,

$$\mathcal{R}_j := \frac{1}{\sqrt{n}_j}\sum_{i\in\mathcal{I}_j}\{\psi(W_i;\theta_a,\hat{\eta}_{a,j}) - \psi(W_i;\theta_a,\eta_a)\} = o_P(1).$$

This term can be broken up into two parts, using the notation in (B.7). Define:

$$\begin{aligned}
\mathcal{R}_{j,1} &:= \mathbb{G}^*_{n,j}[\psi(W;\theta_a,\hat{\eta}_{a,j})] - \mathbb{G}^*_{n,j}[\psi(W;\theta_a,\eta_a)] \\
\mathcal{R}_{j,2} &:= \frac{1}{\sqrt{n}_j}\sum_{i\in\mathcal{I}_j}\Big(E\big\{\psi(W_i;\theta_a,\hat{\eta}_{a,j}) - \psi(W_i;\theta_a,\eta_a)\big|(W_i)_{i\in\mathcal{I}_j^c},\mathcal{A}_n,\mathcal{Z}_n\big\}\Big) \\
&= \frac{1}{\sqrt{n}_j}\sum_{i\in\mathcal{I}_j}\Big(E\big\{\psi(W_i;\theta_a,\hat{\eta}_{a,j})\big|(W_i)_{i\in\mathcal{I}_j^c},\mathcal{A}_n,\mathcal{Z}_n\big\} - E\big\{\psi(W_i;\theta_a,\eta_a)\big|\mathcal{A}_n,\mathcal{Z}_n\big\}\Big).
\end{aligned}$$

Then it suffices to show that $\mathcal{R}_{j,1} = o_P(1)$ and $\mathcal{R}_{j,2} = o_P(1)$ because $|\mathcal{R}_j| \leq |\mathcal{R}_{j,1}| + |\mathcal{R}_{j,2}|$. Starting with $\mathcal{R}_{j,1}$, define $\phi(w) := \psi(w;\theta_a,\hat{\eta}_{a,j}) - \psi(w;\theta_a,\eta_a)$. Then, defining the short-hand notation $E^*(\cdot) = E(\cdot|(W_i)_{i\in\mathcal{I}_j},\mathcal{A}_n,\mathcal{Z}_n)$, we have

$$\begin{aligned}
E^*\{\mathcal{R}_{j,1}^2\} &= E^*\{|\mathbb{G}^*_{n,j}[\phi(W)]|^2\} \\
&= \frac{1}{n_j}E^*\left\{\left(\sum_{i\in\mathcal{I}_j}(\phi(W_i) - E^*(\phi(W_i)))\right)^2\right\}.
\end{aligned}$$

Notice that if we expand the square, the cross-terms cancel out through conditioning on $(\mathcal{A}_n,\mathcal{Z}_n)$ in E^* so that elements (W_i,W_j) are independent. The cross-terms are $[E^*(\phi(W_i) - E^*(\phi(W_i))]\,[E^*(\phi(W_j) - E^*(\phi(W_j))]$ which are zero. Therefore, we can con-

tinue the above as:

$$\begin{aligned} E^*\{\mathcal{R}_{j,1}^2\} &= \frac{1}{n_j} E^* \left\{ \sum_{i \in \mathcal{I}_j} (\phi(W_i) - E^*(\phi(W_i)))^2 \right\} \\ &\leq \frac{1}{n_j} \sum_{i \in \mathcal{I}_j} E^*(\phi(W_i)^2) \\ &\leq \frac{1}{n_j} \sum_{i \in \mathcal{I}_j} \sup_{\eta \in \mathcal{T}_N} E(\phi(W_i)^2 | \mathcal{A}_n, \mathcal{Z}_n). \end{aligned}$$

We would like to show that the summation above is $o_P(1)$. We do this by breaking up $\phi(W_i)^2$. Note first that

$$\begin{aligned} \underbrace{[\psi(W;\theta_a,\eta) - \psi(W;\theta_a,\eta_a)]}_{\upsilon(\eta,\eta_a)}{}^2 \leq \underbrace{[I_a Y_a[\pi^{-1} - \pi_a^{-1}]]^2}_{\upsilon_1(\eta,\eta_a)} + \underbrace{[\mu^*(X) - \mu_a^*(X)]^2}_{\upsilon_2(\eta,\eta_a)} \\ + \underbrace{[I_a[\mu^*(X)\pi^{-1} - \mu_a^*(X)\pi_a^{-1}]]^2}_{\upsilon_3(\eta,\eta_a)}. \end{aligned}$$

We will bound each of the υ terms separately. Since $\pi_a > 0$ for all a, we know that there exists an $\epsilon > 0$ such that $|\mu^*(X)\pi^{-1} - \mu_a^*(X)\pi_a^{-1}| < \epsilon^{-2}|\mu^*(X)\pi_a - \mu_a^*(X)\pi| = \epsilon^{-2}|\pi_a(\mu^*(X) - \mu_a^*(X)) + \mu_a^*(\pi - \pi_a)|$. By the generalized mean inequality, $\exists C > 0$ such that

$$\begin{aligned} (\epsilon^{-2}|\pi_a(\mu^*(X) - \mu_a^*(X)) + \mu_a(\pi - \pi_a)|)^2 &\leq \epsilon^{-4} C \left(\pi_a^2(\mu^*(X) - \mu_a^*(X))^2 + \mu_a^2(X)(\pi - \pi_a)^2 \right) \\ &\leq \epsilon^{-4} C \left(\mu^*(X) - \mu_a^*(X))^2 + \mu_a^2(X)(\pi - \pi_a)^2 \right). \end{aligned}$$

Turning to each of the υ terms, and recalling that we use $\hat{\pi}_a$ (in the whole sample) as our estimator for π_a (and that it can be brought out of the conditional expectation because it is

deterministic conditional on the whole set of indicators $(\mathcal{A}_n)$), we have:

$$(1) \quad \frac{1}{n_j}\sum_{i\in\mathcal{I}_j}\sup_{\eta\in\mathcal{T}_n} E(\upsilon_1|\mathcal{A}_n,\mathcal{Z}_n) \leq [\hat{\pi}_{a,j}^{-1}-\pi_a^{-1}]^2\left[\frac{1}{n_j}\sum_{i\in\mathcal{I}_j}E([Y_{a,i}]^2|Z_i)\right]$$

$$(2) \quad \frac{1}{n_j}\sum_{i\in\mathcal{I}_j}\sup_{\eta\in\mathcal{T}_n} E(\upsilon_2|\mathcal{A}_n,\mathcal{Z}_n) \leq \frac{1}{n_j}\sum_{i\in\mathcal{I}_j}\sup_{\mu^*\in\mathcal{F}^*_{a,n}} E([\mu^*(X_i)-\mu_a^*(X_i)]^2|Z_i)$$

$$(3) \quad \frac{1}{n_j}\sum_{i\in\mathcal{I}_j}\sup_{\eta\in\mathcal{T}_n} E(\upsilon_3|\mathcal{A}_n,\mathcal{Z}_n) \leq \frac{1}{n_j}\sum_{i\in\mathcal{I}_j}\sup_{\mu^*\in\mathcal{F}^*_{a,n}} \epsilon^{-4}CE[\mu^*(X_i)-\mu_a^*(X_i))^2|Z_i]$$
$$+(\hat{\pi}_{a,j}-\pi_a)^2\frac{1}{n_j}\sum_{i\in\mathcal{I}_j}\epsilon^{-4}CE[\mu_a^2(X_i)|Z_i]$$

For (1) to be $o_P(1)$, it suffices that $\hat{\pi}_{a,j}^{-1}\to\pi_a$ in probability (Assumption 2.1), and for all strata Z, $E([Y_a]^2|Z)<\infty$. This must hold since we assume that Y has a finite second-order moment, unconditionally. For (2) to be $o_P(1)$, it suffices that for each strata Z, $\sup_{\mu^*\in\mathcal{F}^*_{a,n}} E([\mu^*(X_i)-\mu_a^*(X_i)]^2|Z_i)=o_P(1)$. This holds by Assumption 2.4 combined with Lemmas B.13 and B.14. For (3) to be $o_P(1)$ the preceding two conditions must hold, with the additional condition that $E(\mu_a^*(X)^2|Z)<\infty$ for each strata Z (Assumption 2.4 specifies this unconditionally for μ_a, which implies that $E(\mu_a^*(X)^2|Z)<\infty$ as well; it must also hold conditionally by strata, since each strata has positive probability). Therefore, $E^*\{\mathcal{R}_{j,1}^2\}=o_P(1)$. By Lemma 6.1 in Chernozhukov et al. (2017), this implies that $\mathcal{R}_{j,1}=o_P(1)$.

For $\mathcal{R}_{j,2}$, we define the function for $r\in[0,1]$:

$$f_{i,j}(r)=E\big\{\psi(W_i;\theta_a,\eta_a+r(\hat{\eta}_{a,j}-\eta_a))\big|(W_i)_{i\in\mathcal{I}_j^c},\mathcal{A}_n,\mathcal{Z}_n\big\}-E\big\{\psi(W_i;\theta_a,\eta_a)\big|\mathcal{A}_n,\mathcal{Z}_n\big\}$$

and define $\mathcal{R}_{j,2}(r)=\frac{1}{\sqrt{n_j}}\sum_{i\in\mathcal{I}_j}f_{j,i}(r)$. Then $\mathcal{R}_{j,2}=\mathcal{R}_{j,2}(1)$. We take a Taylor expansion of $\mathcal{R}_{j,2}(r)$ around $r=0$ by first expanding $f_{j,i}(r)$ as follows, where we note that we can remove that conditioning on $(W_i)_{i\in\mathcal{I}_j^c}$ whenever we do not have a $\hat{\eta}_{a,j}$ since the observations $(W_i)_{i\in\mathcal{I}_j}$

are independent through conditioning on $(\mathcal{A}_n, \mathcal{Z}_n)$:

$$
\begin{aligned}
f_{j,i}(0) &= E\big\{\psi(W_i;\theta_a,\eta_a)\big|(W_i)_{i\in\mathcal{I}_j^c},\mathcal{A}_n,\mathcal{Z}_n\big\} - E\big\{\psi(W_i;\theta_a,\eta_a)\big|\mathcal{A}_n,\mathcal{Z}_n\big\} = 0\\
f'_{j,i}(0) &= \frac{d}{dr}E\big\{\psi(W_i;\theta_a,\eta_a + r(\hat{\eta}_{a,j}-\eta_a))\big|(W_i)_{i\in\mathcal{I}_j^c},\mathcal{A}_n,\mathcal{Z}_n\big\}\Big|_{r=0}\\
f''_{j,i}(\tilde{r}) &= \frac{d^2}{dr^2}E\big\{\psi(W_i;\theta_a,\eta_a + r(\hat{\eta}_{a,j}-\eta_a))\big|(W_i)_{i\in\mathcal{I}_j^c},\mathcal{A}_n,\mathcal{Z}_n\big\}\Big|_{r=\tilde{r}}.
\end{aligned}
$$

We have that

$$
\begin{aligned}
\psi(W_i;\theta_a,\eta_a + r(\hat{\eta}_{a,j}-\eta_a)) = \frac{I_{a,i}}{\pi_a + r(\hat{\pi}_{a,j}-\pi_a)}&\{Y_{a,i} - \mu_a^*(X_i) - r(\hat{\mu}_{a,j}^*(X_i) - \mu_a^*(X_i))\}\\
&+ \mu_a^*(X_i) + r(\hat{\mu}_{a,j}^*(X_i) - \mu_a^*(X_i)) - \theta_a.
\end{aligned}
$$

Taking the expectation, then two derivatives with respect to r, we have

$$
\begin{aligned}
\frac{d}{dr}E^*\big\{\psi(W_i;\theta_a,\eta_a + r(\hat{\eta}_{a,j}-\eta_a))\big\} &= -\frac{I_{a,i}E^*(\hat{\mu}_{a,j}^*(X_i) - \mu_a^*(X_i))}{\pi_a + r(\hat{\pi}_{a,j}-\pi_a)}\\
&\quad - \frac{I_{a,i}E^*[Y_{a,i} - \mu_a^*(X_i) - r(\hat{\mu}_{a,j}^*(X_i) - \mu_a^*(X_i))][\hat{\pi}_{a,j}-\pi_a]}{[\pi_a + r(\hat{\pi}_{a,j}-\pi_a)]^2}\\
&\quad + E^*(\hat{\mu}_{a,j}^*(X_i) - \mu_a^*(X_i))\\
\frac{d^2}{dr^2}E^*\big\{\psi(W_i;\theta_a,\eta_a + r(\hat{\eta}_{a,j}-\eta_a))\big\} &= 2\frac{(\hat{\pi}_{a,j}-\pi_a)I_{a,i}E^*(\hat{\mu}_{a,j}^*(X_i) - \mu_a^*(X_i))}{[\pi_a + r(\hat{\pi}-\pi_a)]^2}\\
&\quad + \frac{I_{a,i}E^*[Y_{a,i} - \mu_a^*(X_i) - r(\hat{\mu}_{a,j}^*(X_i) - \mu_a^*(X_i))][\hat{\pi}_{a,j}-\pi_a]^2}{[\pi_a + r(\hat{\pi}_{a,j}-\pi_a)]^3}.
\end{aligned}
$$

Evaluating f' at $r = 0$, we have

$$
f'_{j,i}(0) = E^*(\hat{\mu}_{a,j}^*(X_i) - \mu_a^*(X_i))\left(1 - \frac{I_{a,i}}{\pi_a}\right) - \frac{I_{a,i}}{\pi_a^2}[\hat{\pi}_{a,j}-\pi_a]E(Y_{a,i} - \mu_a^*(X_i)|Z_i).
$$

We know that $\frac{1}{\sqrt{n_j}}\sum_{i\in\mathcal{I}_j} f_{j,i}(r) = o_P(1)$, using both Assumption 2.1 and 2*. Then,

$$\begin{aligned}\frac{1}{\sqrt{n_j}}\sum_{i\in\mathcal{I}_j} f'_{j,i}(0) &= \pi_a\sqrt{n_j}\sum_{l=1}^{L}\frac{1}{n_k(l)}\hat{p}_{j,l}\left(\sup_{\mu^*\in\mathcal{F}^*_{a,n}} E(\mu^*(X_i)-\mu^*_a(X_i)|Z_i=l)\right)\sum_{i\in I_{k,l}}(\pi_a - I_{a,i}) \\ &\quad + \frac{1}{\pi_a^2}[\hat{\pi}_{a,j}-\pi_a]\frac{1}{\sqrt{n_j}}\sum_{i\in\mathcal{I}_j} I_{a,i}E(Y_{a,i}-\mu^*_a(X_i)|Z_i) \\ &= \pi_a\sum_{l=1}^{L}\sqrt{\hat{p}_{j,l}}\left(\sup_{\mu^*\in\mathcal{F}^*_{a,n}} E(\mu^*(X_i)-\mu^*_a(X_i)|Z_i=l)\right)\frac{1}{\sqrt{n_j(l)}}\sum_{i\in\mathcal{I}_{j,l}}(\pi_a - I_{a,i}) \\ &\quad + o_P(1)\end{aligned}$$

where $\hat{p}_{j,l} = n_j(l)/n_j$, i.e., the fraction within strata l out of the sample j, and $\mathcal{I}_{j,l}$ are the indexes within $\mathcal{I}_j$ such that they are also in strata l. Within each strata l, by Assumption 2.1 (see equation (B.3)), we have that $\frac{1}{\sqrt{n_{k,l}}}\sum_{i\in I_{k,l}}(\pi_a - I_{a,i}) = O_P(1)$. Furthermore, by Assumption 2.4, $\left(\sup_{\mu^*\in\mathcal{F}^*_{a,n}} E(\mu^*(X_i)-\mu^*_a(X_i)|Z_i=l)\right) = o_P(1)$. Therefore, since we have a finite number of strata, $\frac{1}{\sqrt{n_j}}\sum_{i\in\mathcal{I}_j} f_{j,i}(0) = o_P(1)O_P(1) = o_P(1)$. For the second derivative, we have (again applying the prediction unbiasedness, and recalling that $\tilde{r}\leq 1$):

$$\begin{aligned}\left|\frac{1}{\sqrt{n}_j}\sum_{i\in\mathcal{I}_j} f''_{j,i}(\tilde{r})\right| &\leq 2\left|\frac{(\hat{\pi}_{a,j}-\pi_a)\frac{1}{\sqrt{n_j}}\sum_{i\in\mathcal{I}_j}\sup_{\mu^*\in\mathcal{F}^*_{a,n}} E(\mu^*(X_i)-\mu^*_a(X_i)|Z_i)}{[\pi_a+\tilde{r}(\hat{\pi}-\pi_a)]^2}\right| \\ &\quad + \left|[\hat{\pi}_{a,j}-\pi_a]^2\frac{\frac{1}{\sqrt{n_j}}\sum_{i\in\mathcal{I}_j}\sup_{\mu^*\in\mathcal{F}^*_{a,n}} E(\mu^*(X_i)-\mu^*_a(X_i)|Z_i)}{[\pi_a+\tilde{r}(\hat{\pi}_{a,j}-\pi_a)]^3}\right| \\ &\quad + \left|[\hat{\pi}_{a,j}-\pi_a]^2\frac{\frac{1}{\sqrt{n_j}}\sum_{i\in\mathcal{I}_j} I_{a,i}E(Y_{a,i}-\mu^*_a(X_i)|Z_i)}{[\pi_a+\tilde{r}(\hat{\pi}_{a,j}-\pi_a)]^3}\right|.\end{aligned}$$

To show that $\frac{1}{\sqrt{n_j}}\sum_{i\in\mathcal{I}_j} f''_{j,i}(\tilde{r}) = o_P(1)$, we first show that $\pi_a+\tilde{r}(\hat{\pi}_{a,j}-\pi_a)$ is bounded below by a constant. Notice that $\pi_a+\tilde{r}(\hat{\pi}_{a,j}-\pi_a) = (1-\tilde{r})\pi_a + (\tilde{r})\hat{\pi}_{a,j}$ is a weighed average of π_a and $\hat{\pi}_{a,j}$. Since both π_a and $\hat{\pi}_{a,j}$ must be larger than ϵ, their average must also be larger

than ϵ. Therefore, recalling that $\hat{\pi}_{a,j} - \pi_a = O_P(1/\sqrt{n})$ we have

$$\begin{aligned}
\left|\frac{1}{\sqrt{n}_j}\sum_{i\in\mathcal{I}_j} f''_{j,i}(\tilde{r})\right| &\leq 2O_P(1/\sqrt{n})\epsilon^2\left|\frac{1}{\sqrt{n}_j}\sum_{i\in\mathcal{I}_j}\sup_{\mu^*\in\mathcal{F}^*_{a,n}} E(\mu^*(X_i)-\mu^*_a(X_i)|Z_i)\right| \\
&\quad + O_P(1/n)\epsilon^3\left|\frac{1}{\sqrt{n}_j}\sum_{i\in\mathcal{I}_j}\sup_{\mu^*\in\mathcal{F}^*_{a,n}} E(\mu^*(X_i)-\mu^*_a(X_i)|Z_i)\right| \\
&\quad + O_P(1/n)\epsilon^3\left|\frac{1}{\sqrt{n}_j}\sum_{i\in\mathcal{I}_j} I_{a,i}E(Y_{a,i}-\mu^*_a(X_i)|Z_i)\right| \\
&\leq 2O_P(1/\sqrt{n})\epsilon^2\left|\frac{1}{\sqrt{n}_j}\sum_{i\in\mathcal{I}_j}\sup_{\mu^*\in\mathcal{F}^*_{a,n}} E(\mu^*(X_i)-\mu^*_a(X_i)|Z_i)\right| + O_P(1/n)\epsilon^3 O_P(1/\sqrt{n_j})
\end{aligned}$$

where the last term comes from either (B) or (U) holding, as before. Furthermore, recalling Assumption 2.4 as we did for f', we have:

$$\begin{aligned}
\frac{1}{\sqrt{n}_j}\sum_{i\in\mathcal{I}_j}\sup_{\mu^*\in\mathcal{F}^*_{a,n}} E(\mu^*(X_i)-\mu^*_a(X_i)|Z_i) &= \frac{1}{\sqrt{n}_j}\sum_{l=1}^{L}\sum_{i\in\mathcal{I}_{j,l}}\sup_{\mu^*\in\mathcal{F}^*_{a,n}} E(\mu^*(X_i)-\mu^*_a(X_i)|Z_i) \\
&= \sum_{l=1}^{L}\sqrt{\hat{p}_{j,l}}\sqrt{n_j(l)}\sup_{\mu^*\in\mathcal{F}^*_{a,n}} E(\mu^*(X_i)-\mu^*_a(X_i)|Z_i) \\
&\leq \sum_{l=1}^{L}\sqrt{n_j(l)}\sup_{\mu^*\in\mathcal{F}^*_{a,n}} E(\mu^*(X_i)-\mu^*_a(X_i)|Z_i) \\
&= \sum_{l=1}^{L}\sqrt{n_j(l)}\cdot o_P(1) \\
&= o_P\left(\sqrt{n_j(l)}\right).
\end{aligned}$$

Therefore, $\left|\frac{1}{\sqrt{n}_j}\sum_{i\in\mathcal{I}_j} f''_{k,i}(\tilde{r})\right| = O_P(1/\sqrt{n})o_P(\sqrt{n_j(l)}) + o_P(1) = o_P(1)$. □

B.3 Asymptotic Normality (AN) of AIPW Estimators under CARs with Condition (B)

Condition (B) is a condition on the limiting distribution of the within-strata treatment assignment imbalances. It is a critical ingredient for deriving asymptotic normality. In Section B.3.1, we break up the influence function that we derived in Theorem 2.1 (i). The asymptotic behavior of these pieces can be studied separately. Then, using tools in Section B.3.3, we can add the distributions together to show asymptotic normality in Section B.3.2. In this section, recall the definitions of P_n, P_0 and P_0^n from Section B.2.1.

B.3.1 Asymptotic Variance Decomposition

To show asymptotic normality of the AIPW estimator that has the influence function in Theorem (i) and compute its asymptotic variance under condition (B). First, we rewrite ψ_a as the sum of five components. Recall that our expression for ψ_a was

$$\psi_a(A_i, Y_i, X_i) = \frac{I_{a,i}}{\pi_a}\{Y_{a,i} - \tilde{\mu}_a(X_i)\} + \tilde{\mu}_a(X_i) - \theta_a.$$

where $\tilde{\mu}_a(X_i) = \mu_a(X_i) + \theta_a - P_0\mu_a(X_i)$. We can split up this influence function into five parts that will make it easier to derive the asymptotic distribution of $\hat{\theta}_{\text{AIPW}}$. Let

$$\begin{aligned}
\psi_a^1 &: (A_i, X_i, Y_i, Z_i) \to \frac{I_{a,i}}{\pi_a}\left[Y_{a,i} - \tilde{\mu}_a(X_i) - E\left\{Y_{a,i} - \tilde{\mu}_a(X_i)\middle| Z_i\right\}\right] \\
\psi_a^2 &: (A_i, X_i, Y_i, Z_i) \to \frac{(I_{a,i} - \pi_a)}{\pi_a} E\left\{Y_{a,i} - \tilde{\mu}_a(X_i)\middle| Z_i\right\} \\
\psi_a^3 &: (X_i, Y_i, Z_i) \to E\left\{Y_{a,i} - \tilde{\mu}_a(X_i)\middle| Z_i\right\} \\
\psi_a^4 &: (X_i, Z_i) \to \tilde{\mu}_a(X_i) - E\left\{\tilde{\mu}_a(X_i)|Z_i\right\} \\
\psi_a^5 &: (X_i, Z_i) \to E\left\{\tilde{\mu}_a(X_i)|Z_i\right\} - \theta_a.
\end{aligned}$$

Then by construction, $\psi_a = \psi_a^1 + \cdots + \psi_a^5$. Let $\psi^j = (\psi_1^j, ..., \psi_K^j)^T$ be the vector of the jth component of ψ_a across all treatment groups $a = 1, \ldots, K$. Similarly, let $Y = (Y_1, ..., Y_K)^T$,

and $\tilde{\mu}(X) = (\tilde{\mu}_1(X), ..., \tilde{\mu}_K(X))^T$. Conditional on $(\mathcal{A}_n, \mathcal{Z}_n)$, we have that $P_0^n(\phi^1 + \phi^4) = 0$. Using the Lindeberg-Feller central limit theorem (see Lemma B.18),

$$\sqrt{n}P_0^n(\phi^1 + \phi^4) \xrightarrow{d} N\left(0, V^{(1,4)}\right)$$

where $V^{(1,4)}$ is a matrix such that $\text{Var}(\phi^1 + \phi^4|\mathcal{A}_n, \mathcal{Z}_n) \to V^{(1,4)}$ in probability. The variance of two random vectors (since ϕ^1 and ϕ^4 are conditionally mean zero), is given by

$$\begin{aligned}
\text{Var}(\phi^1 + \phi^4|\mathcal{A}_n, \mathcal{Z}_n) &= E\left\{(\phi^1 + \phi^4)(\phi^1 + \phi^4)^T\middle|\mathcal{A}_n, \mathcal{Z}_n\right\} \\
&= E\left\{\phi^1(\phi^1)^T\middle|\mathcal{A}_n, \mathcal{Z}_n\right\} + E\left\{\phi^1(\phi^4)^T\middle|\mathcal{A}_n, \mathcal{Z}_n\right\} \\
&\quad + E\left\{\phi^4(\phi^1)^T\middle|\mathcal{A}_n, \mathcal{Z}_n\right\} + E\left\{\phi^4(\phi^4)^T\middle|\mathcal{A}_n, \mathcal{Z}_n\right\} \\
&= \text{Var}(\phi^1|\mathcal{A}_n, \mathcal{Z}_n) + \text{Var}(\phi^4|\mathcal{A}_n, \mathcal{Z}_n) + \\
&\quad \text{Cov}(\psi_1, \psi_4|\mathcal{A}_n, \mathcal{Z}_n) + \text{Cov}(\psi_4, \psi_1|\mathcal{A}_n, \mathcal{Z}_n).
\end{aligned}$$

Since conditional on $(\mathcal{A}_n, \mathcal{Z}_n)$, the product of $I_{j,i}$ and $I_{k,i}$ is zero for $j \neq k$, then $\text{Var}(\phi^1|\mathcal{A}_n, \mathcal{Z}_n)$ is a diagonal matrix with (a, a)th entry given by

$$\pi_a^{-2} I_{a,i} \text{Var}(Y_{a,i} - \tilde{\mu}_a(X_i)|Z_i) \to \pi_a^{-1} E(\text{Var}(Y_a - \tilde{\mu}_a(X)|Z)).$$

Similarly, $\text{Var}(\phi^4|\mathcal{A}_n, \mathcal{Z}_n) \to E(\text{Var}(\tilde{\mu}(X)|Z))$, $\text{Cov}(\phi^1, \phi^4|\mathcal{A}_n, \mathcal{Z}_n) \to E(\text{Cov}(Y - \tilde{\mu}(X), \tilde{\mu}(X)|Z))$. Therefore,

$$\begin{aligned}
V^{(1,4)} &= \text{diag}\{\pi_a^{-1} E(\text{Var}(Y_a - \tilde{\mu}_a(X)|Z))\} + E(\text{Var}(\tilde{\mu}(X)|Z)) \\
&\quad + E(\text{Cov}(Y - \tilde{\mu}(X), \tilde{\mu}(X)|Z)) + E(\text{Cov}(\tilde{\mu}(X), Y - \tilde{\mu}(X)|Z)).
\end{aligned}$$

Furthermore, we have that, since Z_i are i.i.d. random variables, we have

$$\sqrt{n}P_n(\phi^3 + \phi^5) \xrightarrow{d} N\left(0, V^{(3,5)}\right)$$

where, similarly to the steps for $V^{(1,4)}$ (though there is no need to condition on $(\mathcal{A}_n, \mathcal{Z}_n)$),

$$\begin{aligned} V^{(3,5)} &= \mathrm{Var}(\phi^3) + \mathrm{Var}(\phi^5) + \mathrm{Cov}(\phi^3, \phi^5) + \mathrm{Cov}(\phi^5, \phi^3) \\ &= E(E(Y - \tilde{\mu}|Z)E(Y - \tilde{\mu}|Z)^T) + \mathrm{Var}(E(\tilde{\mu}|Z)) \\ &\quad + \mathrm{Cov}(E(Y - \tilde{\mu}|Z), E(\tilde{\mu}|Z)) + \mathrm{Cov}(E(\tilde{\mu}|Z), E(Y - \tilde{\mu}|Z)) \\ &= E(R(Z)\pi\pi^T R(Z)) + \mathrm{Var}(E(\tilde{\mu}|Z)) \\ &\quad + \mathrm{Cov}(E(Y - \tilde{\mu}|Z), E(\tilde{\mu}|Z)) + \mathrm{Cov}(E(\tilde{\mu}|Z), E(Y - \tilde{\mu}|Z)) \end{aligned}$$

where $R(Z) = \mathrm{diag}\{\pi_a^{-1} E(Y_a - \tilde{\mu}_a(X)|Z); a = 1, \ldots, K\}$.

B.3.2 Proof of Theorem 2.1 (ii)

We now derive asymptotic normality and variance under condition (B). This completes the proof of Theorem 2.1 (ii), and also applies to the cross-fitted estimator since we have shown they have the same influence function.

Under (B), we have that

$$\sqrt{n}\left(\tfrac{n_a(z_l)}{n(z_l)} - \pi_a,\ {\scriptstyle a=1,\ldots,K \atop \scriptstyle l=1,\ldots,L}\right)^T \Big| \mathcal{Z}_n \xrightarrow{d} N\left(0, \mathrm{diag}\left\{\frac{\Omega(z_l)}{P(Z = z_l)},\ l = 1, ..., L\right\}\right). \tag{B.8}$$

We first note that, with $\hat{p}_{z_l} = n(z_l)/n$,

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} I_{a,i}\pi_a^{-1} E(Y_{a,i} - \tilde{\mu}_a(X_i)|Z_i) &= \sum_{l=1}^{L} \frac{n_a(z_l)}{n(z_l)} \hat{p}_{z_l} R_a(z_l) \\ \frac{1}{n}\sum_{i=1}^{n} E(Y_{a,i} - \tilde{\mu}_a(X_i)|Z_i) &= \sum_{l=1}^{L} \pi_a \hat{p}_{z_l} R_a(z_l). \end{aligned}$$

where $R_a(Z) = \pi_a^{-1} E(Y_{a,i} - \tilde{\mu}_a(X_i)|Z)$. Therefore, we can find the distribution of $\psi_2|\mathcal{Z}_n$ by performing a linear transformation on the normal distribution in (B.8). Define $R(Z) =$

$\text{diag}\{R_1(Z), ..., R_k(Z)\}$. Then, letting $\mathbb{L}_n := \left(\frac{n_a(z_l)}{n(z_l)} - \pi_a, \ {}^{a=1,...,K}_{l=1,...,L}\right)^T$, we have that

$$P_n(\psi_2) = [\hat{p}_{z_1} R(z_1), ..., \hat{p}_{z_L} R(z_L)]\, \mathbb{L}_n.$$

Remark B.3. *We need (B) to show asymptotic normality and derive its limiting variance, but we do not need (B) to show that $P_n(\psi_2)|\mathcal{Z}_n = O_P(1/\sqrt{n})$. This is apparent only from Assumption 2.1 (iii), which states that each element of the vector $\sqrt{n}\mathcal{L}_n = O_P(1)$, so the linear transformation of $\mathbb{L}_n$ above is $O_P(1/\sqrt{n})$.*

Therefore, $\sqrt{n}P_n(\psi_2)|\mathcal{Z}_n \xrightarrow{d} N(0, V^{(2)})$ where $V^{(2)}$ is the limit in probability of the variance-covariance matrix of the linear transformation conditioning on $\mathcal{Z}_n$:

$$\begin{aligned}
V^{(2)}_{\mathcal{Z}_n} &= [\hat{p}_{z_1} R(z_1), ..., \hat{p}_{z_L} R(z_L)]\, \text{diag}\left\{p_{z_l}^{-1}\Omega(z_l); l = 1, ...L\right\} [\hat{p}_{z_1} R(z_1), ..., \hat{p}_{z_L} R(z_L)]^T \\
&= \sum_{l=1}^{L} (\hat{p}_{z_l})^2 p_{z_l}^{-1} R(z_l)\Omega(z_l)R(z_l) \\
&= \sum_{l=1}^{L} \hat{p}_{z_l} R(z_l)\Omega(z_l)R(z_l) + o_P(1) \xrightarrow{p} E(R(Z)\Omega(Z)R(Z)) \equiv V^{(2)}.
\end{aligned}$$

Define the following scaled random variables that are the empirical averages of the influence function components: $M_n^{(1,4)} := \left[V^{(1,4)}\right]^{-1} \sqrt{n}P_n(\phi^1 + \phi^4)$, $M_n^{(2)} := \left[V^{(2)}\right]^{-1} \sqrt{n}P_n(\phi^2)$, and $M_n^{(3,5)} := \left[V^{(3,5)}\right]^{-1} \sqrt{n}P_n(\phi^3 + \phi^5)$. Let I_K be the identity matrix with dimension K. Then by the continuous mapping theorem, $M_n^{(1,4)}|\mathcal{A}_n, \mathcal{Z}_n \xrightarrow{d} N(0, I_K)$, $M_n^{(2)}|\mathcal{Z}_n \xrightarrow{d} N(0, I_K)$, and $M_n^{(3,5)} \xrightarrow{d} N(0, I_K)$. Furthermore, these three normal distributions are asymptotically independent. Applying Lemma B.17, $(M_n^{(1,4)}, M_n^{(2)}, M_n^{(3,5)}) \xrightarrow{d} (M^{(1,4)}, M^{(2)}, M^{(5)})$, where $M^{(1,4)}$, $M^{(2)}$, and $M^{(3,5)}$ are independent multivariate normal random variables. Therefore, by the continuous mapping theorem, we can scale each of them by their respective variance-covariance matrices, and add the three random variables together to get that under (B),

$$\sqrt{n}(\hat{\theta}_{AIPW} - \theta) \equiv \sqrt{n}P_n(\phi^1 + \phi^2 + \phi^3 + \phi^4 + \phi^5) + o_P(1) \xrightarrow{d} N(0, \Sigma_{(B)})$$

where $\Sigma_{(B)}$ is given by

$$\begin{aligned}
\Sigma_{(B)} &= V^{(1,4)} + V^{(2)} + V^{(3,5)} \\
&= \text{diag}\{\pi_a^{-1} E(\text{Var}(Y_a - \tilde{\mu}_a(X)|Z))\} + E(\text{Var}(\tilde{\mu}(X)|Z)) \\
&\quad + E(\text{Cov}(Y - \tilde{\mu}(X), \tilde{\mu}(X)|Z)) + E(\text{Cov}(\tilde{\mu}(X), Y - \tilde{\mu}(X)|Z)) \\
&\quad + E(R(Z)\Omega(Z)R(Z)) + E(R(Z)\pi\pi^T R(Z)) + \text{Var}(E(\tilde{\mu}|Z)) \\
&\quad + \text{Cov}(E(Y - \tilde{\mu}(X)|Z), E(\tilde{\mu}(X)|Z)) + \text{Cov}(E(\tilde{\mu}(X)|Z), E(Y - \tilde{\mu}(X)|Z)) \\
&= \text{diag}\{\pi_a^{-1} E(\text{Var}(Y_a - \tilde{\mu}_a(X)|Z))\} + E(R(Z)\Omega(Z)R(Z)) + E(R(Z)\pi\pi^T R(Z)) \\
&\quad + \text{Var}(\tilde{\mu}(X)) + \text{Cov}(Y - \tilde{\mu}(X), \tilde{\mu}(X)) + \text{Cov}(\tilde{\mu}(X), Y - \tilde{\mu}(X)) \\
&= \underbrace{\text{diag}\{\pi_a^{-1} E(\text{Var}(Y_a - \tilde{\mu}_a(X)|Z))\}}_{(I)} + E(R(Z)\Omega(Z)R(Z)) + E(R(Z)\pi\pi^T R(Z)) \\
&\quad - \text{Var}(\tilde{\mu}(X)) + \text{Cov}(Y, \tilde{\mu}(X)) + \text{Cov}(\tilde{\mu}(X), Y)
\end{aligned}$$

where we have simplified the expressions using the law of total variance and law of total covariance. Under simple randomization $\Omega_{SR} = \text{diag}\{\pi_a; a = 1, \ldots, K\} - \pi\pi^T$. Then,

$$\begin{aligned}
(I) &= \text{diag}\{\pi_a^{-1}\text{Var}(Y_a - \tilde{\mu}_a(X))\} - \text{diag}\{\pi_a^{-1}\text{Var}(E(Y_a - \tilde{\mu}_a(X)|Z))\} \\
&= \text{diag}\{\pi_a^{-1}\text{Var}(Y_a - \tilde{\mu}_a(X))\} - \text{diag}\{\pi_a^{-1}\text{Var}(R_a(Z))\} \\
&= \text{diag}\{\pi_a^{-1}\text{Var}(Y_a - \tilde{\mu}_a(X))\} - E(R(Z)\text{diag}\{\pi_a\}R(Z))
\end{aligned}$$

where the last equality is because $E(R(Z)) = 0$. Putting these together, we have

$$\begin{aligned}
\Sigma_{(B)} &= \text{diag}\{\pi_a^{-1}\text{Var}(Y_a - \tilde{\mu}_a(X))\} \\
&\quad - E(R(Z)\text{diag}\{\pi_a\}R(Z)) + E(R(Z)\Omega(Z)R(Z)) + E(R(Z)\pi\pi^T R(Z)) \\
&\quad - \text{Var}(\tilde{\mu}(X)) + \text{Cov}(Y, \tilde{\mu}(X)) + \text{Cov}(\tilde{\mu}(X), Y) \\
&= \text{diag}\{\pi_a^{-1}\text{Var}(Y_a - \tilde{\mu}_a(X))\} - E(R(Z)\{\Omega_{SR} - \Omega\}R(Z)) \\
&\quad - \text{Var}(\tilde{\mu}(X)) + \text{Cov}(Y, \tilde{\mu}(X)) + \text{Cov}(\tilde{\mu}(X), Y).
\end{aligned}$$

Note that $\tilde{\mu}_a$ can be replaced with μ_a in the above expression because $\text{Var}(P_0\mu_a(X) - \theta_a)$ is a constant. When $\mu_a(X)$ is linear in x, the variance agrees with Ye et al. (2023)'s Theorem 3.

B.3.3 Tools for Proof of Theorem 2.1 (ii)

We will use the following lemma that uses asymptotic independence of the various components of the influence function that we have defined in order to claim asymptotic normality.

Lemma B.17 (Asymptotic Independence)**.** *Let Q_n, R_n, and S_n be random vectors of the same dimension such that $Q_n|\mathcal{A}_n, \mathcal{Z}_n \xrightarrow{d} N(0, I)$, $R_n|\mathcal{Z}_n \xrightarrow{d} N(0, I)$, and $S_n \xrightarrow{d} N(0, I)$ where I is the identity matrix with the correct dimension. Then jointly, $(Q_n, R_n, S_n) \xrightarrow{d} (Q, R, S)$, where Q, R, and S are independent, standard multivariate normal random variables. (Also see Ye et al. (2023) Proof of Theorem 3).*

Proof. Applying iterated expectation, and the dominated convergence theorem to switch limit and expectation, we have

$$
\begin{aligned}
&\lim_{n\to\infty} P(Q_n \le Q, R_n \le R, S_n \le S) \\
&= \lim_{n\to\infty} E\left\{E\left\{E\left\{1(Q_n \le Q)1(R_n \le R)1(S_n \le S)\middle|\mathcal{A}_n, \mathcal{Z}_n\right\}|\mathcal{Z}_n\right\}\right\} \\
&= \lim_{n\to\infty} E\left\{E\left\{E\left\{1(Q_n \le Q)\middle|\mathcal{A}_n, \mathcal{Z}_n\right\}1(R_n \le R)\middle|\mathcal{Z}_n\right\}1(S_n \le S)\right\} \\
&= E\left\{E\left\{\lim_{n\to\infty} E\left\{1(Q_n \le Q)\middle|\mathcal{A}_n, \mathcal{Z}_n\right\}\lim_{n\to\infty} 1(R_n \le R)\middle|\mathcal{Z}_n\right\}\lim_{n\to\infty} 1(S_n \le S)\right\} \\
&= \Phi(Q)E\left\{E\left\{\lim_{n\to\infty} 1(R_n \le R)\middle|\mathcal{Z}_n\right\}\lim_{n\to\infty} 1(S_n \le s)\right\} \\
&= \Phi(Q)E\left\{\lim_{n\to\infty} E\left\{1(R_n \le R)\middle|\mathcal{Z}_n\right\}\lim_{n\to\infty} 1(S_n \le s)\right\} \\
&= \Phi(Q)\Phi(R)E\left\{\lim_{n\to\infty} 1(S_n \le s)\right\} \\
&= \Phi(Q)\Phi(R)\Phi(S) \\
&= P(Q \le Q, R \le R, S \le S)
\end{aligned}
$$

Therefore $(Q_n, R_n, S_n) \xrightarrow{d} (Q, R, S)$, where Q, R, and S are independent, standard multivariate normal random variables. □

Lemma B.18 (Verification of Lindeberg-Feller Condition)**.** *Define ψ_1 and ψ_4 as in Section B.3.1, that is $\psi_1 = (\psi_1^1, ..., \psi_k^1)$, and $\psi_4 = (\psi_1^4, , ..., \psi_k^4)$ with*

$$\psi_a^1 : (A_i, X_i, Y_i, Z_i) \to \frac{I_{a,i}}{\pi_a}\left[Y_{a,i} - \tilde{\mu}_a(X_i) - E\left\{Y_{a,i} - \tilde{\mu}_a(X_i)\middle| Z_i\right\}\right]$$
$$\psi_a^4 : (X_i, Z_i) \to \tilde{\mu}_a(X_i) - E\left\{\tilde{\mu}_a(X_i)|Z_i\right\}.$$

Then $\sqrt{n}P_n(\psi^1 + \psi^4)\big|\mathcal{A}_n, \mathcal{Z}_n \xrightarrow{d} N\left(0, V^{(1,4)}\right)$ where $V^{(1,4)}$ is a matrix such that $\mathrm{Var}(\psi^1 + \psi^4|\mathcal{A}_n, \mathcal{Z}_n) \to V^{(1,4)}$ in probability.

Proof. We aim to verify the Lindeberg condition for any linear combination of the $2K$ vector $P_n(\psi_1^1, ..., \psi_k^1, \psi_1^4, ..., \psi_k^4)c$ with $c \in \mathbb{R}^{2K}$, conditional on $(\mathcal{A}_n, \mathcal{Z}_n)$. By the Cramer-Wold device, and the Continuous Mapping Theorem, this will complete the proof. For simplicity of notation, let $W_i = (A_i, X_i, y_{1,i}, ..., y_{k,i}, Z_i)$. Let

$$K_i = \frac{1}{n}\left(\psi_1^1(W_i), ..., \psi_k^1(W_i), \psi_1^4(W_i), ..., \psi_k^4(W_i)\right)c.$$

We are interested in $\sum_{i=1}^n K_i$. We know that $E(K_i|\mathcal{A}_n, \mathcal{Z}_n) = 0$, since each component $\psi_a^1(W_i)$ and $\psi_a^4(W_i)$ have conditional expectation zero. Furthermore,

$$\mathrm{Var}(K_i|\mathcal{A}_n, \mathcal{Z}_n) = \frac{1}{n^2}c^T\mathrm{Var}\left(\psi_1^1(W_i), ..., \psi_k^1(W_i), \psi_1^4(W_i), ..., \psi_k^4(W_i)|\mathcal{A}_n, \mathcal{Z}_n\right)c.$$

Lindeberg's condition for the random variables $K_i|(\mathcal{A}_n, \mathcal{Z}_n)$ holds if, for all $\epsilon > 0$, (following exactly the steps in Ye et al. (2023) proof of their Theorem 2)

$$\sum_{i=1}^n E\left[\frac{K_i^2}{\mathrm{Var}(\sum_{i=1}^n K_i|\mathcal{A}_n, \mathcal{Z}_n)} I\left(\frac{K_i^2}{\mathrm{Var}(\sum_{i=1}^n K_i|\mathcal{A}_n, \mathcal{Z}_n)} > \epsilon\right)\middle|\mathcal{A}_n, \mathcal{Z}_n\right] \to 0 \tag{B.9}$$

as $n \to \infty$. Let $\tau_{i,n} := \frac{\mathrm{Var}(K_i|\mathcal{A}_n, \mathcal{Z}_n)}{\mathrm{Var}(\sum_{i=1}^n K_i|\mathcal{A}_n, \mathcal{Z}_n)}$. By (conditional) independence, $\sum_{i=1}^n \tau_{i,n} = 1$, since

the sum of the variances is the variance of the sums. Then we can rewrite (B.9) as:

$$\begin{aligned}
(\text{B.9}) &= \sum_{i=1}^{n}(\tau_{i,n})E\left[\frac{K_i^2}{\text{Var}(K_i|\mathcal{A}_n,\mathcal{Z}_n)}I\left(\frac{K_i^2}{\text{Var}(\sum_{i=1}^{n}K_i|\mathcal{A}_n,\mathcal{Z}_n)}>\epsilon\right)\middle|\mathcal{A}_n,\mathcal{Z}_n\right] \\
&\leq \max_i E\left[\frac{K_i^2}{\text{Var}(K_i|\mathcal{A}_n,\mathcal{Z}_n)}I\left(\frac{K_i^2}{\text{Var}(\sum_{i=1}^{n}K_i|\mathcal{A}_n,\mathcal{Z}_n)}>\epsilon\right)\middle|\mathcal{A}_n,\mathcal{Z}_n\right] \\
&= \max_i E\left[\frac{K_i^2}{\text{Var}(K_i|\mathcal{A}_n,\mathcal{Z}_n)}I\left(\frac{K_i^2}{\text{Var}(K_i|\mathcal{A}_n,\mathcal{Z}_n)}>\frac{\epsilon}{\tau_{i,n}}\right)\middle|\mathcal{A}_n,\mathcal{Z}_n\right]
\end{aligned}$$

Note that $W_i := \frac{K_i}{\sqrt{\text{Var}(K_i|\mathcal{A}_n,\mathcal{Z}_n)}}$ is a random variable which, conditional on $\mathcal{A}_n, \mathcal{Z}_n$, has mean zero and has unit variance. It is important to recall that we have finite strata levels $\{1, ..., L\}$. Therefore, quantities conditional on strata levels and treatment indicators have maximums that are achieved. For any i, we have that $E(W_i^2 I(W_i^2 > \epsilon \max_i(\tau_{i,n}^{-1}))|\mathcal{A}_n, \mathcal{Z}_n) \to 0$ by the dominated convergence theorem. This is because point-wise, the function $w^2 I(w^2 > \epsilon \max_i(\tau_{i,n}^{-1})) \to 0$ as $n \to \infty$. Also, $w^2 I(w^2 > \epsilon \max_i(\tau_{i,n}^{-1})) \leq w^2$, and $E(w^2|\mathcal{A}_n, \mathcal{Z}_n) = 1 < \infty$ is bounded, so we can switch limit and expectation. Therefore, $\lim_{n\to\infty} E(W_i^2 I(W_i^2 > \epsilon \max_i(\tau_{i,n}^{-1}))|\mathcal{A}_n, \mathcal{Z}_n) = E(\lim_{n\to\infty} W_i^2 I(W_i^2 > \epsilon \max_i(\tau_{i,n}^{-1}))|\mathcal{A}_n, \mathcal{Z}_n) = 0$.

The covariance elements of $\text{Var}\left(\psi_1^1(W_i), ..., \psi_k^1(W_i), \psi_1^4(W_i), ..., \psi_k^4(W_i)|\mathcal{A}_n, \mathcal{Z}_n\right)$ are given by:

$$\begin{aligned}
\text{Cov}(\psi_a^1(W_i), \psi_a^1(W_i)|\mathcal{A}_n, \mathcal{Z}_n) &= \pi_a^{-2} I_{a,i} \text{Var}(Y_{a,i} - \tilde{\mu}_a(X_i)|Z_i) \\
\text{Cov}(\psi_a^1(W_i), \psi_b^1(W_i)|\mathcal{A}_n, \mathcal{Z}_n) &= 0 \\
\text{Cov}(\psi_a^4(W_i), \psi_b^4(W_i)|\mathcal{A}_n, \mathcal{Z}_n) &= \text{Cov}(\tilde{\mu}_a(X_i), \tilde{\mu}_b(X_i)|Z_i) \\
\text{Cov}(\psi_a^1(W_i), \psi_b^4(W_i)|\mathcal{A}_n, \mathcal{Z}_n) &= \frac{1}{\pi_a} I_{a,i} \text{Cov}(Y_{a,i} - \tilde{\mu}_a(X_i), \tilde{\mu}_b(X_i)|Z_i)
\end{aligned}$$

The elements are all functions of (A_i, Z_i) which have finite levels. Therefore, the quadratic form of the covariance matrix, $c^T \text{Var}\left(\psi_1^1(W_i), ..., \psi_k^1(W_i), \psi_1^4(W_i), ..., \psi_k^4(W_i)|\mathcal{A}_n, \mathcal{Z}_n\right) c \leq C_{Z_i,A_i}$, which is positive semi-definite by nature of being a covariance matrix, can be bounded above and below by positive constants. Thus, $\max_i \text{Var}(K_i|\mathcal{A}_n, \mathcal{Z}_n) =$

$n^{-2}\max_{z_l;l\in\{1,...,L\},a\in\{1,...,k\}}(C_{z_l,a})$. Proceeding in a similar fashion,

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=1}^{n}K_i|\mathcal{A}_n,\mathcal{Z}_n\right) &\geq n^{-2}\sum_{i=1}^{n}c^T\mathrm{Var}\left(\psi_1^1(W_i),...,\psi_k^1(W_i),\psi_1^4(W_i),...,\psi_k^4(W_i)|\mathcal{A}_n,\mathcal{Z}_n\right)\\ &\geq n^{-1}\min_{z_l;l\in\{1,...,L\},a\in\{1,...,k\}}(C_{z_l,a}).\end{aligned}$$

Putting these pieces together,

$$\tau_{i,n}\leq\max\tau_{i,n}\leq n^{-1}\left(\max_{z_l;l\in\{1,...,L\},a\in\{1,...,k\}}(C_{z_l,a})/\min_{z_l;l\in\{1,...,L\},a\in\{1,...,k\}}(C_{z_l,a})\right)=o(1).$$

□

B.4 Conditions for Guaranteed Efficiency Gain

In this section, we derive conditions for guaranteed efficiency gain of an AIPW estimator over the sample mean. The first condition is technical; it is given in Theorem B.1 below. Since it is quite difficult to interpret, we provide a much simpler condition for guaranteed efficiency gain in Theorem 2.3 in the main text. This result relies on a *universality* condition of the asymptotic variance: the asymptotic variance does not depend on the CAR scheme. We provide the proof of Theorem 2.3 at the end of this section.

Theorem B.1 (Guaranteed efficiency gain under CAR)**.** *For all randomization schemes satisfying condition* (B)*, a sufficient condition for* $\hat{\theta}_{\mathrm{AIPW}}$ *having guaranteed efficiency gain over the sample mean* $\bar{Y}$ *under CAR and simple randomization is*

$$\begin{aligned}&\mathrm{diag}\left\{\pi_a^{-1}\mathrm{Cov}\{Y_a-\mu_a(X),\mu_a(X)\},a=1,...,k\right\}-\mathrm{Cov}(Y^{(,)}\mu)+\mathrm{Var}(\mu)\\ &\qquad=E\left[\{R_Y(Z)-R_X(Z)\}\{\Omega_{\mathrm{SR}}-\Omega(Z)\}R_X(Z)\right].\end{aligned}\tag{B.10}$$

Proof. We will show that under condition (B) we have a guaranteed efficiency gain if the

following condition holds:

$$\begin{aligned}&\text{diag}\left\{\pi_a^{-1}\text{Cov}\{Y_a-\mu_a(X),\mu_a(X)\},a=1,...,k\right\}-\text{Cov}\{Y^{(-)}\mu(X),\mu(X)\}\\&=E\left[\{R_Y(Z)-R_X(Z)\}\{\Omega_{\text{SR}}-\Omega(Z)\}R_X(Z)\right].\end{aligned}$$

where $R_Y(Z)=\text{diag}\{\pi_a^{-1}E(Y_a-\theta_a|Z)\}$ and $R_X(Z)=\text{diag}\{\pi_a^{-1}E(\mu_a(X)-E(\mu_a(X))|Z)\}$, so that $R(Z)=R_Y(Z)-R_X(Z)$. In order to achieve a guaranteed efficiency gain over ANOVA, a simple sample mean, we would need the difference between the covariance matrix of ANOVA and $\hat{\theta}_{\text{AIPW}}$ to be positive definite. In what follows, we show that a sufficient condition for this difference being positive definite is that (B.10) holds. Let $\Sigma_{(B)}^{\bar{Y}}$ be the variance-covariance matrix for the sample mean $\bar{Y}$ under (B), given by the result in Theorem 2.1 (ii) with $\tilde{\mu}=0$ and $R(Z)=R_Y(Z)$:

$$\Sigma_{(B)}^{\bar{Y}}=\text{diag}\{\pi_a^{-1}\text{Var}(Y_a)\}-E(R_Y(Z)\{\Omega_{SR}-\Omega\}R_Y(Z)).$$

Then,

$$\begin{aligned}&\Sigma_{(B)}^{\bar{Y}}-\Sigma_{(B)}\\&=\underbrace{\text{diag}\{\pi_a^{-1}\text{Var}(Y_a)\}-\text{diag}\{\pi_a^{-a}\text{Var}\{Y_a-\mu_a(X)\}}_{(I)}\\&\quad\underbrace{-\text{Cov}\{Y,\mu(X)\}-\text{Cov}\{\mu(X),Y\}+\text{Var}\{\mu(X)\}}_{(II)}\\&\quad\underbrace{-E(R_Y(Z)\{\Omega_{SR}-\Omega(Z)\}R_Y(Z))+E\left[\{R_Y(Z)-R_X(Z)\}\{\Omega_{\text{SR}}-\Omega(Z)\}\{R_Y(Z)-R_X(Z)\}\right]}_{(III)}.\\&=\text{diag}\{\pi_a^{-1}\text{Var}(\mu_a(X))\}+2\text{diag}\{\pi_a^{-a}\text{Cov}\{\mu_a(X),Y_a-\mu_a(X)\}\\&\quad-\text{Cov}\{Y-\mu(X),\mu(X)\}-\text{Cov}\{\mu(X),Y-\mu(X)\}-\text{Var}\{\mu(X)\}\\&\quad-E(R_Y(Z)\{\Omega_{SR}-\Omega\}R_Y(Z))+E\left[\{R_Y(Z)-R_X(Z)\}\{\Omega_{\text{SR}}-\Omega(Z)\}\{R_Y(Z)-R_X(Z)\}\right]\end{aligned}$$

We can rewrite each of the above terms. Using the variance of a difference and the law of

total variance:

$$\begin{aligned}(I) &= \text{diag}\{\pi_a^{-1}\text{Var}(\mu_a(X))\} + 2\text{diag}\{\pi_a^{-1}\text{Cov}\{\mu_a(X), Y_a - \mu_a(X)\}\} \\ &= \text{diag}\{\pi_a^{-1}E(\text{Var}(\mu_a(X)|Z))\} + \text{diag}\{\pi_a^{-1}\text{Var}(E(\mu_a(X)|Z))\} \\ &\quad + 2\text{diag}\{\pi_a^{-1}E(\text{Cov}\{\mu_a(X), Y_a - \mu_a(X)|Z\})\} \\ &\quad + 2\text{diag}\{\pi_a^{-1}\text{Cov}\{E(\mu_a(X)|Z), E(Y_a - \mu_a(X)|Z)\}\}.\end{aligned}$$

For (II), we expand terms and then add and subtract the quadratic form of R_X:

$$\begin{aligned}&(II) \\ &= -E(R_X(Z)\{\Omega_{SR} - \Omega(Z)\}R_Y(Z)) \\ &\quad - E(R_Y(Z)\{\Omega_{SR} - \Omega(Z)\}R_X(Z)) + E(R_X(Z)\{\Omega_{SR} - \Omega(Z)\}R_X(Z)) \\ &= -E(R_X(Z)\{\Omega_{SR} - \Omega(Z)\}\{R_Y(Z) - R_X(Z)\}) - E(\{R_Y(Z) - R_X(Z)\}\{\Omega_{SR} - \Omega(Z)\}R_X(Z)) \\ &\quad - E(R_X(Z)\{\Omega_{SR} - \Omega(Z)\}R_X(Z)).\end{aligned}$$

For (III), we use law of total covariance and variance, and then use the definition of R_X and R_Y. As an example, we have

$$\begin{aligned}\text{Var}(\mu(X)) &= E(\text{Var}(\mu(X)|Z)) + \text{Var}(E(\mu(X)|Z)) \\ &= E(\text{Var}(\mu(X)|Z)) + E(E(\mu(X)|Z)E(\mu(X)|Z)^T) - E(\mu(X))E(\mu(X))^T \\ &= E(\text{Var}(\mu(X)|Z)) + E(R_X(Z)\pi\pi^T R_X(Z)).\end{aligned}$$

The same steps hold for the covariance terms. Therefore,

$$\begin{aligned}(III) = &-E(\text{Cov}\{Y - \mu(X), \mu(X)|Z\}) - E(\{R_Y(Z) - R_X(Z)\}\pi\pi^T R_X(Z)|Z) \\ &- E(\text{Cov}\{\mu(X), Y - \mu(X)|Z\}) - E(R_X(Z)\pi\pi^T\{R_Y(Z) - R_X(Z)\}|Z) \\ &- E(\text{Var}(\mu(X)|Z)) - E(R_X(Z)\pi\pi^T R_X(Z)).\end{aligned}$$

We now can combine terms across $(I)+(II)+(III)$. For example, considering the variance terms for μ, and recalling that $\Omega_{SR} = \text{diag}(\pi) - \pi\pi^T$, we have:

$$\begin{aligned}
&\text{diag}\{\pi_a^{-1}E(\text{Var}(\mu_a(X)|Z))\} + \text{diag}\{\pi_a^{-1}\text{Var}(E(\mu_a(X)|Z))\} \\
&\quad - E(R_X(Z)\{\Omega_{SR} - \Omega(Z)\}R_X(Z)) - E(\text{Var}(\mu(X)|Z)) - E(R_X(Z)\pi\pi^T R_X(Z)) \\
&= \text{diag}\{\pi_a^{-1}E(\text{Var}(\mu_a(X)|Z))\} - E(\text{Var}(\mu(X)|Z)) + E(R_X(Z)\Omega(Z)R_X(Z)) \\
&\quad + E(R_X(Z)\text{diag}(\pi)R_X(Z)) - E(R_X(Z)\pi\pi^T R_X(Z)) - E(R_X(Z)\Omega_{SR}R_X(Z)) \\
&= \text{diag}\{\pi_a^{-1}E(\text{Var}(\mu_a(X)|Z))\} - E(\text{Var}(\mu(X)|Z)) + E(R_X(Z)\Omega(Z)R_X(Z)).
\end{aligned}$$

We have a similar result for each of the covariance terms. Using the law of total covariance, and the fact that again $\Omega_{SR} = \text{diag}(\pi) - \pi\pi^T$, we have:

$$\begin{aligned}
&\text{diag}\{\pi_a^{-1}E(\text{Cov}(Y_a - \mu_a(X), \mu_a(X)|Z))\} - E(\text{Cov}(Y - \mu(X), \mu(X)|Z)) \\
&\quad = \text{diag}(\pi_a^{-1}\text{Cov}(Y_a - \mu_a(X), \mu_a(X))\} - \text{Cov}(Y - \mu(X), \mu(X)) \\
&\quad\quad - E(\{R_Y(Z) - R_X(Z)\}\Omega_{SR}R_X(Z)).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\text{diag}\{\pi_a^{-1}E(\text{Cov}(Y_a - \mu_a(X), \mu_a(X)|Z))\} - E(\text{Cov}(Y - \mu(X), \mu(X)|Z)) \\
&\quad + E(\{R_Y(Z) - R_X(Z)\}\Omega(Z)R_X(Z)) \\
&= \text{diag}\{\pi_a^{-1}\text{Cov}(Y_a - \mu_a(X), \mu_a(X))\} \\
&\quad - \text{Cov}(Y - \mu(X), \mu(X)) - E(\{R_Y(Z) - R_X(Z)\}\{\Omega_{SR} - \Omega(Z)\}R_X(Z)).
\end{aligned}$$

This also holds for the transpose matrix, i.e., dealing with terms like $\text{Cov}(\mu(X), Y - \mu(X))$.

Putting all of these parts together, we have the following result:

$$\begin{aligned}(I)+(II)+(III) = \text{diag}\{\pi_a^{-1}E(\text{Var}(\mu_a(X)|Z))\} - E(\text{Var}(\mu(X)|Z)) + E(R_X(Z)\Omega(Z)R_X(Z)) \\ + 2\text{diag}(\pi_a^{-1}E(\text{Cov}(\mu_a(X), Y_a - \mu_a(X)|Z)) \\ - \text{Cov}(Y-\mu(X),\mu(X)) - \text{Cov}(Y-\mu(X),\mu(X))^T \\ - E(\{R_Y(Z)-R_X(Z)\}\{\Omega_{SR}-\Omega(Z)\}R_X(Z)) \\ - E(\{R_Y(Z)-R_X(Z)\}\{\Omega_{SR}-\Omega(Z)\}R_X(Z))^T.\end{aligned}$$

With the condition given in 1 (d), all terms cancel out besides the first line. In other words, under 1 (d),

$$\Sigma_{(B)}^{\bar{Y}} - \Sigma_{(B)} = \text{diag}\{\pi_a^{-1}E(\text{Var}(\mu_a(X)|Z))\} - E(\text{Var}(\mu(X)|Z)) + E(R_X(Z)\Omega(Z)R_X(Z)).$$

Note that $\Omega(Z)$ is positive semidefinite. Furthermore, Lemma 1 from Ye et al. (2023) states that if M is a matrix with columns $m_1, ..., m_k$ and $\pi_1, ..., \pi_k$ are nonnegative constants that sum to one, then $\text{diag}(\pi_t^{-1}m_t^T m_t) - M^T M$ is positive semidefinite. This means that $\text{diag}\{\pi_a^{-1}E(\text{Var}(\mu_a(X)|Z))\} - E(\text{Var}(\mu(X)|Z))$ is positive semidefinite, as is its expectation. Therefore, $\Sigma_{(B)}^{\bar{Y}} - \Sigma_{(B)}$ is positive semidefinite and we have guaranteed efficiency gain. □

B.4.1 Proof of Theorem 2.3: Guaranteed Efficiency Gain through Universal Variance

Under (U), we have that $E\{Y^{(-)}\mu(X) \mid Z\} = \theta - E\{\mu(X)\}$ almost surely. Since we defined $\tilde{\mu} = \mu(X) - E(\mu(X)) - \theta$, this means that under (U), $E\{Y^{(-)}\tilde{\mu}(X) \mid Z\} = 0$. As a result, $\psi^2 = \psi^3 = 0$ almost surely. Therefore, $V^{(5)} := V^{(3,5)} = \text{Var}(E(\tilde{\mu}|Z))$.

Define $M_n^{(1,4)} := \left[V^{(1,4)}\right]^{-1}\sqrt{n}P_n(\psi^1 + \psi^4)$ and $M_n^{(5)} := \left[V^{(3,5)}\right]^{-1}\sqrt{n}P_n(\psi^5)$. Then by the continuous mapping theorem, $M_n^{(1,4)}|\mathcal{A}_n, \mathcal{Z}_n \xrightarrow{d} N(0, I_K)$ and $M_n^{(5)} \xrightarrow{d} N(0, I_K)$. Furthermore, these two normal distributions are asymptotically independent. Applying Lemma B.17, $(M_n^{(1,4)}, M_n^{(5)}) \xrightarrow{d} (M^{(1,4)}, M^{(5)})$, where $M^{(1,4)}$ and $M^{(5)}$ are independent multivariate normal random variables. Therefore, by the continuous mapping theorem, we can both scale

them by their respective variance-covariance matrices, and add the two random variables together to get that under (B),

$$\sqrt{n}(\hat{\theta}_{\text{AIPW}} - \theta) \equiv \sqrt{n}P_n(\psi^1 + \psi^4 + \psi^5) + o_P(1) \xrightarrow{d} N(0, \Sigma_{(U)})$$

where $\Sigma_{(U)} = V^{(1,4)} + V^{(5)}$. Note that $\Sigma_{(U)}$ has no dependence on the covariate-adaptive randomization scheme, so the asymptotic distribution (including the asymptotic variance) of $\hat{\theta}_{\text{AIPW}}$ is universal under (U).

It is now straightforward to show guaranteed efficiency gain under (U) + (G). Under (U), $R(Z) = 0$. Therefore, the RHS of (B.10) is 0. If (G) holds, then both covariance terms on the LHS of (G) are also 0. Since both sides of (G) are 0, (G) is satisfied.

B.5 Joint Calibration

In Section B.5.1, we first show that Joint Calibration, as in Theorem 2.4, satisfies Assumptions 2.2 and 2.3, which is the key requirement for Theorem 2.1 to hold. Therefore, Theorem 2.1 applies to the Joint Calibration estimator, which we formalize in Section B.5.2. In Section B.5.3, we include more details about Example 2.1 for using Joint Calibration with GLMs. In Section B.5.4, we show that using the Joint Calibration with cross-fitted $\hat{\mu}_a$'s still has the same influence function.

B.5.1 Joint Calibration Satisfies Assumptions 2.2 and 2.3

Here we show that Assumptions 2.2 and 2.3 are satisfied for the jointly calibrated $\hat{\mu}_a^*(X_i)$ and a limiting function $\mu_a^*(X_i)$. Recall that for joint calibration, we have defined

$$\hat{\mu}_a^*(X_i) = \hat{\alpha}_a^T Z_i + \hat{\beta}_a^T \hat{\mu}(X_i),$$
$$(\hat{\alpha}_a, \hat{\beta}_a) = \arg\min_{\alpha_a, \beta_a} \sum_{i:A_i=a} \left\{Y_{a,i} - \alpha_a^T Z_i - \beta_a^T \hat{\mu}(X_i)\right\}^2.$$

Define $\mu_a^*(X_i) = \alpha_a^{*T} Z_i + \beta_a^{*T} \mu(X_i)$, where $\mu(X_i) = (\mu_1(X_i), ..., \mu_k(X_i))^T$ is the vector of all μ functions. Define the true regression parameters as the solution to the corresponding population problem:

$$(\alpha_a^*, \beta_a^*) = \arg\min_{\alpha_a, \beta_a} E\left\{\left(Y_a - \alpha_a^T Z - \beta_a^T \mu^*(X)\right)^2\right\}. \tag{B.11}$$

By assumption, Assumptions 2.2 and 2.3 are satisfied for $\hat{\mu}_a(X)$ using some $\mu_a(X)$ for each a. Our goal is to show that, given Assumptions 2.2 and 2.3 for each $\hat{\mu}_a(X)$, Assumptions 2.2 and 2.3 are satisfied for $\hat{\mu}_a^*(X)$ using the above $\mu_a^*(X)$.

Consistency of Regression Coefficients

We first establish that $\hat{\alpha}_a \to \alpha_a^*$ and $\hat{\beta}_a \to \beta_a^*$ in probability. If $\hat{\mu}$ were not random this would be trivial as they are just coefficients from linear regression. Following the same logic as in

Section 3.1 of Ye et al. (2023),

$$\begin{pmatrix}\hat{\alpha}_a\\ \hat{\beta}_a\end{pmatrix} = \frac{n}{n_a}\left\{\sum_{i=1}^{n}\begin{pmatrix}Z_i\\ \hat{\mu}(X_i)\end{pmatrix}\begin{pmatrix}Z_i\\ \hat{\mu}(X_i)\end{pmatrix}^T\right\}^{-1}\sum_{i:A_i=a}\begin{pmatrix}Z_iY_i\\ \hat{\mu}(X_i)Y_i\end{pmatrix}$$

Therefore, to prove that $\hat{\alpha}_a \to \alpha_a^*$, and $\hat{\beta}_a \to \beta_a^*$, it suffices to show that, in probability,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\begin{smallmatrix}Z_i\\ \hat{\mu}(X_i)\end{smallmatrix}\right)\left(\begin{smallmatrix}Z_i\\ \hat{\mu}(X_i)\end{smallmatrix}\right)^T \to E\left\{\left(\begin{smallmatrix}Z\\ \mu(X)\end{smallmatrix}\right)\left(\begin{smallmatrix}Z\\ \mu(X)\end{smallmatrix}\right)^T\right\}$$
$$\frac{1}{n_a}\sum_{i:A_i=a}\left(\begin{smallmatrix}Z_iY_{a,i}\\ \hat{\mu}(X_i)Y_{a,i}\end{smallmatrix}\right) \to E\left\{\left(\begin{smallmatrix}ZY_a\\ \mu(X)Y_a\end{smallmatrix}\right)\right\}$$

The components without $\hat{\mu}$ are straightforward. Specifically, $\frac{1}{n}\sum_{i=1}^{n} Z_iZ_i^T \to E(ZZ^T)$ by the law of large numbers since Z_i are iid. We cite Ye et al. (2023)'s work in Lemma 3 of their supplemental materials to claim that $\frac{1}{n_a}\sum_{i:A_i=a} Z_iY_{a,i} \to E(ZY_a)$ in probability. For terms with $\hat{\mu}$, we start with $\frac{1}{n_a}\sum_{i:A_i=a} Y_{a,i}\hat{\mu}(X_i)$:

$$\frac{1}{n_a}\sum_{i:A_i=a} Y_{a,i}\hat{\mu}(X_i) - E(Y_i\mu(X_i)) = \underbrace{\frac{1}{n_a}\sum_{i:A_i=a} Y_{a,i}\mu(X_i) - E(Y_{a,i}\mu(X_i))}_{(1)}$$
$$+\underbrace{\frac{1}{n_a}\sum_{i:A_i=a} Y_{a,i}\hat{\mu}(X_i) - \frac{1}{n_a}\sum_{i:A_i=a} Y_{a,i}\mu(X_i)}_{(2)}.$$

Term (1) is $o_P(1)$ by the same logic of Ye et al. (2023) Lemma 3. For term (2), we have similar types of empirical process terms as we dealt with in the proof of Theorem 2.1, when finding the influence function. However, the difference here is that we do not need a specific

rate; we just need it to be $o_P(1)$. We have:

$$\begin{aligned}
(2) &= \frac{1}{n_a}\sum_{i:A_i=a} Y_{a,i}\hat{\mu}(X_i) - \frac{1}{n_a}\sum_{i:A_i=a} Y_{a,i}\mu(X_i) \\
&= \hat{\pi}_a^{-1}\frac{1}{n}\sum_{i:A_i=a}\left[Y_{a,i}\hat{\mu}(X_i) - P_0(Y_{a,i}\hat{\mu}(X_i))\right] - \hat{\pi}_a^{-1}\frac{1}{n}\sum_{i=1}^{n} I_{a,i}[Y_{a,i}\mu(X_i) - P_0(Y_{a,i}\mu(X_i))] \\
&\quad + P_0(Y_{a,i}(\hat{\mu}(X_i) - \mu(X_i))) \\
&= \hat{\pi}_a^{-1}\frac{1}{n}\sum_{i=1}^{n} I_{a,i}\left\{Y_{a,i}\hat{\mu}(X_i) - P_0 Y_{a,i}\hat{\mu}(X_i) - Y_{a,i}\mu(X_i) + P_0 Y_{a,i}\mu(X_i)\right\} + P_0(Y_{a,i}(\hat{\mu}(X_i) - \mu(X_i))).
\end{aligned}$$

Since $\hat{\pi}_a^{-1} \to \pi_a^{-1}$ in probability, by Slutsky's Lemma, to show that (2) is $o_P(1)$, it suffices to show that

(I) $P_0(Y_a\hat{\mu}(X) - Y_a\mu(X)) = o_P(1)$. By the Cauchy-Schwartz Inequality, we have

$$\begin{aligned}
|P_0(Y_a\hat{\mu}(X) - Y_{a,i}\mu(X))| &\leq \|Y_a\|_{L_2}\|\hat{\mu}(X) - \mu(X)\|_{L_2} \\
&= o_P(1)
\end{aligned}$$

by assumption since $\|\hat{\mu}(X_i) - \mu(X_i)\|_{L_2} \to 0$ in probability, and since Y_a has a finite second moment. Note that this also holds replacing $P_0(Y_a[\hat{\mu}(X) - \mu(X)])$ with $P_0(|Y_a||\hat{\mu}(X) - \mu(X)|)$, therefore $P_0(|Y_a||\hat{\mu}(X) - \mu(X)|)$ is also $o_P(1)$.

(II) $\frac{1}{n}\sum_{i=1}^{n} I_{a,i}\left\{Y_{a,i}\hat{\mu}(X_i) - P_0 Y_{a,i}\hat{\mu}(X_i) - Y_{a,i}\mu(X_i) + P_0 Y_{a,i}\mu(X_i)\right\} = o_P(1)$. This term is similar to the term $\mathcal{M}$ in the proof of Theorem 2.1 (i), but instead of $\hat{\mu}_a(X_i)$ and $\mu_a(X_i)$, we have $Y_{a,i}\hat{\mu}(X_i)$ and $Y_{a,i}\mu(X_i)$ respectively. However, we do not want to follow the techniques of $\mathcal{M}$ in the proof of Theorem 2.1 (i) because its assumptions are too strict. We only need the term to be $o_P(1)$ rather than $o_P(1/\sqrt{n})$. So in this case, we can use some more basic techniques. Term (II) can be broken up into two terms, each of which is $o_P(1)$ by applications of Assumption 2.2, Cauchy-Schwartz, and Markov's Inequality:

(i) $\frac{1}{n}\sum_{i=1}^{n} I_{a,i}\left\{Y_{a,i}[\hat{\mu}(X_i)-\mu(X_i)]\right\}$. Here we can apply the finite version of Cauchy-Schwartz:

$$\left|\frac{1}{n}\sum_{i=1}^{n} I_{a,i}\left\{Y_{a,i}[\hat{\mu}(X_i)-\mu(X_i)]\right\}\right| \leq \sqrt{\frac{1}{n}\sum_{i:A_i=a}[Y_{a,i}]^2}\sqrt{\frac{1}{n}\sum_{i:A_i=a}[\hat{\mu}(X_i)-\mu(X_i)]^2}$$
$$\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}[Y_{a,i}]^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}(X_i)-\mu(X_i)]^2}.$$

We know that $\sqrt{\frac{1}{n}\sum_{i=1}^{n}[Y_{a,i}]^2} - \|Y_{a,i}\|_{L_2} = o_P(1)$ by the law of large numbers and continuous mapping theorem. We also know that $\sqrt{\frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}(X_i)-\mu(X_i)]^2} = o_P(1)$ by Assumption:

$$\begin{aligned}
\lim_{n\to\infty} P\left(\frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}(X_i)-\mu_a(X_i)]^2 > \epsilon\right) &= \lim_{n\to\infty} E\left(P\left(\frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}(X_i)-\mu(X_i)]^2 > \epsilon \middle| \hat{\mu}\right)\right) \\
\text{(DCT)} \quad &= E\left(\lim_{n\to\infty} P\left(\frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}(X_i)-\mu(X_i)]^2 > \epsilon \middle| \hat{\mu}\right)\right) \\
\text{(Markov's Inequality)} \quad &\leq E\left(\lim_{n\to\infty} P\left(\frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}(X_i)-\mu(X_i)]^2 \middle| \hat{\mu}\right)\right)/\epsilon \\
\text{(Assumption 2.2)} \quad &= E(0)/\epsilon = 0.
\end{aligned}$$

Therefore, by applying Slutsky's Theorem, we can see that the whole term is $o_P(1)$.

(ii) $\frac{1}{n}\sum_{i=1}^{n} I_{a,i}P_0\left\{Y_{a,i}[\hat{\mu}(X_i)-\mu(X_i)]\right\}$. To see that this term is $o_P(1)$, note that the P_0 term is not dependent on i, therefore, it becomes $\hat{\pi}_a P_0\left\{Y_a[\hat{\mu}(X)-\mu(X)]\right\}$.

Furthermore,

$$\begin{aligned}|\hat{\pi}_a P_0 \{Y_{a,i}[\hat{\mu}(X) - \mu(X)]\}| &\leq \hat{\pi}_a P_0 |Y_a[\hat{\mu}(X) - \mu(X)]| \\ &\leq \hat{\pi}_a \|\hat{\mu}(X) - \mu(X)\|_{L_2} \|Y_a\|_{L_2} \\ &= \hat{\pi}_a o_P(1) \\ &= (\hat{\pi}_a - \pi_a) o_P(1) + \pi_a o_P(1) = o_P(1)\end{aligned}$$

where we have used Cauchy-Schwartz, the fact that $\|\hat{\mu}(X) - \mu(X)\|_{L_2} = o_P(1)$ by Assumption 2.2, and that Y_a has a bounded second moment.

Next, we can see that $\frac{1}{n}\sum_{i=1}^n Z_i\hat{\mu}(X_i) \to E(Z\mu(X))$ by using the following expansion:

$$\frac{1}{n}\sum_{i=1}^n Z_i\hat{\mu}(X_i) - E(Z\mu(X)) = (P_n - P_0)Z_i\mu(X_i) + P_0(Z_i[\hat{\mu}(X_i) - \mu(X_i)]) + (P_n - P_0)(Z_i\hat{\mu} - Z_i\mu)$$

The first term is $o_P(1)$ by the law of large numbers. The second term is $o_P(1)$ for the same reason as (I) above replacing $Y_{a,i}$ with Z_i, noting that Z_i is just an indicator function. The last term is $o_P(1)$ because the functions $(Z_i, X_i) \to Z_i\hat{\mu}(X_i)$ and $(Z_i, X_i) \to Z_i\mu(X_i)$ satisfy 2.3, the Donsker condition, assuming that it is already satisfied for $\hat{\mu}(X_i)$ and $\mu(X_i)$.

Finally, we would like to show that

$$\frac{1}{n}\sum_{i=1}^n \hat{\mu}(X_i)\hat{\mu}(X_i)^T \to E(\mu(X)\mu(X)^T) \tag{B.12}$$

in probability. We focus on each of the components of the matrix $E(\mu(X_i)\mu(X_i)^T)$: we would like to show that $\frac{1}{n}\sum_{i=1}^n \hat{\mu}_a(X_i)\hat{\mu}_b(X_i) \to E(\mu_a(X_i)\mu_b(X_i))$ in probability, where $a = 1, \ldots, K$, $b = 1, ..., K$. Let $\hat{g}_{a,b} = \hat{\mu}_a(X_i)\hat{\mu}_b(X_i)$ and $g_{a,b} = \mu_a(X_i)\mu_b(X_i)$. Then,

$$\frac{1}{n}\sum_{i=1}^n \hat{\mu}_a(X_i)\hat{\mu}_b(X_i) - E(\mu_a(X)\mu_b(X)) = (P_n - P_0)g_{a,b} + P_n(\hat{g}_{a,b} - g_{a,b}).$$

$(P_n - P_0)g_{a,b} = o_P(1)$ by the law of large numbers. Therefore, for (B.12) to hold, we need $P_n(\hat{g}_{a,b} - g_{a,b}) = o_P(1)$. We can expand this term as:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\underbrace{[\hat{\mu}_a(X_i) - \mu_a(X_i)]\,[\hat{\mu}_b(X_i) - \mu_b(X_i)]}_{(III)} + \underbrace{\mu_b(X_i)\,[\hat{\mu}_a(X_i) - \mu_a(X_i)]}_{(IV)} + \underbrace{\mu_a(X_i)\,[\hat{\mu}_b(X_i) - \mu_b(X_i)]}_{(V)}\right)$$

We will use Cauchy-Schwartz Inequality. In (II) (i) from before, we showed that $\sqrt{\frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}(X_i) - \mu(X_i)]^2} = o_P(1)$, so this holds for the μ_a and μ_b components of the μ vector. Therefore, we have

$$(III) \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_a(X_i) - \mu_a)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_b(X_i) - \mu_b)^2} = o_P(1)$$

$$\begin{aligned}(IV) &\leq \left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mu_b^2(X_i)} - \|\mu_b\|_{L_2(P)}\right)\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_a(X_i) - \mu_a)^2} + \|\mu_b\|_{L_2(P)}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_a(X_i) - \mu_a)^2} \\ &= o_P(1).\end{aligned}$$

Term (V) is identical to term (IV) with μ_a and μ_b switched. Therefore, $P_n(\hat{g}_{a,b} - g_{a,b}) = o_P(1)$.

Assumption 2.2 for Joint Calibration: Stability

Assume that for every a, there exists a function $\mu_a \in L_2(P)$ on $\mathscr{X}$ such that, as $n \to \infty$ (Assumption 2.2), $\|\hat{\mu}_a - \mu_a\|_{L_2} \to 0$ in probability. First note that $\mu_a^* \in L_2(P)$ as well, because:

$$\|\mu_a^*\|_{L_2} \leq \left\|\alpha_a^{*T}\right\|_2 + \sum_{a'=1}^{k}|\beta_{a,a'}^{*T}|\,\|\mu_a'\|_{L_2} < \infty.$$

Then:

$$\begin{aligned}\|\hat{\mu}_a^* - \mu_a^*\|_{L_2} &\leq \|\hat{\alpha}_a - \alpha_a^*\|_{L_2} + \sum_{a'=1}^{k} \left\| \hat{\beta}_{a,a'}\hat{\mu}_{a'} - \beta_{a,a'}^*\mu_{a'}^* \right\|_{L_2} \\ &\leq \|\hat{\alpha}_a - \alpha^*{}_a\|_2 + \sum_{a=1}^{K} |\hat{\beta}_{a,a'} - \beta_{a,a'}| \, \|\hat{\mu}_{a'}\|_{L_2} + |\beta_{a,a'}| \, \|\hat{\mu}_{a'} - \mu_{a'}^*\|_{L_2} \\ &\leq \|\hat{\alpha}_a - \alpha^*{}_a\|_2 + \sum_{a'=1}^{k} |\hat{\beta}_{a,a'} - \beta_{a,a'}| \left(\|\mu_{a'}\|_{L_2} + \|\hat{\mu}_{a'} - \mu_{a'}\|_{L_2} \right) + |\beta_{a,a'}| \, \|\hat{\mu}_{a'} - \mu_{a'}^*\|_{L_2} .\end{aligned}$$

This entire expression on the RHS is $o_P(1)$ because of the consistency of the regression coefficients that we proved in Section B.5.1 and because of Assumption 2.2 on $\hat{\mu}_a$. Therefore, Assumption 2.2 holds for $\hat{\mu}_a^*$ and μ_a^*.

2.3 for Joint Calibration: Donsker Condition

Assume that 2.3 (i-iii) holds for $\mu_a(X_i)$ and $\hat{\mu}_a(X_i)$, $a = 1, \ldots, K$. Define

$$\mathcal{F}_a^* := \{\alpha^T Z_i + \beta^T \mu(X_i) : \mu_a \in \mathcal{F}_a, |\alpha_a - \alpha_a^*| < 1, |\beta_a - \beta_a^*| < 1, a = 1, \ldots, K\}.$$

Then,

(i) $\mu_a^* \in \mathcal{F}_a^*$. By assumption, each $\mu_a \in \mathcal{F}_a$, therefore, for $\mu_a^* \in \mathcal{F}_a^*$.

(ii) $P(\hat{\mu}_a^* \in \mathcal{F}_a^*) \to 1$ as $n \to \infty$. We can write

$$P(\hat{\mu}_a^* \notin \mathcal{F}_a^*) \leq \sum_{a'=1}^{k} \left\{ P(\hat{\mu}_{a'} \notin \mathcal{F}_{a'}) + P(|\hat{\alpha}_{a,a'} - \alpha_{a,a'}^*| \geq 1) + P(|\hat{\beta}_{a,a'} - \beta_{a,a'}^*| \geq 1) \right\} \to 0$$

by 2.3 on $\hat{\mu}_{a'}$ and by definition of convergence in probability.

(iii) $\int_0^1 \sup_Q \sqrt{\log N(\mathcal{F}_a^*, \|\cdot\|_{L_2(Q)}, s)} \, ds < \infty$. Let T_a^μ be a minimal ϵ-cover for $\mathcal{F}_a$, T_l^α be a minimal ϵ-cover for the unit ball in 1-dimension of $\alpha_l - \alpha_l^*$ (letting $\alpha_a = (\alpha_1, ..., \alpha_L)$ and ignoring the a subscript on α_a for simplicity), and similarly T_a^β be a minimal ϵ-cover

for the unit ball in 1-dimension of $\beta_a - \beta_a^*$ (letting $\beta_a = (\beta_1, ..., \beta_K)$ and ignoring the a subscript on β_a for simplicity).

Take any function $f \in \mathcal{F}_a^*$. Select the elements of T_a^μ, $T_l^\alpha, l = 1, ..., L$, and $T_a^\beta, a = 1, \ldots, K$ such that they cover the α_l's, β_a's and μ_a's of the chosen f, and denote them as α_l', β_a', and μ_a', respectively. All of these functions are contained within the class

$$T := \{(\alpha_1, ..., \alpha_L)^T Z_i + (\beta_1, ..., \beta_K)^T (\mu_1(X_i), ..., \mu_k(X_i)) : \alpha_l - \alpha_l^* \in T_l^\alpha, \beta_a - \beta_a^* \in T_a^\beta, \mu_a \in T_a^\mu\}.$$

We define the function $f'(Z_i, X_i) := (\alpha_1', ..., \alpha_L')^T Z_i + (\beta_1', ..., \beta_K')^T (\mu_1'(X_i), ..., \mu_k'(X_i))$ as the covering function for f. Then:

$$\begin{aligned} \|f - f'\|_{L_2} &\leq \sum_{l=1}^{L} |\alpha_l - \alpha_l'| + \sum_{a=1}^{K} |\beta_a - \beta_a'| \, \|\mu_a\|_{L_2} + |\beta_a'| \, \|\mu_a - \mu_a'\|_{L_2} \\ &< L\epsilon + 2KC\epsilon \end{aligned}$$

where $C > \max_a\{|1 + \beta_a^*|, \|\mu_a\|_{L_2}, a = 1, \ldots, K\}$, which we know is $< \infty$ because $\beta_a^* < \infty$ and functions in $\mathcal{F}_a$ are square integrable. Therefore, we can find a constant $1 \leq C' < \infty$ such that $\|f - f'\|_{L_2} < C'\epsilon$.

Now that we have shown that T is a $C'\epsilon$ cover for $\mathcal{F}_a^*$, we will show that the uniform entropy integral also holds. We know that $N(\mathcal{F}_a^*, \|\cdot\|_{L_2}, C'\epsilon) \leq \prod_{l=1}^{L} |T_l^\alpha| \prod_{a=1}^{K} |T_a^\beta||T_a^\mu|$. Therefore,

$$\log N(\mathcal{F}_a^*, \|\cdot\|_{L_2}, C'\epsilon) \leq \sum_{a=1}^{K} \log N(\mathcal{F}_a, \|\cdot\|_{L_2}, \epsilon) + \sum_{a=1}^{K} (M/\epsilon) + \sum_{l=1}^{L} (M/\epsilon)$$

for some constant $M < \infty$, where the covering numbers for the unit ball in 1-dimension is given by Example 5.8 in Wainwright (2019). Taking the square root, the supremum,

then the integral, we have:

$$\int_0^1 \sup_Q \sqrt{\log N(\mathcal{F}_a^*, \|\cdot\|_{L_2(Q)}, C'\epsilon)}d\epsilon \leq \sum_{a=1}^{K} \int_0^1 \sup_Q \sqrt{\log N(\mathcal{F}_a, \|\cdot\|_{L_2}, \epsilon)}d\epsilon + (K+L)M \int_0^1 \sqrt{(1/\epsilon)}d\epsilon.$$

$\int_0^1 \sqrt{(1/\epsilon)}d\epsilon$ is finite, and $\int_0^1 \sup_Q \sqrt{\log N(\mathcal{F}_a, \|\cdot\|_{L_2}, \epsilon)}d\epsilon$ for all a is finite by 2.3. Therefore, the RHS of the above is finite. Finally, for the LHS, let $\epsilon' = C'\epsilon$. Then, LHS is equivalent to

$$(C')^{-1} \int_0^{C'} \sup_Q \sqrt{\log N(\mathcal{F}_a^*, \|\cdot\|_{L_2(Q)}, \epsilon')}d\epsilon'$$

Since $1 \leq C' < \infty$, we can see that $\int_0^1 \sup_Q \sqrt{\log N(\mathcal{F}_a^*, \|\cdot\|_{L_2(Q)}, \epsilon)}d\epsilon < \infty$.

B.5.2 Proof of Theorem 2.4

We have already shown that Assumptions 2.2 and 2.3 hold for $\hat{\mu}_a^*$ with respect to the function $\mu_a^*(X_i) = \alpha_a^{*T} Z_i + \beta_a^{*T} \mu(X_i)$. Let $\hat{W} = (Z^T, \hat{\mu}(X))$, $W = (Z^T, \mu(X))$, and $\Sigma_W = \text{Var}(W)$. Let $\gamma_a = (\alpha_a^{*T}, \beta_a^{*T})^T$ and $\hat{\gamma}_a = (\hat{\alpha}_a^T, \hat{\beta}_a^T)^T$. With this simplified notation, $\hat{\mu}_a^*(X) = \hat{\gamma}_a^T \hat{W}$, and $\mu_a^*(X) = \gamma_a^T W$. Based on the fact that we have shown that our OLS regression coefficients are consistent for their population minimizers based on (B.11), we know that $\hat{\gamma}_a \to \gamma_a \equiv \text{Var}(W)^{-1}\text{Cov}(W, Y_a)$.

To claim universality with (U), it suffices to show that for every a, $E\{Y_a - \mu_a^*(X) + P_0\mu_a^* - \theta_a \mid Z\} = 0$. This is satisfied by the population score equations based on how we have defined α_a^* and β_a^*, because for each $l = 1, ..., L$, we have $E[1(Z_i = l)\{Y_a - \mu_a^*(X)\}] = 0$. Therefore, we can directly apply Theorem 2.3 to obtain the variance of $\hat{\mu}_a^*$. Furthermore, now that we have established universality, we only need the following to hold in order to claim guaranteed efficiency gain $\text{Cov}\{Y_a - \mu_a^*(X), \mu_b^*(X)\} = 0$, for all $a = 1, \ldots, K$, $b = 1, ..., K$. By the definition of covariance, it suffices to show that both $E((Y_a - \mu_a^*(X))\mu_b^*(X)) = 0$

and $E(Y_a - \mu_a^*(X)) = 0$. The latter is satisfied by population score equations for group a since we have a Z-specific intercept, and because we can take expectation over Z. For $E((Y_a - \mu_a^*(X))\mu_b^*(X))$, we have

$$\begin{aligned} E((Y_a - \mu_a^*(X))\mu_b^*(X)) &= E((Y_a - \mu_a^*(X))\{\alpha_b^{*T} Z + \beta_b^{*T}\mu(X)\}) \\ &= \alpha_b^{*T} E(Z(Y_a - \mu_a^*(X))) + \beta_b^{*T} E(\mu(X)(Y_a - \mu_a^*(X))). \end{aligned}$$

Both of the terms above are also zero by the population score equations for group a. Using the above result, we can use our results from Theorem 2.3 applied under (G) that the asymptotic variance of $\hat{\mu}_a^*$ is

$$\begin{aligned} V &= \text{diag}\left[\pi_a^{-1}\text{Var}\{Y_a - \mu_a^*(X)\}, a = 1, \ldots, K\right] + \text{Cov}\{\mu^*(X), Y\} \\ &= \text{diag}\left[\pi_a^{-1}\text{Var}\{Y_a - \gamma_a^T W\}, a = 1, \ldots, K\right] + \text{Cov}((\mu_1^*(X), ..., \mu_K^*(X))^T, (Y_1, \ldots, Y_K)^T). \end{aligned}$$

The (a, b)th entry of $\text{Cov}((\mu_1^*(X), ..., \mu_K^*(X))^T, (y_1, ..., y_K)^T)$ is given by

$$\begin{aligned} \text{Cov}(\mu_1^*(X), Y_a) &= \text{Cov}(\gamma_a^T W, Y_b) \\ &= \gamma_a^T \text{Cov}(W, Y_b) \\ &= \text{Cov}(W, Y_a)^T \text{Var}(W)^{-1} \text{Cov}(W, Y_b) \\ &= \text{Cov}(W, Y_a)^T \text{Var}(W)^{-1} \text{Var}(W) \text{Var}(W)^{-1} \text{Cov}(W, Y_b) \\ &= \gamma_a^T \text{Var}(W) \gamma_b \end{aligned}$$

Therefore, the whole matrix $\text{Cov}((\mu_1^*(X), ..., \mu_K^*(X))^T, (Y_1, \ldots, Y_K)^T)$ can be written as $\Gamma^T \Sigma_W \Gamma$ where $\Gamma = (\gamma_1, \ldots, \gamma_K)$. Thus,

$$V = \text{diag}\left[\pi_a^{-1}\text{Var}\{Y_a - \mu_a^*(X)\}, a = 1, \ldots, K\right] + \Gamma^T \Sigma_W \Gamma.$$

Based on the fact that we have shown that (U) and (G) are satisfied, this means that

$\hat{\theta}_{\text{JC}}$ has guaranteed efficiency gain over the sample mean $\bar{Y}$, as that is the direct result of Theorem 2.3.

B.5.3 Example: Joint Calibration for Generalized Linear Models

Consider a GLM as in Example 1 using a canonical link function g, and where the parameters for the GLM are fitted using working models separately for each treatment group, i.e., the limiting working model is given by

$$\mu_a(X) = g^{-1}(\alpha_a + \beta_a^T X).$$

The AIPW estimator for group a is $\hat{\theta}_{\text{AIPW,a}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_a(X_i)$, because prediction unbiasedness is satisfied for GLMs with canonical link functions. Also, due to the population score equations, $E(Y_a - \mu_a(X)) = 0$. We can use this to simplify the resulting asymptotic variance formula. Let $r_a(X) = Y_a - g^{-1}(\alpha_a + \beta_a^T(X))$ be the "residual" from the GLM prediction. Let $R(Z) = \text{diag}\{\pi_a^{-1}E(r_a(X)|Z), a = 1\ldots, k\}$. The asymptotic variance is readily available from Theorem 2.1, i.e.,

$$\begin{aligned} V_{\text{AIPW}} = \text{diag}[\pi_a^{-1}\text{Var}(r_a(X)), a = 1, \ldots, K] \\ + \text{Cov}([\mu_1(X), \ldots, \mu_K(X)]) \\ + \text{Cov}([Y_1, \ldots, Y_K]^T[\mu_1(X), \ldots, \mu_K(X)]) + \text{Cov}([\mu_1(X), \ldots, \mu_K(X)]^T[Y_1, \ldots, Y_K]) \\ + \underbrace{E[R(Z)\{\Omega_{SR} - \Omega(Z)\}R(Z)]}_{(\star)}. \end{aligned}$$

If all joint levels of Z are included as covariates in X, then as mentioned in Example 1, the population score equations show that $E(r_a(X)|Z) = 0$ for all a, therefore $R(Z) = 0$. Then, term $(\star)$ above is 0, making V_{AIPW} invariant to the CAR scheme. As stated in the text, if $r_a(X)$ is not a very good predictor for Y_a, then it is possible that V_{AIPW} could be larger than the variance of the sample mean, even under simple randomization. This motivates

joint calibration. If we apply the joint calibration estimator on top of our GLM as stated in §4, with covariates $W = (Z^T, \mu_1(X), \ldots, \mu_K(X))^T$, and treatment-group specific linear regression parameters $\gamma_a \in \mathbb{R}^{L+K}$ (one intercept per joint strata level and a coefficient for each working model from the first stage, we have the new limiting function $\mu_a^*(W) = \gamma_a^T W$. Because this is itself a GLM, and Z are included as covariates, $(\star)$ is once again 0, with $R(Z)$ defined with $r_a^*(W) = Y_a - \gamma_a^T W$ in place of $r_a(X)$. Furthermore, because of the population score equations with linear regression and treatment-by-covariate interactions, joint calibration satisfies condition (G) (see proof of Theorem 2.4 for more details), which results in the following asymptotic variance formula:

$$\begin{aligned}
V_{\text{JC}} &= \text{diag}\left[\pi_a^{-1}\text{Var}(Y_a - \gamma_a^T W), a = 1, \ldots, K\right] + \text{Cov}([\gamma_1^T, \ldots, \gamma_K^T]^T W, [Y_1, \ldots, Y_K]) \\
&= \text{diag}[\pi_a^{-1}\left\{\text{Var}(Y_a) + \gamma_a^T \text{Var}(W)\gamma_a - 2\text{Cov}(Y_a, \gamma_a^T W)\right\}, a = 1, \ldots, K] \\
&\quad + \text{Cov}([\text{Cov}(W, Y_1), \ldots, \text{Cov}(W, Y_K)]^T \Sigma_W^{-1} W, [Y_1, \ldots, Y_K]) \\
&= \text{diag}\left[\pi_a^{-1}\left\{\text{Var}(Y_a) - \gamma_a^T \Sigma_W \gamma_a\right\}, a = 1, \ldots, K\right] + (\gamma_1^T, \ldots, \gamma_K^T)^T \Sigma_W (\gamma_1, \ldots, \gamma_K),
\end{aligned}$$

since the definition of $\gamma_a^T = \sum_W^{-1} \text{Cov}(W, Y_a)$. We can then construct a variance estimator for V_{JC} as follows:

$$\hat{V}_{\text{JC}} = \text{diag}\left[\hat{\pi}_a^{-1}\left\{\hat{\sigma}_a - \hat{\gamma}_a^T \hat{\Sigma}_W \hat{\gamma}_a\right\}, a = 1, \ldots, K\right] + (\hat{\gamma}_1^T, \ldots, \hat{\gamma}_K^T)^T \hat{\Sigma}_W (\hat{\gamma}_1, \ldots, \hat{\gamma}_K)$$

where $\hat{\sigma}_a$ is the sample variance of Y_a among individuals with $A_i = a$, $\hat{\Sigma}_W$ is the sample covariance matrix for $\hat{W} = (Z^T, \hat{\mu}_1(X), \ldots, \hat{\mu}_K(X))^T$, and $\hat{\gamma}_a = (\hat{W}_a^T \hat{W}_a)^{-1} \hat{W}_a^T Y_a\{A_i = a\}$ is the estimated coefficient for the ath group, where $\hat{W}_a$ is the design matrix stacking $\hat{W}$ for those in group a, and $Y_a\{A_i = a\}$ is the vector of the outcome Y_a but only for those in group a. Alternatively, the diagonal terms in the first term of $\hat{V}_{\text{JC}}$ can be estimated using $\hat{\pi}_a^{-1} \frac{1}{n_a - 1} \sum_{i:A_i=a} (Y_{a,i} - \hat{\gamma}_a^T (Z_i^T, \hat{\mu}_1(X_i), \ldots, \hat{\mu}_K(X_i))^T)^2$, which results in a variance covariance-matrix that is guaranteed to be positive definite even in finite samples.

B.5.4 Joint Calibration with Cross-Fitting

Corollary B.1. *The estimator $\hat{\theta}_{\mathrm{JC,a}}$ that uses the cross-fit $\check{\mu}_a(X_i) = \sum_{j=1}^{J} I\{i \in \mathcal{I}_j\}\hat{\mu}_{a,j}(X_i)$ as defined in Lemma B.12 rather than the standard $\hat{\mu}_a(X_i)$, has influence function given by*

$$\psi_a(A_i, X_i, Y_i) := \frac{I_{a,i}}{\pi_a}\{Y_{a,i} - \mu_a^*(X_i) - (\theta_a - P_0\mu_a^*(X_i))\} + \mu_a^*(X_i) - P_0\mu_a^*(X_i) \tag{B.13}$$

where $\mu_a^*(X_i) = \sum_{l=1}^{L} I(Z_i = z_l)\alpha_{a,l} + \beta_a^T\mu(X_i)$.

Proof. Let $\check{\mu} = (\check{\mu}_1, ..., \check{\mu}_K)$, Recall that with $\hat{\mu}_a^*(X_i) = \sum_{l=1}^{L} I(Z_i = z_l)\hat{\alpha}_{al} + \hat{\beta}_a^T\check{\mu}(X_i)$,

$$\hat{\theta}_{\mathrm{JC,a}} - \theta_a = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{I_{a,i}}{\hat{\pi}_a}\{Y_{a,i} - \hat{\mu}_a^*(X_i)\} + \hat{\mu}_a^*(X_i)\right] - \theta_a.$$

Then, assuming that $\hat{\alpha}_{a,l}$ and $\hat{\beta}_a$ are consistent for $\alpha_{a,l}$ and β_a respectively, we have

$$\begin{aligned}
\hat{\theta}_{\mathrm{JC,a}} - \theta_a &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{I_{a,i}}{\hat{\pi}_a}\left\{Y_{a,i} - \sum_{l=1}^{L} I(Z_i = z_l)\hat{\alpha}_{al} - \hat{\beta}_a^T\check{\mu}(X_i)\right\}\right. \\
&\qquad \left. + \sum_{l=1}^{L} I(Z_i = z_l)\hat{\alpha}_{al} + \hat{\beta}_a^T\check{\mu}(X_i)\right] - \theta_a \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{I_{a,i}}{\hat{\pi}_a}\left\{Y_{a,i} - \sum_{l=1}^{L} I(Z_i = z_l)\alpha_{al} - \beta_a^T\check{\mu}(X_i)\right\}\right. \\
&\qquad \left. + \sum_{l=1}^{L} I(Z_i = z_l)\alpha_{al} + \beta_a^T\check{\mu}(X_i)\right] - \theta_a \\
&\quad \underbrace{- \frac{1}{n}\sum_{i=1}^{n}\left[\frac{I_{a,i}}{\hat{\pi}_a}\left(\sum_{l=1}^{L} I(Z_i = z_l)(\hat{\alpha}_{al} - \alpha_{al}) + (\hat{\beta}_a - \beta_a)^T\check{\mu}(X_i)\right)\right]}_{R_1} \\
&\quad + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{l=1}^{L} I(Z_i = z_l)(\hat{\alpha}_{al} - \alpha_{al}) + (\hat{\beta}_a - \beta_a)^T\check{\mu}(X_i)\right]}_{R_2}.
\end{aligned}$$

First, it suffices to show that $R_1 - R_2$ is $o_P(1/\sqrt{n})$. Starting with the $(\hat{\alpha}_{al} - \alpha_{al})$ terms, we

have

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{I_{a,i}}{\hat{\pi}_a}-1\right]\left[\sum_{l=1}^{L}I(Z_i=z_l)(\hat{\alpha}_{al}-\alpha_{al})\right]=\sum_{l=1}^{L}\frac{1}{n}\sum_{i:Z_i=z_l}\left[\frac{I_{a,i}}{\hat{\pi}_a}-1\right](\hat{\alpha}_{al}-\alpha_{al}).$$

Therefore, we aim to show that in each level of l, we have

$$(\hat{\alpha}_{al}-\alpha_{al})\frac{1}{n}\sum_{i:Z_i=z_l}\left[\frac{I_{a,i}}{\hat{\pi}_a}-1\right]=o_P(1/\sqrt{n}).$$

Since $(\hat{\alpha}_{al}-\alpha_{al})=o_P(1)$, it suffices to show that $\frac{1}{n}\sum_{i:Z_i=z_l}\left[\frac{I_{a,i}}{\hat{\pi}_a}-1\right]=O_P(1/\sqrt{n})$:

$$\begin{aligned}
\frac{1}{n}\sum_{i:Z_i=z_l}\left[\frac{I_{a,i}}{\hat{\pi}_a}-1\right] &= (\hat{\pi}_a)\frac{1}{n}\sum_{i:Z_i=z_l}(I_{a,i}-\hat{\pi}_a)\\
&= (\hat{\pi}_a)(n(l)/n)\frac{1}{n(l)}\sum_{i:Z_i=z_l}(I_{a,i}-\hat{\pi}_a)\\
&= (\hat{\pi}_a)(n(l)/n)\left(n_a(l)/n(l)-\hat{\pi}_a\right)\\
&= (\hat{\pi}_a)(n(l)/n)\left([n_a(l)/n(l)-\pi_a]+[\pi_a-\hat{\pi}_a]\right)\\
&= (\hat{\pi}_a)(n(l)/n)O_P(1/\sqrt{n})\\
&= \pi_a P(Z_i=l)O_P(1/\sqrt{n})+o_P(1/\sqrt{n})\\
&= O_P(1/\sqrt{n}).
\end{aligned}$$

For the $(\hat{\beta}_a-\beta_a)^T$ term, we will use the results of previous lemmas and theorems. We have that the following vector can be re-written by adding and subtracting the scalar $(\bar{Y}_a-\theta_a)$:

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{I_{a,i}}{\hat{\pi}_a}\check{\mu}(X_i)-\check{\mu}(X_i)\right]=-\frac{1}{n}\sum_{i=1}^{n}\left[\frac{I_{a,i}}{\hat{\pi}_a}(Y_{a,i}-\check{\mu}(X_i))+\check{\mu}(X_i)-\theta_a\right]+(\bar{Y}_a-\theta_a).$$

The above term is a vector, i.e., with $(\bar{Y}_a-\theta_a)=(\bar{Y}_a-\theta_a)1_K$ and $Y_{a,i}=Y_{a,i}1_K$. From Theorem 2.1, we know that $(\bar{Y}_a-\theta_a)=O_P(1/\sqrt{n})$. Using the equivalence result in Lemma

B.12, and Remark B.2,

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{I_{a,i}}{\hat{\pi}_a}(Y_{a,i}-\check{\mu}(X_i))+\check{\mu}(X_i)-\theta_a\right]$$

is also $O_P(1/\sqrt{n})$ by Theorem 2.1. Although the vector $\check{\mu}$ is $(\check{\mu}_1,...,\check{\mu}_K)$ which may not be prediction unbiased for $j=1,...,K$, $j\neq a$, we still have the desired influence function because we are using the AIPW estimator, so we have prediction unbiasedness guaranteed. Therefore, we have

$$\begin{aligned}(\hat{\beta}_a-\beta_a)^T\frac{1}{n}\sum_{i=1}^{n}\left[\frac{I_{a,i}}{\hat{\pi}_a}\check{\mu}(X_i)-\check{\mu}(X_i)\right] &= (\hat{\beta}_a-\beta_a)^T O_P(1/\sqrt{n})+(\hat{\beta}_a-\beta_a)^T O_P(1/\sqrt{n})\\ &= o_P(1/\sqrt{n}).\end{aligned}$$

Finally, we have

$$\begin{aligned}\hat{\theta}_{\mathrm{JC,a}}=\frac{1}{n}\sum_{i=1}^{n}\Bigg[&\frac{I_{a,i}}{\hat{\pi}_a}\left\{Y_{a,i}-\sum_{l=1}^{L}I(Z_i=z_l)\alpha_{al}-\beta_a^T\check{\mu}(X_i)\right\}\\ &+\sum_{l=1}^{L}I(Z_i=z_l)\alpha_{al}+\beta_a^T\check{\mu}(X_i)\Bigg]+o_P(1/\sqrt{n}).\end{aligned}$$

We will show that $\check{\mu}_a^*(X_i):=\sum_{l=1}^{L}I(Z_i=z_l)\alpha_{al}+\beta_a^T\check{\mu}(X_i)$ satisfies Assumption 2.4 as long as $\hat{\mu}_{a,j}$ satisfies Assumption 2.4, making the above equivalent to using a different cross-fitting estimator. Specifically, we know that with probability $1-\Delta_n$, $\|\hat{\mu}_{a,j}-\mu_a\|_{L^2(P_0)}\leq\delta_n$ for all a treatment groups. Then also with probability $1-\Delta_n$,

$$\|\check{\mu}_a^*(X_i)-\mu_a^*\|_{L_2(P_0)}\leq\beta_a^T\|\check{\mu}-\mu\|\leq\beta_a^T 1_K\delta_n.$$

Since $\delta_n=o(1)$, $\beta_a^T 1_K\delta_n=o(1)$, so Assumption 2.4 is satisfied and $\hat{\theta}_{\mathrm{JC,a}}$ has the desired influence function. □

B.6 Augmenting Estimating Equations under CAR

The results in our Theorem 2.1 concern the vector of treatment means θ and the AIPW estimator $\hat{\theta}_{\text{AIPW}}$. The delta method can then be used to construct an estimate of an treatment contrast. A more general approach is that of Zhang et al. (2008), where the estimand $\delta \in \mathbb{R}^r$ is defined as the solution to the augmented estimating equation, m^*:

$$m^*(Y, A, X; \delta, f, \pi) = m(Y, A, \delta) - \sum_{a=1}^{K} \{I(A = a) - \pi_a\} f_a(X)$$

where $f = (f_1, \ldots, f_K)$ are arbitrary functions of X, and where $m(Y, A; \delta)$ is an unbiased estimating equation for δ_0, with $E(m(Y, A; \delta_0)) = 0$. The estimator $\hat{\delta}$ is the vector-valued solution to the empirical, augmented estimating equation, $\frac{1}{n}\sum_{i=1}^{n} m^*(Y_i, A_i, X_i; \delta, \hat{f}, \hat{\pi}) = 0$. This approach is very general as it allows for directly augmenting a treatment contrast using covariate adjustment, rather than augmenting the mean vector and then constructing a contrast (the AIPW estimator in (1) is a special case). In this section, we show that under Assumptions 2.1–2.3, with μ_a replaced by f_a, and $\hat{\mu}_a$ replaced by $\hat{f}_a$, and standard regularity conditions on the estimating function m, that the results in Theorem 2.1 apply to $\hat{\delta}$, with a different influence function and asymptotic variance given in the Corollary.

Corollary B.2 (Augmentation of Estimating Equations)**.** *Consider the augmented estimating equation m^*, with nuisance functions f:*

$$m^*(Y, A, X; \delta, f, \pi) = m(Y, A, \delta) - \sum_{a=1}^{K} \{I(A = a) - \pi_a\} f_a(X). \tag{B.14}$$

Define the estimand δ_0 as the solution to $P_0 m(\cdot; \delta) = 0$. Define the estimator of δ_0, $\hat{\delta}_n$, to be the solution to $P_n m^*(\cdot; \delta, \hat{f}, \hat{\pi}) = 0$, where $\hat{f} = (\hat{f}_1, \ldots, \hat{f}_K)$ each satisfy Assumptions 2.2 and 2.3, for some $f_0 = (f_1, \ldots, f_K)$, $\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_K)$, $\hat{\pi}_a = n_a/n$, and $\pi_0 = (\pi_1, \ldots, \pi_K)$. Under standard regularity conditions on m (see proof for details), Assumption 2.1, and Assumptions 2.2 and 2.3 on f_a:

$$\hat{\delta} - \delta_0 = \frac{1}{n}\sum_{i=1}^{n}\left[-\left(\nabla_\delta P_0 m\left(\cdot;\delta\right)|_{\delta=\delta_0}\right)^{-1}\right] m^*(\cdot;\delta_0, f_0 - P_0 f_0, \pi_0) + o_P(1/\sqrt{n}).$$

Furthermore, under condition (B) on the randomization scheme,

$$\begin{aligned}
\sqrt{n}(\hat{\delta} - \delta_0) &\xrightarrow{d} N(0, V_{\delta_0}) \\
V_{\delta_0} &= AV_\Psi A^T \\
A &= -\left(\nabla_\delta P_0 m\left(\cdot;\delta\right)|_{\delta=\delta_0}\right)^{-1}(\pi_1 I_r, \ldots, \pi_K I_r) \\
V_\Psi &= \operatorname{diag}\{\pi_a^{-1}\operatorname{Var}(m(Y(a);\delta_0) - \tilde{f}_a(X))\} - E(R(Z)\left\{\Omega_{\mathrm{SR}} - \Omega(Z)\right\}R(Z)^T) \\
&\quad - \operatorname{Var}(\tilde{f}(X)) + \operatorname{Cov}(m(Y;\delta_0), \tilde{f}(X)) + \operatorname{Cov}(\tilde{f}(X), m(Y;\delta_0)).
\end{aligned}$$

where I_r is the identity matrix of size $r \times r$, $\tilde{f}_a = f_a - P_0 f_a$ are the centered augmentation functions, and the vectors $f = (f_1^T, \ldots, f_K^T)^T$ and $m(Y;\delta_0) = (m(Y(1),1;\delta_0)^T, \ldots, m(Y(K),K;\delta_0)^T)^T$ are the stacked vectors of the augmentation functions and potential outcome estimating equations, respectively, and both are vector-valued with output dimension kr. The matrix $R(Z) = \operatorname{diag}\{R_1(Z), \ldots, R_k(Z)\}$, where $R_a(Z) = \pi_a^{-1}E(m(Y(a),a,\delta_0) - \tilde{f}_a(X)|Z) \in \mathbb{R}^r$.

Proof of Corollary B.2. We will first show consistency of the estimated $\hat{\delta}$ to δ_0 under CAR, then derive the influence function under CAR, and finally, asymptotic normality under CAR. Throughout, we consider the following regularity conditions on m:

(i) $P_0 m(\cdot;\delta)$ is continuously differentiable at δ_0, $P_0 m(\cdot;\delta)$ is continuous in δ, and that $\hat{\delta} \to \delta_0$ implies that $m(\hat{\delta}) - m(\delta_0) = o_P(1)$. This last condition implies that $P(m(\hat{\delta}) - m(\delta_0))^2 \to 0$ in probability when $\hat{\delta}$ is a consistent estimator for δ_0. See van der Vaart (1998) Section 5.3 for a discussion of this condition. An example of when this holds is when $\hat{\delta}$ is the MLE from a smooth parametric family.

(ii) Identifiability condition: for all $\epsilon > 0$, $\inf_\delta \{||P_0 m(\cdot;\delta)|| : ||\delta - \delta_0|| \geq \epsilon\} > 0$. That is,

the root of the estimating equation is unique.

(iii) $\sup_\delta ||(P_n - P_0)m(\cdot;\delta)|| \to 0$ in probability, i.e., m is a P_0 Glivenko-Cantelli class. This holds for M-estimators under CAR by the consistency result in the proof of Theorem 1 in Wang et al. (2021). Although they only consider two treatment groups, their consistency result can be extended to multiple treatment groups. Furthermore, even though they only consider stratified permuted block and biased coin randomization schemes in their Theorem 1, the proof of consistency only requires regularity conditions that are satisfied by our Assumption 2.2 (which also applies to Pocock and Simon's minimization).

(iv) The estimand δ_0 is invariant to the CAR scheme; that is, the solution to $P_0 m(\cdot;\delta)$ is the same regardless of the CAR scheme that is used to construct P_0. This generally holds; see Wang et al. (2021) Section 3.3 for a discussion of this condition.

Consistency of $\hat{\delta}$ to δ_0. We first prove consistency of $\hat{\delta}$ to δ_0, which will be used in deriving the influence function. We aim to show that for all $\epsilon > 0$, $P(||\hat{\delta} - \delta_0|| > \epsilon) \to 0$ as $n \to \infty$. Using (ii), there exists some ϵ^* such that:

$$\begin{aligned} P(||\hat{\delta} - \delta_0|| > \epsilon) &\leq P(||P_0 m(\cdot;\hat{\delta})|| > \epsilon^*) \\ &= P(||P_0 m^*(\cdot;\hat{\delta}, f_0, \pi_0) - P_0 m^*(\cdot;\delta_0, f_0, \pi_0)|| > \epsilon^*) \end{aligned}$$

since $P_0 m^*(\cdot;\delta, f, \pi_0) = P_0 m(\cdot;\delta)$ for any (δ, f) by the independence of A and X. Then, it suffices to show that $P_0 m^*(\cdot;\hat{\delta}, f_0, \pi_0) - P_0 m^*(\cdot;\delta_0, f_0, \pi_0) = o_P(1)$. For simplicity of notation, let $\eta = (\pi, f)$. We have the following decomposition:

$$\begin{aligned} &P_0 m^*(\cdot;\hat{\delta}, f_0, \pi_0) - P_0 m^*(\cdot;\delta_0, f_0, \pi_0) \\ &\quad = \underbrace{P_0 m^*(\cdot;\delta_0, \eta_0) - P_n m^*(\cdot;\delta_0, \hat{\eta})}_{(a)} + \underbrace{P_n m^*(\cdot;\delta_0, \hat{\eta}) - P_n m^*(\cdot;\hat{\delta}, \hat{\eta})}_{(b)} + \underbrace{P_n m^*(\cdot;\hat{\delta}, \hat{\eta}) - P_0 m^*(\cdot;\hat{\delta}, \eta_0)}_{(c)}. \end{aligned}$$

We can write $(a)+(c) = (P_n - P_0)(m(\cdot;\hat{\theta}) - m(\cdot;\theta))$, and by (iii), this is $o_P(1)$. Furthermore, since $P_n m^*(\cdot;\hat{\delta},\hat{\eta}) = 0$ and $P_0 m^*(\cdot;\delta_0,\eta_0) = 0$, (b) can be written as

$$\begin{aligned}(b) &= P_n\left[m^*(\cdot;\delta_0,\hat{\eta}) - m^*(\cdot;\delta_0,\eta_0)\right] + (P_n - P_0)m^*(\cdot;\delta_0,\eta_0) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\left\{I_{a,i} - \hat{\pi}_a\right\}\hat{f}_a(X_i) - \left\{I_{a,i} - \pi_a\right\}f_a(X_i)\right] + (P_n - P_0)m(\cdot;\delta_0)\end{aligned}$$

The term on the right is $o_P(1)$ by (iii). The term on the left is $o_P(1)$ by using Assumption 2.2 with an application of Markov's Inequality and Slutsky's Lemma, similar to term (II) in the proof in Supplemental Material Section B.5.1.

Asymptotic Linearity. First, define the new functions $\hat{h}_a = \hat{f}_a - P_0 f_a$ and $h_a = f_a - P_0 f_a$, as the estimated and limiting augmentation functions, both centered by the mean of the *limiting* augmentation function $P_0 f_a$. For any δ, $P_n m^*(\delta,\hat{f},\hat{\pi}) = P_n m^*(\delta,\hat{h},\hat{\pi})$ because $P_n\left\{I_{a,i} - \hat{\pi}_a\right\}P_0 h = 0$. we have the same thing for the population moment: $P_0 m^*(\delta,f,\pi_0) = P_0 m^*(\delta,h,\pi_0)$ because $P_0\left\{I_{a,i} - \pi_a\right\}P_0 h = 0$. In other words, we can center the augmentation function and not change the estimation equation. We work with those centered versions in the expansion below, and re-define $\eta = (\pi, h)$.

Based on the definitions of $\hat{\delta}$ and δ_0 as solving the empirical and population estimating equations, respectively, we have the following expansion:

$$0 = \underbrace{(P_n - P_0)m^*(\delta_0,\eta_0)}_{(I)} + \underbrace{P_0\left[m^*(\hat{\delta},\hat{\eta}) - m^*(\delta_0,\eta_0)\right]}_{(II)} + \underbrace{(P_n - P_0)\left[m^*(\hat{\delta},\hat{\eta}) - m^*(\delta_0,\eta_0)\right]}_{(III)}.$$

Term (I) is the linear term that will be part of the influence function. Term (II) can be

further broken down into two components:

$$(II) = \underbrace{P_0\left[m(\hat{\delta}) - m(\delta_0)\right]}_{(a)} - \underbrace{\sum_{a=1}^{K} P_0\left[\{I_{a,i} - \hat{\pi}_a\}\left\{\hat{f}_a(X_i) - P_0 f_a\right\} - \{I_{a,i} - \pi_a\}\{f_a(X_i) - P_0 f_a\}\right]}_{(b)}.$$

For term (a), we can do a first-order Taylor expansion, to write $(a) = (\hat{\delta} - \delta_0)\nabla_\delta P_0 m(\delta)|_{\delta=\tilde{\delta}}$ for some $\tilde{\delta} \in \left[\delta_0, \hat{\delta}\right]$. Re-arranging terms, and also using the continuous mapping theorem, condition (i) on $P_0 m$, and the consistency of $\hat{\delta} \to \delta_0$ (which implies $\tilde{\delta} \to \delta_0$), we have

$$\begin{aligned}(\hat{\delta} - \delta_0) = &\left(-\left[\nabla_\delta P_0 m(\delta)|_{\delta=\delta_0}\right]^{-1} + o_P(1)\right)(P_n - P_0)m^*(\delta_0, \eta_0) \\ &+ \left(-\left[\nabla_\delta P_0 m(\delta)|_{\delta=\delta_0}\right]^{-1} + o_P(1)\right)\{(III) - (b)\}.\end{aligned}$$

If we can show that $(P_n - P_0)m^*(\delta_0, \eta_0)$ is $O_p(1/\sqrt{n})$ (which we will do in the asymptotic normality section), and that (III) and (b) are both $o_p(1/\sqrt{n})$, we have our desired influence function. Let $\hat{g}_a(A_i) = I(A_i = a) - \hat{\pi}_a$, and $g_a(A_i) = I_{a,i} - \pi_a$. Term (b) can be rewritten to show that it is asymptotically negligible:

$$\begin{aligned}(b) = \sum_{a=1}^{K} P_0\left\{\hat{g}_a\hat{h}_a - g_a h_a\right\} &= \sum_{a=1}^{K} P_0\left\{\hat{g}_a\hat{h}_a - \hat{g}_a h_a + \hat{g}_a h_a - g_a h_a\right\} \\ \text{(Lemma B.1)} &= \sum_{a=1}^{K}(\hat{\pi}_a - \pi_a)P_0\left(\hat{h}_a - h_a\right) + (\hat{\pi}_a - \pi_a)P_0 h_a \\ \text{(Assumption 2.1)} &= \sum_{a=1}^{K} O_P(1/\sqrt{n})\left[P_0\left(\hat{f}_a - f_a\right)\right] \\ \text{(Assumption 2.2)} &= \sum_{a=1}^{K} o_P(1/\sqrt{n}) = o_P(1/\sqrt{n}).\end{aligned}$$

For term (III), we have

$$(III) = (P_n - P_0)\left[m(\hat{\delta}) - m(\delta_0)\right] + \sum_{a=1}^{K} \underbrace{(P_n - P_0)\left[\{I_{a,i} - \hat{\pi}_a\}\left\{\hat{f}_a(X_i) - P_0 f_a\right\}\right]}_{(d)}$$
$$- \sum_{a=1}^{K} \underbrace{(P_n - P_0)\left[\{I_{a,i} - \pi_a\}\{f_a(X_i) - P_0 f_a\}\right]}_{(e)}.$$

We focus first on the empirical process term for the un-augmented estimating equation m. By the consistency of $\hat{\delta} \to \delta_0$, and the continuity of $P_0 m$, $P_0(m(\hat{\delta}) - m(\delta_0))^2 \to 0$. Furthermore, m is a parametric function. So, Assumptions 2.2 and 2.3 are satisfied for functions $\hat{m} \equiv m(\hat{\delta})$ and $m_0 \equiv m(\delta_0)$.

Under simple randomization, $(P_n - P_0)\left[m(\hat{\delta}) - m(\delta_0)\right] = o_P(1/\sqrt{n})$ by an application of van der Vaart (1998) Lemma 19.24, since m is a parametric function of θ, and $A \perp Y$, so we have independent and identically distributed data. Under CAR, we can deal with the dependence in the way that we do in the proof of Theorem 2.1, when we have empirical process terms with $\hat{\mu}_a(X)$ and $\mu_a(X)$. The difference in this case is that $m(\delta) \equiv m(Y, A; \delta)$, and (Y, A) are not independent and identically distributed like the covariates X are marginally. Therefore, the results about $(P_n - P_0)(\hat{\mu}_a - \mu_a)$ do not immediately apply in this case. However, we can use the same trick as in the appendix of Wang et al. (2021) to rewrite the estimating equation as:

$$m(Y, A; \delta) = \sum_{a=1}^{K} I_{a,i} m(Y(a), a; \delta)$$

In other words, for a given a, $m(Y(a), a; \delta)$ is a function only of the potential outcomes $Y(a)$, which are independent and identically distributed. Define the functions $s_a : Y \to m(Y(a), a; \delta_0)$, and $\hat{s}_a : Y \to m(Y(a), a; \hat{\delta})$. Thus, to show that $(P_n - P_0)\left[m(\hat{\delta}) - m(\delta_0)\right]$, we only need to show that, for a given $a = 1, \ldots, K$, $(P_n - P_0)1_a(\hat{s}_a - s_a) = o_P(1/\sqrt{n})$, where $1_a : A \to I(A = a)$. We use the same decomposition as in the proof of Theorem 2.1

in the Supplemental Material B.2.2, with $P_0^n(\cdot) = \frac{1}{n}\sum_{i=1}^n P_0(\cdot|\mathcal{A}_n, \mathcal{Z}_n)$, where $\mathcal{A}_n = \{A_i; i = 1, \ldots, n\}$ and $\mathcal{Z}_n = \{Z_i; i = 1, \ldots, n\}$:

$$(P_n - P_0)1_a(\hat{s}_a - s_a) = (P_n - P_0^n)1_a(\hat{s}_a - s_a) + (P_0^n - P_0)1_a(\hat{s}_a - s_a)$$

The term with $(P_n - P_0^n)$ on the RHS is identical to term $\mathcal{M}_1$ in the proof of Theorem 2.1, with $\hat{\mu}_a$ replaced by $\hat{s}_a$ and μ_a replaced by s_a. Since Assumptions 2.2 and 2.3 are satisfied for $(\hat{m}, m_0)$, they are also satisfied for $(\hat{s}_a, s_a)$ for each a (i.e., as a function of δ, $m(Y(a), a; \delta)$ is parametric, and continuity tells us that $P_0(\hat{s}_a - s_a)^2 \to 0$ in probability). Therefore, Lemma B.11 applies, and $(P_n - P_0^n)1_a(\hat{s}_a - s_a) = o_P(1/\sqrt{n})$. For the second term, we can follow similar steps to term $\mathcal{M}_2$ in the proof of Theorem 2.1 (also applying the independence of A and $Y(a)$):

$$\begin{aligned}
(P_0^n - P_0)1_a(\hat{s}_a - s_a) &= \frac{1}{n}\sum_{i=1}^n I_{a,i} P_0(\hat{s}_a - s_a|\mathcal{A}_n, \mathcal{Z}_n) - \pi_a P_0(\hat{s}_a - s_a) \\
&= \frac{1}{n}\sum_{i=1}^n \sum_{l=1}^L [I_{a,i} I(Z_i = l) P_0(\hat{s}_a - s_a|Z_i = l) - \pi_a P(Z = l) P(\hat{s}_a - s_a|Z = l)] \\
&= \sum_{l=1}^L \left\{ \frac{n_a(l)}{n} - \pi_a P(Z = l) \right\} P_0(\hat{s}_a - s_a|Z = l) \\
&= o_P(1) \sum_{l=1}^L \left\{ \frac{n_a(l)}{n} - \pi_a P(Z = l) \right\} \\
&= o_P(1) \left\{ n_a/n - \pi_a \right\} = o_P(1/\sqrt{n})
\end{aligned}$$

where we have used the fact that $P_0(\hat{s}_a - s_a|Z = l) = o_P(1)$ for each l as we have in term $\mathcal{M}_2$ in the proof of Theorem 2.1.

Finally, we show that for a given a, $(d) - (e) = o_P(1/\sqrt{n})$. Within term (e), we can apply Lemma B.1 to show that $P_0\left[\{I_{a,i} - \pi_a\}\{f_a(X_i) - P_0 f_a\}\right] = 0$. Next, in term (d), the $P_0 f_a$

can be replaced with $P_0\hat{f}_a$ because

$$(P_n - P_0)\{I_{a,i} - \hat{\pi}_a\}\left\{P_0\hat{f}_a - P_0 f_a\right\} = P_0\{I_{a,i} - \hat{\pi}_a\}P_0(\hat{f}_a - f_a)$$
$$= (\pi_a - \hat{\pi}_a)o_P(1) = o_P(1/\sqrt{n}).$$

Similarly to term (e), in term (d), $P_0\left[\{I_{a,i} - \hat{\pi}_a\}\left\{\hat{f}_a(X_i) - P_0\hat{f}_a\right\}\right] = 0$. Thus, we are left with:

$$(d) - (e) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} I_{a,i}\left\{\hat{f}_a(X_i) - f_a(X_i) - P_0\hat{f}_a + P_0 f_a\right\}}_{\mathcal{M}}$$
$$+ \{\pi_a + o_P(1)\}(P_n - P_0)(\hat{f}_a - f_a).$$

We recognize $(d) - (e)$ as the same as the term (I) (multiplied by a constant) from our proof of Theorem 2.1 (i) in Section B.2.2, with $\hat{\mu}_a$ replaced by $\hat{f}_a$ and μ_a replaced by f_a. Therefore, the whole term (III) is $o_P(1/\sqrt{n})$, as we showed in the proof of Theorem 2.1.

Asymptotic Normality. Recall that $\delta \in \mathbb{R}^r$. Then using the same trick of breaking up the estimating equation into its potential outcome components as we did earlier (as in Wang et al. (2021)), the influence function that we derived can be written equivalently as, with $C = -\left(\nabla_\delta P_0 m\left(\cdot;\delta\right)|_{\delta=\delta_0}\right)^{-1}$,

$$\psi(Y_i, A_i, X_i) = C\sum_{a=1}^{K}\pi_a \underbrace{\left[\frac{I_{a,i}}{\pi_a}\{m(Y(a), a; \delta_0) - f_a(X) + P_0 f_a\} - f_a(X) + P_0 f_a(X)\right]}_{\psi_a(\cdot)}.$$

The function ψ_a is identical to the influence function Theorem 2.1, but with μ_a replaced by f_a, and with the derived outcome $m(Y(a), a; \delta_0)$ replacing $Y(a)$, and where f_a and $m(Y(a), a; \delta_0)$ are vector-valued in $\mathbb{R}^r$, rather than scalar functions. This means that $\psi_a \in \mathbb{R}^r$. Let $\Psi(\cdot) = (\psi_1^T(\cdot), \ldots, \psi_K^T(\cdot))^T$. We can follow the proof of Theorem 2.1 in the Supplemental Materials B.3.1 (Asymptotic Variance Decomposition), but in vector form, and converted to the new

derived outcome. Let $\tilde{f}_a = f_a - P_0 f_a$, and let $f = (f_1^T, \ldots, f_K^T)^T \in \mathbb{R}^{K \cdot r}$ be the stacked vectors of augmentation functions. Let $m(Y; \delta_0) = (m(Y(1), 1; \delta_0)^T, \ldots, m(Y(K), K; \delta_0)^T)^T$ be the stacked vectors of potential outcome estimating equations. Define

$$\epsilon_{a,\delta_0}(x) : x \to m(Y(a), a, \delta_0) - \tilde{f}_a(x).$$

Let $R_a(Z) = \pi_a^{-1} E(\epsilon_{a,\delta_0}(X)|Z) \in \mathbb{R}^r$, and $R(Z) = \text{diag}\{R_1(Z), \ldots, R_k(Z)\}$, which is a block-diagonal matrix where each block is of dimension $r \times 1$. These are equivalent to the univariate case of $R_a(Z)$ and $R(Z)$ when $r = 1$ for the AIPW estimator. Likewise, let $Q_a(Z) = \pi_a^{-1} \text{Var}(r_a(X)|Z)$, and $Q(Z) = \text{diag}\{Q_1(Z), \ldots, Q_K(Z)\}$, which is also block-diagonal but where each block has dimension $r \times r$.

By the proof of Theorem 2.1, $\sqrt{n}(P_n - P_0)\Psi \Rightarrow N(0, V_\Psi)$ where:

$$V_\Psi = V^{(1,4)} + V^{(3,5)} + V^{(2)}$$

$$\begin{aligned} V^{(1,4)} = {} & E(Q(Z)) + E(\text{Var}(\tilde{f}(X)|Z)) \\ & + E(\text{Cov}(m(Y; \delta_0) - \tilde{f}(X), \tilde{f}(X)|Z)) + E(\text{Cov}(\tilde{f}(X), m(Y; \delta_0) - \tilde{f}(X)|Z)) \end{aligned}$$

$$\begin{aligned} V^{(3,5)} = {} & E(R(Z)(\pi\pi^T)R(Z)^T) + \text{Var}(E(\tilde{f}(X)|Z)) \\ & + \text{Cov}(E(m(Y; \delta_0) - \tilde{f}(X)|Z), E(\tilde{f}(X)|Z)) + \text{Cov}(E(\tilde{f}(X)|Z), E(m(Y; \delta_0) - \tilde{f}(X)|Z)) \end{aligned}$$

$$V^{(2)} = E(R(Z)\Omega(Z)R(Z)^T)$$

We derive $V^{(2)}$ by following the same steps as in Section B.3 of the Supplemental Material; the only difference between this result and that in A.5 is that $R(Z)$ is not a symmetric matrix so we explicitly note its transpose. A similar difference is noted for $V^{(3,5)}$. By the

same arguments in Section B.3, we have the simplified form of

$$\begin{aligned} V_\Psi &= E(Q(Z)) + E(R(Z)\Omega(Z)R^T(Z)) + E(R(Z)\pi\pi^T R(Z)^T) \\ &\quad - \mathrm{Var}(\tilde{f}) + \mathrm{Cov}(m(Y;\delta_0), \tilde{f}) + \mathrm{Cov}(\tilde{f}, m(Y;\delta_0)) \\ &= \mathrm{diag}\{\pi_a^{-1}\mathrm{Var}(m(Y(a);\delta_0) - \tilde{f}_a(X))\} - E(R(Z)\left\{\Omega_{\mathrm{SR}} - \Omega(Z)\right\}R(Z)^T) \\ &\quad - \mathrm{Var}(\tilde{f}(X)) + \mathrm{Cov}(m(Y;\delta_0), \tilde{f}(X)) + \mathrm{Cov}(\tilde{f}(X), m(Y;\delta_0)). \end{aligned}$$

The key difference between this result and the asymptotic normality result in Theorem 2.1 is that the dimension of the variance matrix is $(rK \times rK)$, rather than $(K \times K)$. Ultimately, we are interested in a linear combination of the elements of Ψ where we sum across the different treatment groups. Our influence function is given by $\psi(\cdot) = C\sum_{a=1}^{K}\pi_a\psi_a(\cdot)$. This can be expressed as a linear transformation with matrix $A = -\left(\nabla_\delta P_0 m\left(\cdot;\delta\right)|_{\delta=\delta_0}\right)^{-1}\left(\pi_1 I_r, \ldots, \pi_K I_r\right)$ of the influence vector $\Psi(\cdot)$, i.e., $\psi(\cdot) = A\Psi(\cdot)$. Then,

$$\sqrt{n}\left(\hat{\delta} - \delta_0\right) \Rightarrow N(0, AV_\Psi A^T).$$

Robust Variance Estimation. To construct a variance estimator that is robust to model mis-specification, we can use the same approach as in Section 2.5 of the main text. A robust variance estimator for δ_0 is $\hat{A}\hat{V}_\Psi\hat{A}^T$ where:

- $\hat{A} = -\left(\nabla_\delta P_n m(\cdot;\delta)|_{\delta=\hat{\delta}}\right)^{-1}\left(\hat{\pi}_1 I_r, \ldots, \hat{\pi}_K I_r\right)$. We can use $P_n m(\cdot;\delta)$ for consistent estimation of $P_0 m(\cdot;\delta)$ for any δ by regularity condition (iii); then we can plug in $\hat{\delta}$ for δ_0 by the continuous mapping theorem.

- $\hat{V}_\psi$ is consistently estimated by the following:

$$\begin{aligned} \hat{V}_\Psi &= \mathrm{diag}\left[\hat{\pi}_a^{-1}(S_a^2 - 2\hat{Q}_{aa} + \hat{\Sigma}_{aa}),\, a = 1, ..., K\right] + \hat{Q} + \hat{Q}^T - \hat{\Sigma} \\ &\quad - \sum_z \frac{n(z)}{n}\hat{R}(z)\{\Omega_{\mathrm{SR}} - \Omega(z)\}\hat{R}(z), \end{aligned}$$

where:

- S_a^2 is the sample covariance matrix of $m(Y_{a,i}; \hat{\delta})$ for individuals in treatment arm a.
- $\hat{Q}$ is the $rK \times rK$ matrix with the (a, b)th component $r \times r$ block given by the sample covariance matrix of $m(Y_{a,i}; \hat{\delta})$ and $\hat{f}_b(X_i)$ (both vector-valued functions in $\mathbb{R}^r$), and $\hat{Q}_{a,a}$ is its (a, a)th $r \times r$ block component.
- $\hat{\Sigma}$ is the $rK \times rK$ sample covariance matrix of $\hat{f}(X_i)$ based on the entire sample.
- $\hat{R}(z)$ is the diagonal matrix with block-diagonal components $\hat{R}_a(z)$ where

$$\hat{R}_a(z) = \hat{\pi}_a^{-1} \bar{\epsilon}_{a,\hat{\delta}}(z)$$

  where $\bar{\epsilon}_{a,\delta}(z)$ is the sample mean of $\epsilon_{a,\delta}(X_i)$ for individuals within group a and strata z.

□

Remark B.4. *(Equivalence of Corollary B.2 and Theorem 2.1). If we set $m(Y, A, \delta)$ to be a vector of length k, with the ath entry $I_{a,i}\,(Y_i - \delta_a)$, then $\delta_a = E(Y_a)$ is the solution to the population estimating equation and $\hat{\delta}_a = \bar{Y}_a$ is the solution to the empirical estimating equation. If we also set $\hat{f}_a(X)$ to be the vector with ath entry $\hat{\mu}_a(X)$ and 0's elsewhere, then $\hat{\delta}_a$ as the solution to the empirical estimating equation $\sum_{i=1}^n m^*(Y_i, A_i, X_i, \delta, \hat{f}, \hat{\pi}) = 0$ is exactly the AIPW estimator. Furthermore, the variance in Corollary B.2 agrees with the variance in Theorem 2.1. This is apparent by noting that $A = \text{diag}\{1/\pi_a; a = 1, \ldots, K\}(\pi_1 I_K, \ldots, \pi_K I_K)$. This results in substantial cancellation when applied to the matrices in V_Ψ, which are sparse: the function vectors m and $\tilde{f}$ each are of length K^2 but only have K non-zero entries. We are left with the variance in Theorem 2.1.*

B.7 Full Simulation Results

In Tables B.1 and B.2, we include the full results from our simulation study. This table is identical to Table 1 in the main text, but with additional columns that show the average SE, coverage, and power to reject the null hypothesis of no treatment effect, using naive SE. It also includes rows for non-stratified random forest working models.

Table B.1: Simulation results for estimating $\theta_2 - \theta_1 = 0.123$ under case I based on $n = 500$ and 5,000 replicates. Entries are multiplied by 100. The naive SE and CP ignore CAR, with empty entries under simple randomization, oracle, or JC. Under minimization, some entries are marked with "–" because the correct SE and CP are not available.

		Working				Correct			Naive		
	CAR	model	Method	Bias	SD	SE	CP	P	SE	CP	P
Case I	SR		Mean	0.13	4.49	4.44	94.64	78.39			
		Oracle	AIPW	0.04	3.82	3.75	94.60	89.61			
		Logistic	AIPW	0.04	3.96	3.89	94.56	87.81			
			LC	0.01	3.91	3.84	94.60	88.05			
			JC	-0.00	3.87	3.77	94.34	88.97			
		Random	CF	0.09	4.06	4.02	94.74	85.99			
		forest	JC	0.05	3.87	3.81	94.60	89.01			
		Stratified	CF	0.06	3.92	3.90	94.64	88.11			
		random forest	JC	0.05	3.90	3.87	94.86	88.35			
	PB		Mean	-0.01	4.01	3.98	95.14	85.87	4.44	97.12	81.23
		Oracle	AIPW	0.00	3.79	3.75	94.90	89.53			
		Logistic	AIPW	-0.02	3.89	3.88	94.94	87.67	3.89	94.98	87.57
			LC	-0.03	3.83	3.81	94.92	88.63	3.83	95.12	88.39
			JC	-0.05	3.82	3.76	94.64	89.15			
		Random	CF	-0.01	3.87	4.02	96.00	86.25	4.02	95.94	86.35
		forest	JC	0.01	3.83	3.80	94.80	88.83			
		Stratified	CF	-0.00	3.90	3.89	95.08	87.21	3.90	95.06	87.31
		random forest	JC	0.00	3.88	3.86	94.86	87.53			
	Mini		Mean	0.18	4.00	–	–	–	4.44	96.78	82.88
		Oracle	AIPW	0.16	3.74	3.75	95.04	90.42			
		Logistic	AIPW	0.16	3.88	–	–	–	3.88	94.96	88.74
			LC	0.13	3.82	–	–	–	3.83	95.20	89.44
			JC	0.11	3.79	3.76	94.98	89.88			
		Random	CF	0.17	3.85	–	–	–	4.01	95.88	87.60
		forest	JC	0.17	3.79	3.80	95.08	89.76			
		Stratified	CF	0.17	3.86	–	–	–	3.89	95.12	88.78
		random forest	JC	0.17	3.83	3.86	95.08	89.26			

Table B.2: Simulation results for estimating $\theta_2 - \theta_1 = 0.132$ under case II, based on $n = 500$ and 5,000 replicates. Entries are multiplied by 100. The naive SE and CP ignore CAR, with empty entries under simple randomization, oracle, or JC. Under minimization, some entries are marked with "–" because the correct SE and CP are not available.

Case II	SR		Mean	-0.10	4.68	4.65	94.64	81.06		
		Oracle	AIPW	-0.05	3.74	3.53	93.88	95.55		
		Logistic	AIPW	-0.07	4.65	4.57	94.44	82.48		
			LC	-0.07	4.66	4.57	94.32	82.46		
			JC	-0.08	4.50	4.35	93.94	84.85		
		Random	CF	-0.08	4.11	4.08	94.88	89.43		
		forest	JC	-0.08	4.05	4.00	94.72	90.55		
		Stratified	CF	-0.12	4.05	4.01	94.74	90.45		
		random forest	JC	-0.11	4.05	3.99	94.72	90.65		
	PB		Mean	0.09	4.41	4.41	94.85	85.79	4.65	95.98	82.95
		Oracle	AIPW	0.04	3.67	3.53	93.71	95.90		
		Logistic	AIPW	0.09	4.40	4.35	94.63	86.13	4.57	95.62	84.20
			LC	0.08	4.40	4.35	94.57	86.23	4.56	95.58	84.20
			JC	0.10	4.40	4.34	94.59	86.47			
		Random	CF	0.07	3.99	4.08	95.56	90.99	4.08	95.50	91.12
		forest	JC	0.05	3.98	3.99	95.24	91.72			
		Stratified	CF	0.08	3.97	4.00	95.08	92.04	4.00	95.26	91.86
		random forest	JC	0.08	3.97	3.99	95.20	92.02			
	Mini		Mean	-0.00	4.44	–	–	–	4.64	95.93	82.91
		Oracle	AIPW	-0.03	3.69	3.53	93.78	95.69		
		Logistic	AIPW	0.02	4.41	–	–	–	4.56	95.67	84.02
			LC	0.02	4.42	–	–	–	4.55	95.57	84.04
			JC	0.03	4.42	4.34	94.38	85.88			
		Random	CF	-0.01	4.00	–	–	–	4.08	95.09	90.75
		forest	JC	-0.01	3.98	3.99	94.77	91.56			
		Stratified	CF	-0.04	3.98	–	–	–	4.00	95.01	91.26
		random forest	JC	-0.04	3.98	3.98	94.89	91.50			

B.8 Considerations for Finite Sample Performance of Variance Estimators

There are several reasonable ways to construct a robust variance estimator, and each strategy will have different finite sample performance. The estimator listed in Section 2.5 is given as follows. Let S_a^2 be the sample variance of $Y_{a,i}$ for patients in treatment arm a, $\hat{Q}$ be the $K \times K$ matrix whose (a,b)th component is the sample covariance of $Y_{a,i}$ and $\hat{\mu}_b(X_i)$ for patients in arm a, $\hat{\Sigma}$ be the sample covariance matrix of $(\hat{\mu}_1(X_i), \ldots, \hat{\mu}_K(X_i))^T$ based on the entire sample, $\hat{Q}_{aa}$ and $\hat{\Sigma}_{aa}$ be respectively the (a,a)th component of $\hat{Q}$ and $\hat{\Sigma}$, $\hat{R}_Y(z) = \text{diag}\big[\hat{\pi}_a^{-1}\{\bar{Y}_a(z) - \hat{\theta}_a\}, a = 1, \ldots, K\big]$, $\hat{R}_X(z) = \text{diag}\big[\hat{\pi}_a^{-1}\{\bar{\mu}_a(z) - \bar{\mu}_a\}, a = 1, \ldots, K\big]$, $\bar{Y}_a(z)$ be the sample mean of $Y_{a,i}$ for patients in treatment arm a and stratum z, $\hat{\theta}_a$ be the ath component in $\hat{\theta}_{\text{AIPW}}$ (or the cross-fitted version), $\bar{\mu}_a(z)$ be the sample mean of $\hat{\mu}_a(X_i)$'s for patients in stratum z, and $\bar{\mu}_a$ be the sample mean of $\hat{\mu}_a(X_i)$'s for all patients. Then, V can be consistently estimated by

$$\hat{V}_1 = \text{diag}\left[\hat{\pi}_a^{-1}(S_a^2 - 2\hat{Q}_{aa} + \hat{\Sigma}_{aa}),\, a = 1, ..., K\right] + \hat{Q} + \hat{Q}^T - \hat{\Sigma} - \hat{f}(\Omega)$$

where $\hat{f}(\Omega) = \sum_z \frac{n(z)}{n}\{\hat{R}_Y(z) - \hat{R}_X(z)\}\{\Omega_{\text{SR}} - \Omega(z)\}\{\hat{R}_Y(z) - \hat{R}_X(z)\}$. This is the variance estimator that we use in the simulation study and case study. Alternative variance estimators can be considered by modifying the non-Ω terms of the variance. Rather than deconstruct the $\text{Var}(Y_a - \mu_a(X))$ in the $\text{diag}[\pi_a^{-1}\text{Var}(Y_a - \mu_a(X)), a = 1, \ldots, K]$ of the variance into $\text{Var}(Y_a) + \text{Var}(\mu_a(X)) - 2\text{Cov}(Y_a, \mu_a(X))$ as we do in $\hat{V}_1$, we could instead estimate the variance of this term directly using patients in treatment arm a only, where $\hat{D}_a = \hat{\text{Var}}(Y_{i,a} - \mu_a(X_i))$, and have:

$$\hat{V}_2 = \text{diag}\left[\hat{\pi}_a^{-1}\hat{D}_a,\, a = 1, ..., K\right] + \hat{Q} + \hat{Q}^T - \hat{\Sigma} - \hat{f}(\Omega)$$

There are pros and cons to $\hat{V}_1$ and $\hat{V}_2$. For $\hat{V}_1$, its performance may be affected if the distribution of covariates X_i differs between those with $A_i = a$ and $A_i \neq a$. In $\hat{V}_2$, only those with $A_i = a$ are used, whereas in $\hat{V}_1$, the entire sample is used to estimate the term $\hat{\Sigma}$. In

$\hat{V}_2$, however, if there are many zeroes in $Y_{i,a}$, then this estimator can be unstable. A third estimator that may be suitable and handle some of this instability is:

$$\hat{V}_3 = \text{diag}\left[\hat{\pi}_a^{-1}\bar{\sigma}_a,\, a = 1, ..., K\right] + \hat{\Sigma}_2 - \hat{f}(\Omega)$$

where $\bar{\sigma}_a$ is the sample mean of $[Y_{i,a} - \hat{\mu}_a(X_i)]^2$ among those with $A_i = a$, and $\hat{\Sigma}_2$ is the matrix with (a, a) entries of:

$$2\widehat{\text{Cov}}_a(Y_{i,a} - \hat{\mu}_a(X_i)) + \widehat{\text{Var}}_a(\hat{\mu}_a(X_i))$$

and (a, b), $a \neq b$, entries of:

$$\widehat{\text{Cov}}_a(Y_{i,a}, \hat{\mu}_b(X_i)) + \widehat{\text{Cov}}_b(Y_{i,b}, \hat{\mu}_a(X_i)) - \frac{1}{2}\left[\widehat{\text{Cov}}_a(\hat{\mu}_a(X_i), \hat{\mu}_b(X_i)) + \widehat{\text{Cov}}_b(\hat{\mu}_a(X_i), \hat{\mu}_b(X_i))\right]$$

where $\widehat{\text{Cov}}_a$ is the sample covariance among patients in arm a, and $\widehat{\text{Var}}_a$ is the sample variance among patients in arm a. This estimator focuses on using participants in arms a and b only for the cross-terms, rather than using only those in a, or the entire sample.

Appendix C

SUPPLEMENTAL MATERIAL FOR CHAPTER 3

C.1 Preliminaries

C.1.1 Notation

Here we define some notation that we will use throughout:

- Let $I_{\mathcal{J}}(Z) = I(\pi_k(Z) > 0, k \in \mathcal{J})$ be an indicator of membership to the ECE population. Let $p_{\mathcal{J}} = E(I_{\mathcal{J}}(Z))$ be the probability of this indicator function being 1.

- Let $E_{\mathcal{J}}(X) = E(X|I_{\mathcal{J}}(Z) = 1)$ be the expectation over the ECE population. Sometimes we will also use the similar notation $P_{\mathcal{J}}$ when working with probabilities.

- Let $\mathbb{E}_{\mathcal{J}}$ be the empirical counterpart to $E_{\mathcal{J}}$. That is, for a random variable X_i,

$$\mathbb{E}_{\mathcal{J}}(X_i) = n(\mathcal{I}_{\mathcal{J}})^{-1} \sum_{i \in I_{\mathcal{J}}} X_i.$$

  We will use the notation $\mathbb{P}_{\mathcal{J}}$ when referring to the empirical probability measure that corresponds to the expectation above. Importantly, $I_{\mathcal{J}}(Z_i)$ are independent and identically distributed (iid) for $Z_i = 1, \ldots, n$; therefore, the empirical measure $\mathbb{P}_{\mathcal{J}}$ is using iid samples, and we assume that $n(\mathcal{I}_{\mathcal{J}})$ in our asymptotic regime. As such, standard empirical process results using $\mathbb{P}_{\mathcal{J}}$ and its population counterpart $P_{\mathcal{J}}$ apply in this setting. For short-hand, we use the notation $\epsilon = o_P(1/\sqrt{n^*})$ to convey the property that $\epsilon \cdot \sqrt{n(\mathcal{I}_{\mathcal{J}})} \to 0$ in probability as $n(\mathcal{I}_{\mathcal{J}}) \to \infty$.

- Let $\Lambda_j(s) = P_{\mathcal{J}} N_j(s) / P_{\mathcal{J}} R_j(s)$, where $N_j(s) = I(Y^{(j)} \leq t, \delta^{(j)} = 1)$ is the event counting process, and $R_j(s) = I(Y^{(j)} \geq s)$ be the at-risk process.

- We will use the short-hand notation $p_j^C(s|V)$ to denote the probability of remaining uncensored in arm j within the ECE population, as a function of time, i.e., $p_j^C(s|V) \equiv P_{\mathcal{J}}(C^{(j)} \geq s|V)$.

C.1.2 Identification of the Cumulative Hazard

Lemma C.1. *Under Assumptions 3.1 and 3.2, we have the following result for any function g of the potential outcomes $(T^{(i)}, C^{(j)})$:*

$$E(I_{\mathcal{J}}(Z)I(A=j)\pi_j(Z)^{-1}p_j^C(s|V)^{-1}g(\cdot)) = E(I_{\mathcal{J}}(Z)p_j^C(s|V)^{-1}g(\cdot)).$$

This result allows us to remove the treatment propensity weights from the expectation, given the relationships assumed in Assumption 3.2 regarding the censoring and treatment propensity variables.

Proof.

$$\begin{aligned} E(I(A=j)\pi_j(Z)^{-1}p_j^C(s|V)^{-1}g(\cdot)I_{\mathcal{J}}(Z)) &= E(\pi_j(Z)^{-1}I_{\mathcal{J}}(Z)E(I(A=j)g(\cdot)p_j^C(s|V)|Z)) \\ \text{(Assumption 3.2)} &= E(\pi_j(Z)^{-1}I_{\mathcal{J}}(Z)P(A=j|Z)E(g(\cdot)(p_j^C)^{-1}(V)|Z)) \\ &= E(p_j^C(s|V)^{-1}g(\cdot)I_{\mathcal{J}}(Z)). \end{aligned}$$

□

Lemma C.2 (Identification)**.** *Let $dL_j(s) = N_j(s) - R_j(s)d\Lambda_j(s)$. We have the following identification result:*

$$P_{\mathcal{J}}f_j(s)dL_j(s) = 0$$

where $f_j(W)(t) = I(A=j)\pi_j(Z)^{-1}P_{\mathcal{J}}(C^{(j)} \geq t|V)^{-1}$ contains the double weight based on treatment propensity and censoring processes.

Proof. It suffices to show that $E(f_j(s)I_{\mathcal{J}}(Z)dL_j(s)) = 0$, since dividing by $E(I_{\mathcal{J}}(Z))$ results in $P_{\mathcal{J}}f_j(s)dL_j(s)$. By applying Lemma C.1, we have that:

$$E(f_j(s)I_{\mathcal{J}}(Z)dL_j(s)) = E(P_{\mathcal{J}}(C^{(j)} \geq s|V)^{-1}I_{\mathcal{J}}(Z)dL_j(s))$$

Recall that $dL_j(s) = dN_j(s) - R_j(s)d\Lambda_j(s)$. We have that:

$$\begin{aligned}
E(P_{\mathcal{J}}(C^{(j)} \geq s|V)^{-1} I_{\mathcal{J}}(Z) R_j(s)) &= E(P_{\mathcal{J}}(C^{(j)} \geq s|V)^{-1} I_{\mathcal{J}}(Z) R_j(s)) \\
&= E\left\{ \frac{E\left\{ I(Y^{(j)} \geq s) \middle| I_{\mathcal{J}}(Z), V \right\} E(I_{\mathcal{J}}(Z)|V)}{P(C^{(j)} \geq s|I_{\mathcal{J}}(Z), V)} \right\} \\
&= E\left\{ \frac{P(\min(T^{(j)}, C^{(j)}) \geq s|I_{\mathcal{J}}(Z), V) E(I_{\mathcal{J}}(Z)|V)}{P(C^{(j)} \geq s|I_{\mathcal{J}}(Z), V)} \right\} \\
&= E\left\{ P(T^{(j)} \geq s|C^{(j)} \geq s, I_{\mathcal{J}}(Z), V) E(I_{\mathcal{J}}(Z)|V) \right\} \\
&= E\left\{ P(T^{(j)} \geq s|I_{\mathcal{J}}(Z), V) E(I_{\mathcal{J}}(Z)|V) \right\} \\
&= E(T^{(j)} \geq s, I_{\mathcal{J}}(Z)) \\
&= P_{\mathcal{J}}(T^{(j)} \geq s) p_{\mathcal{J}}.
\end{aligned}$$

Since the denominator of $d\Lambda_j(s)$ is $P_{\mathcal{J}}(T^{(j)} \geq s)$, we are left with showing that

$$E(P_{\mathcal{J}}(C^{(j)} \geq s|V)^{-1} I_{\mathcal{J}}(Z) dN_j(s)) = P_{\mathcal{J}}(dN_j(s)) p_{\mathcal{J}}.$$

This holds using the same logic as the steps above for $R_j(s)$.

□

C.1.3 Functional Delta Method for a Cumulative Hazard

Lemma C.3 (Application of the Functional Delta Method)**.** *Let $\hat{\Lambda}_j(t)$ be an asymptotically linear estimator of the cumulative hazard function $\Lambda_j(t)$ such that $\hat{\Lambda}_j(t) - \Lambda_j(t) = \mathbb{P}_{\mathcal{J}} \psi_j(t) + o_P(1/\sqrt{n^*})$, and such that $\sqrt{n(\mathcal{I}_{\mathcal{J}})}(\hat{\Lambda}_j(t) - \Lambda_j(t))$ weakly converges to a mean-zero Gaussian process. Define $\hat{S}_j(t) = \prod_{0<t' \leq t} \left\{ 1 - \Delta\hat{\Lambda}(t') \right\}$. Then,*

$$\hat{S}_j(t) - S_j(t) = \mathbb{P}_{\mathcal{J}} S_j(t) \int_0^t \frac{d\psi_j(u)}{(1 - \Delta\Lambda_j(u))} + o_P(1/\sqrt{n^*})$$

and $\sqrt{n(\mathcal{I}_{\mathcal{J}})}(\hat{S}_j(t) - S_j(t))$ *converges weakly to a mean zero Gaussian process with variance*

$$V_j(t) = \{S_j(t)\}^2 \operatorname{Var}\left[\int_0^t \frac{d\psi_j(u)}{(1 - \Delta\Lambda_j(u))}\right].$$

Proof. We have the following relationship between the survival function $S_j(t)$ and the cumulative hazard function $\Lambda_j(t)$ outlined in Kosorok (2008) (see discussion following Lemma 12.5) and Slavk (2007) (see Section 1.4): $S_j(t) \equiv \exp(-\Lambda_j(t)) = \prod_{0<t'\leq t}\{1 - \Delta\Lambda_j(t')\}$. Therefore, with the product integral defined as $\Psi(\Lambda)(t) = \prod_{0<t'\leq t}\{1 - \Delta\Lambda(t')\}$, $\hat{S}_j = \Psi(\hat{\Lambda}_j)$ and $S_j = \Psi(\Lambda_j)$, where $\Delta\Lambda(t) = \Lambda(t) - \Lambda(t-)$ where $\Lambda(t-)$ is the limit of $\Lambda(t)$ approaching t from the left. The map Ψ is Hadamard differentiable with derivative $\Psi'_\Lambda(\alpha)(t) = \Psi(\Lambda)(t)\int_0^t d\alpha(u)\{1 - \Delta\Lambda_j(s)\}$. Note that by definition, the functional derivatives are linear. With f continuous, $\Psi'_\Lambda(f)$ is continuous. Let T be the mean-zero Gaussian distribution such that $\sqrt{n(\mathcal{I}_{\mathcal{J}})}(\hat{\Lambda}_j(t) - \Lambda_j(t)) \rightsquigarrow T$. Then, applying the Hadamard differentiability result above, $\sqrt{n(\mathcal{I}_{\mathcal{J}})}(\Psi(\hat{\Lambda}_j) - \Psi(\Lambda_j)) \rightsquigarrow \Psi'_\Lambda(T)$ and more specifically:

$$\begin{aligned}
\sqrt{n(\mathcal{I}_{\mathcal{J}})}(\Psi(\hat{\Lambda}_j(t)) - \Psi(\Lambda_j)(t)) &= \Psi'_{\Lambda_j}(\sqrt{n(\mathcal{I}_{\mathcal{J}})}(\hat{\Lambda}_j(t) - \Lambda_j(t))) + o_P(1) \\
&= \sqrt{n(\mathcal{I}_{\mathcal{J}})}\mathbb{P}_{\mathcal{J}}\Psi'_{\Lambda_j}(\psi_j(t)) + o_P(1) \\
&= \sqrt{n(\mathcal{I}_{\mathcal{J}})}\mathbb{P}_{\mathcal{J}}S_j(t)\int_0^t \frac{d\psi_j(u)}{(1 - \Delta\Lambda_j(u))} + o_P(1).
\end{aligned}$$

This result uses fundamental results from Kosorok (2008) (see also van der Vaart (1998), Theorem 20.8). □

C.1.4 Estimation of the Treatment Propensity

In this section, we give the asymptotically linear form of the treatment propensity weights when they are estimated based on a stratification variable H (rather than using known weights with Z_i).

Remark C.1. *For an h such that $I_{\mathcal{J}}(h) = 1$, $\sum_{i=1}^n I(H_i = h) = \sum_{i\in\mathcal{I}_{\mathcal{J}}^{(h)}} I(H_i = h)$. Therefore, estimating $\hat{\pi}_j(h)$ using $n_j(\mathcal{I}_{\mathcal{J}}^{(h)})/n(\mathcal{I}_{\mathcal{J}}^{(h)})$ versus $n_j(H_i)/n(H_i)$ are equivalent. We are*

interested in asymptotics with respect to the empirical measure $\mathbb{P}_{\mathcal{J}}$, so we use the former expression of $\hat{\pi}_j(h)$ to make this explicit in Lemma C.4. Furthermore, $\pi_j(h) \equiv P(A = j|H = h) = P_{\mathcal{J}}(A = j|H = h)$.

Lemma C.4 (Treatment Propensity Estimation)**.** *Let $\hat{\pi}_j(h) = n_j(\mathcal{I}_{\mathcal{J}}^{(h)})/n(\mathcal{I}_{\mathcal{J}}^{(h)})$. Then, for a given level h of the stratification variable, for which $P(A = j|H = h) > 0$,*

$$\hat{\pi}_j(h)^{-1} - \pi_j(h)^{-1} = \mathbb{P}_{\mathcal{J}}\phi_{j,h}^{A}(W_i) + o_P(1/\sqrt{n^*})$$

where $\phi_{j,h}^{A}(W_i) = \frac{-I(H=h)[I(A=j)-\pi_j(h)]}{P_{\mathcal{J}}(H=h)\pi_j(h)^2}$.

Proof. First, we note that $\hat{\pi}_j(h) = \mathbb{P}_{\mathcal{J}}I(H_i = h, A_i = j)/\mathbb{P}_{\mathcal{J}}I(H_i = h)$. The numerator and denominator can be written in asymptotically linear forms, and then using the delta method using function $g(a, b) = a/b$, and $g'(a, b) = (1/b, -a/b^2)$ applied to $a = P(H = h, A = j)$ and $b = P(H = h)$, we have:

$$\begin{aligned}
&\hat{\pi}_j(h) - \pi_j(h) \\
&= \mathbb{P}_{\mathcal{J}}\left\{\frac{I(H_i = h, A_i = j) - P_{\mathcal{J}}(H = h, A = j)}{P_{\mathcal{J}}(H = h)} - P_{\mathcal{J}}(H = h, A = j)\frac{I(H_i = h) - P_{\mathcal{J}}(H = h)}{P_{\mathcal{J}}(H = h)^2}\right\} \\
&\quad + o_P(1/\sqrt{n^*}) \\
&= \mathbb{P}_{\mathcal{J}}\frac{I(H = h)\left[I(A = j) - \pi_j(h)\right]}{P_{\mathcal{J}}(H = h)} + o_P(1/\sqrt{n^*}).
\end{aligned}$$

Applying the delta method once again to take the inverse of $\hat{\pi}_j(h)$ using function $g(a) = 1/a$ with $g'(a) = -1/a^2$ and $a = \pi_j(h)$ we have:

$$\begin{aligned}
\hat{\pi}_j(h)^{-1} - \pi_j(h)^{-1} &= \mathbb{P}_{\mathcal{J}}\phi_{j,h}^{A}(W_i) + o_P(1/\sqrt{n^*}) \\
\phi_{j,h}^{A}(W_i) &= \frac{-I(H = h)\left[I(A = j) - \pi_j(h)\right]}{P_{\mathcal{J}}(H = h)\pi_j(h)^2}.
\end{aligned}$$

□

C.1.5 IPCW Weights and the Breslow Estimator

In this section we detail the estimator that we use for inverse probability of censoring weighting in Theorem 3.1 when V is non-empty. For a given strata level v (of V^S), let $\mathcal{I}_{\mathcal{J}}(j,v) = \{i \in \mathcal{I}_{\mathcal{J}} : V_i^S = v, A_i = j\}$.

Lemma C.5. *Let $N_j^C(s) = I(Y^{(j)} \leq t, \delta^{(j)} = 0)$ be the censoring counting process. Consider the following estimator of the cumulative hazard of being censored given a covariate profile v_P within strata v:*

$$\hat{\Lambda}_{j,v}^C(t|V^P = v_P) = \exp(\hat{\beta}_{j,v}^T v_P) \sum_{i \in \mathcal{I}_{\mathcal{J}}(v)} \int_0^t \frac{dN_{ij}^C(u)}{\sum_{k \in \mathcal{I}_{\mathcal{J}}(v)} Y_{kj}(u) \exp(\hat{\beta}_{j,v}^T V_k^P)}$$

where $\hat{\beta}_{j,v}$ is the estimated coefficient vector from the Cox proportional hazards model for strata v given the form in Assumption 3.3, estimated only with individuals in $\mathcal{I}_{\mathcal{J}}(j,v)$. Let $\mathbb{P}_{\mathcal{J}(j,v)}$ be the corresponding empirical measure. Under mild regularity conditions, $\hat{\beta}_{j,v} \to \beta_{j,v}$ under the model in Assumption 3.3. Then, we have the following asymptotically linear form:

$$\hat{\Lambda}_{j,v}^C(V^P)(t) - \Lambda_{j,v}^C(V^P)(t) = \mathbb{P}_{\mathcal{J}(j,v)} \psi_{j,v,i}^C(V_P)(t) + o_P(1/\sqrt{n^*})$$
$$\psi_{j,v,i}^C(V^P)(t) = \exp(\beta_{j,v}^T V^P) \left\{ \left[V^P \Lambda_{0,j,v}^C(t) - g_{j,v}(t) \right]^T \Omega_{j,v}^{-1} \cdot \xi_{j,v}^{(1)}(W_i) + \xi_{j,v}^{(2)}(W_i)(t) \right\}$$

where $\Lambda_{0,j,v}^C(t)$ is the baseline hazard for censoring in strata v. For a vector a, define $a^{\otimes 0} = 1$,

$a^{\otimes 1} = a$, *and* $a^{\otimes 2} = aa^T$. *The other terms in the influence function are given by:*

$$\begin{aligned}
\xi_{j,v}^{(1)}(W_i) &= \int_0^\tau \left[V_i^P - c_{j,v}^{(1)}(t)\right] dL_{i,j,v}^C(t) \\
\xi_{j,v}^{(2)}(W_i) &= \int_0^t \frac{dL_{i,j,v}^C(t)}{r_{j,v}^{(0)}(t)} \\
\Omega_{j,v} &= \int_0^\tau \left\{c_{j,v}^{(2)}(t) - c_{j,v}^{(1)}(t)\left[c_{j,v}^{(1)}(t)\right]^T\right\} E_{\mathcal{J}(v)}(Y_j(t)\lambda_{j,v}^C(t))dt \\
dL_{i,j,v}^C(t) &= dN_{i,j}^C(t) - Y_{i,j}(t)\exp(\beta_{j,v}^T V_i^P)d\Lambda_{0j,v}^C(t) \\
g_{j,v}(t) &= \int_0^t c_{j,v}^{(1)}(s)d\Lambda_{0j,v}^C(s) \\
c_{j,v}^{(k)}(t) &= \frac{r_{j,v}^{(k)}(t)}{r_{j,v}^{(0)}(t)} \\
r_{j,v}^{(k)}(t) &= E_{\mathcal{J}(v)}\left[Y_j(t)\exp(\beta_{j,v}^T V^P)(V^P)^{\otimes k}\right] \\
\lambda_{j,v}^C(t) &= \lambda_{0j,v}^C(t)\exp(\beta_{j,v}^T V^P)
\end{aligned}$$

where $E_{\mathcal{J}(v)}$ *is the expectation with respect to strata* $V^S = v$ *within the ECE population, and* $\lambda_{0j,v}(t)$ *is the baseline censoring hazard function for arm* j *and stratification level* v, $\tau \equiv \tau_j(V)$ *is the fixed value from Definition 3.2. Note that each of the* ξ *terms are functions of the observed data vector* $(N_{i,j}^C, Y_{i,j}, V_i^P)$. *Also note that when estimating* $\hat{\Lambda}_{j,v}^C(t)$, *observations in the specific strata* v *which extend further than the maximum follow-up time* $\tau_j(V)$ *from Definition 3.2 should be censored in order for this influence function to be valid.*

Proof. See Zhang and Schaubel (2012) Appendix A.1, but where we have specified using our notation that this is for a specific stratification level v. □

Lemma C.6. *Consider the following estimator of the probability of remaining uncensored given a proportional hazards covariate profile* V^P *within strata* $V^S = v$*:*

$$\hat{p}_{j,v}^C(V^P)(t) = \exp\left\{-\hat{\Lambda}_{j,v}^C(V^P)(t)\right\}.$$

Then we have the following asymptotically linear form for the inverse probability:

$$(\hat{p}^C_{j,v}(V^P)(t))^{-1} - (p^C_{j,v}(V^P)(t))^{-1} = \mathbb{P}_{\mathcal{J}}\phi^C_{j,v,i}(V^P)(t) + o_P(1/\sqrt{n^*})$$
$$\phi^C_{j,v,i}(V^P)(t) = -(p^C_{j,v}(V^P)(t))^{-1}\frac{I(V^S_i = v)I(A_i = j)}{P_{\mathcal{J}}(V^S = v)\pi_j(v)}\int_0^t \frac{d\psi^C_{j,v,i}(V^P)(t)}{\left[1 - \Delta\Lambda^C_{j,v}(V^P)(t)\right]}$$

where $\psi^C_{j,v,i}$ is the influence function from Lemma C.5, and $\pi_j(v) = P_{\mathcal{J}}(A = j|V^S = v)$. Furthermore $(\hat{p}^C_{j,v}(V^P)(t))^{-1}$ is uniformly consistent for $(p^C_{j,v}(V^P)(t))^{-1}$.

Proof. We first apply the results of the functional delta method in Lemma C.3, then again apply the delta method to obtain the inverse of the predicted probabilities. Finally, we use Slutsky's Lemma to convert results in terms of the strata-specific empirical measure $\mathbb{P}_{\mathcal{J}(j,v)}$ to the overall empirical measure $\mathbb{P}_{\mathcal{J}}$. The uniform consistency results are apparent from Zhang and Schaubel (2012) Appendix A.2.

□

C.2 Asymptotic Theory

C.2.1 Empirical Process Theory for Kaplan-Meier Estimators

In this section, we show a general expansion for deriving the asymptotic behavior of Kaplan-Meier estimators, when they are weighted with a weight function $\hat{f}_j(s)$. In some cases (e.g., using known propensity weights and under independent censoring), this weight function will not be estimated and will be time-invariant, and therefore will reduce to f_j. We derive the asymptotic behavior in its most general form, and then show the simplifications that result under various assumptions. This general outline will also be used when we derive the asymptotic behavior of the post-stratification estimators, which are a linear combination of a special case of IPW estimators where the weights are constant.

Let $L_j(s) = N_j(s) - R_j(s)\Lambda_j(s)$, where $\Lambda_j(s) = P_{\mathcal{J}}N_j(s)/P_{\mathcal{J}}R_j(s)$. In all of our asymptotic results, we will only consider $t \in [0, \tau]$ where τ is a time such that $P(T_j \geq \tau) \equiv PY_j(\tau) > 0$. This condition is implied by any τ that is allowed by Definition 3.2.

Let $\mathbb{Y}_{j,n}(s) := I\{\mathbb{P}_{\mathcal{J}}\hat{f}_j(s)R_j(s) > 0\}$ be the event that empirical risk set is non-empty for group j at time s, and $\mathbb{Y}^C_{j,n}(s)$ be its complement. Note that $\mathbb{Y}^C_{j,n}(s) = 0 \implies \mathbb{P}_{\mathcal{J}}\hat{f}_j(s)dN_j(s) = 0$, i.e., there can be no events if no one is in the risk set. Likewise, $\mathbb{Y}^C_{j,n}(s) = 0 \implies \mathbb{P}_{\mathcal{J}}\hat{f}_j(s)dL_j(s) = 0$. Then we can make the following expansion, which uses similar elements to Wang et al. (2020) and Zhang et al. (2015):

$$
\begin{aligned}
&\hat{\Lambda}_j^{\text{IPW}}(t) - \Lambda_j(t) \\
&= \int_0^t \frac{\mathbb{P}_{\mathcal{J}}\hat{f}_j(s)dN_j(s)}{\mathbb{P}_{\mathcal{J}}\hat{f}_j(s)Y_j(s)} - \int_0^t d\Lambda_j(s) \\
&= \int_0^t \left\{ \frac{\mathbb{P}_{\mathcal{J}}\hat{f}_j(s)dN_j(s)}{\mathbb{P}_{\mathcal{J}}\hat{f}_j(s)R_j(s)} - \frac{\mathbb{P}_{\mathcal{J}}\hat{f}_j(s)R_j(s)}{\mathbb{P}_{\mathcal{J}}\hat{f}_j(s)R_j(s)} d\Lambda_j(s)\mathbb{Y}_{j,n}(s) \right\} - \int_0^t \mathbb{Y}^C_{j,n}(s)d\Lambda_j(s) \\
&= \int_0^t \left\{ \frac{\mathbb{P}_{\mathcal{J}}\hat{f}_j(s)[dN_j(s) - R_j(s)d\Lambda_j(s)]}{\mathbb{P}_{\mathcal{J}}w_j\hat{f}_j(s)R_j(s)} \right\} \mathbb{Y}_{j,n}(s) - \int_0^t \mathbb{Y}^C_{j,n}(s)d\Lambda_j(s) \\
&= \int_0^t \left\{ \frac{\mathbb{P}_{\mathcal{J}}\hat{f}_j(s)dL_j(s)}{\mathbb{P}_{\mathcal{J}}\hat{f}_j(s)R_j(s)}\mathbb{Y}_{j,n}(s) + \frac{\mathbb{P}_{\mathcal{J}}\hat{f}_j(s)dL_j(s)}{P_{\mathcal{J}}f_j(s)R_j(s)} - \frac{\mathbb{P}_{\mathcal{J}}\hat{f}_j(s)dL_j(s)}{P_{\mathcal{J}}f_j(s)R_j(s)} \right\} - \int_0^t \mathbb{Y}^C_{j,n}(s)d\Lambda_j(s) \\
&= \int_0^t \left\{ \frac{\mathbb{P}_{\mathcal{J}}\hat{f}_j(s)dL_j(s)}{P_{\mathcal{J}}f_j(s)R_j(s)} \right\} + \int_0^t \mathbb{P}_{\mathcal{J}}\hat{f}_j(s)dL_j(s) \left\{ \frac{1}{\mathbb{P}_{\mathcal{J}}\hat{f}_j(s)R_j(s)} - \frac{1}{P_{\mathcal{J}}f_j(s)R_j(s)} \right\} \mathbb{Y}_{j,n}(s) \\
&\qquad - \int_0^t \mathbb{Y}^C_{j,n}(s)d\Lambda_j(s) \\
&= \underbrace{\int_0^t \frac{dA_n(s)}{B_0(s)}}_{(I)} + \underbrace{\int_0^t C_n(s)dA_n(s)}_{(II)} + \underbrace{\int_0^t \mathbb{Y}^C_{j,n}(s)d\Lambda_j(s)}_{(III)}
\end{aligned}
$$

where

$$A_n(s) = \mathbb{P}_{\mathcal{J}}\hat{f}_j(s)L_j(s) \tag{C.1}$$

$$B_0(s) = P_{\mathcal{J}}f_j(s)R_j(s) \tag{C.2}$$

$$C_n(s) = \left\{ \frac{1}{\mathbb{P}_{\mathcal{J}}\hat{f}_j(s)R_j(s)} - \frac{1}{P_{\mathcal{J}}f_j(s)R_j(s)} \right\} \mathbb{Y}_{j,n}(s) \tag{C.3}$$

In the proofs of Theorem 3.1 and 3.2, the goal will be to derive the influence function that

results from term (I), and then to show that (II) and (III) are asymptotically negligible. We can use results from Kosorok (2008)'s book to claim that (II) is asymptotically negligible, as we outline in Lemma C.7. The reason that Lemma C.7 applies to term (II) above is because C_n is left-continuous with right limits (and it is predictable), it is uniformly $o_P(1)$ on the range we are considering (this can be seen by the fact that on the set $\mathbb{Y}_{j,n}(s)$, the denominators in $C_n(s)$ are always bounded away from zero, and their difference is uniformly $o_P(1)$), and it has uniformly bounded total variation because it is the difference between two functions that have bounded total variation (they are both monotone decreasing functions up to a point, then exactly zero). Also, it is evident through the Proof of Theorem 3.1 that $\sqrt{n}A_n(s)$ converges weakly. Essentially, it is term (I) above, but not divided by $B_0(s)$.

Term (III) is asymptotically negligible following the logic in the Proof of Theorem 2 from Wang et al. (2021). Note that our definition of $\mathbb{Y}^C_{j,n}(s)$ has the weights involved, but on the bounded time range that we consider, this is equivalent to the non-weighted set. In other words, the empirical risk set is non-empty if and only if the weighted empirical risk set is non-empty, because our weights are always bounded away from zero.

Lemma C.7. *Let A_n and C_n be defined as above. If (i) C_n is a cadlag (or caglad) function on $[0,t]$ with a finite ℓ^∞ norm, (ii) $\sup_{s\in[0,t]}|C_n(s)| = o_P(1)$, (iii) C_n has uniformly bounded total variation, and (iv) $\sqrt{n}A_n$ converges weakly to a tight, mean-zero process, then $\int_0^t C_n(s)dA_n(s) = o_P(1/\sqrt{n^*})$.*

Proof. See Kosorok (2008) Lemma 4.2. □

C.2.2 Proof of Theorem 3.1

Here we walk through the steps to prove Thoerem 3.1, aided by the intermediate results in Sections C.1 and C.2.1. Let $\hat{f}_j(W_i) = I(A_i = j)\hat{w}_j(W_i)^{-1}$. Let $f_j(W_i)$ be its limiting counterpart, with $\hat{w}_j$ replaced by w_j. The functions $\hat{w}_j$ and w_j are the combination of known/estimated propensity weights and censoring weights, if applicable, defined in Theorem 3.1.

By Lemma C.2, $P_{\mathcal{J}} f_j(s) dL_j(s) = 0$, so dA_n from (C.1) can be rewritten as $\mathbb{P}_{\mathcal{J}} \hat{f}_j(s) dL_j(s) - P_{\mathcal{J}} f_j(s) dL_j(s)$, and we can make a further expansion from there:

$$dA_n = \underbrace{(\mathbb{P}_{\mathcal{J}} - P_{\mathcal{J}}) f_j(s) dL_j(s)}_{(a)} + \underbrace{P_{\mathcal{J}}(\hat{f}_j(s) - f_j(s)) dL_j(s)}_{(b)} + \underbrace{(\mathbb{P}_{\mathcal{J}} - P_{\mathcal{J}})(\hat{f}_j(s) - f_j(s)) dL_j(s)}_{(c)}. \tag{C.4}$$

Terms (a) and (b) may contribute to the influence function. Term (c) is $o_P(1/\sqrt{n})$ when integrated from $0 \to t$, and dividing by $B_0(s)$ because $\sup_{s,i} |\hat{f}_{i,j}(s) - f_{i,j}(s)| = o_P(1)$ for all v since $w_{i,j}$ (and its estimated counterpart) are bounded away from zero, and the censoring weight estimator is uniformly consistent (see Lemma C.6). One note is that the proportional hazards censoring covariates V_P (if used) must have bounded support, which is likely satisfied in practice.

Inverse Probability Weighting: Theorem 3.1 (i)

We focus first on the case where V is empty, so we do not do any IPCW. In this case, $\hat{w}_j(W_i) = \pi_j(Z_i)^{-1}$ when the propensities are known, and $\hat{w}_j(W_i) = \hat{\pi}_j(H_i)^{-1}$ when they are estimated. We will work through the case where the π_j are estimated, and treat the case where π_j are known as a special case. For term (a), we have:

$$(a) = \mathbb{P}_{\mathcal{J}} \phi_a(W)(s)$$
$$\phi_a(W) : s \to I(A = j) \pi_j(H)^{-1} dL_j(s)$$

When π_j is known, then the influence function for the cumulative hazard is given by:

$$\hat{\Lambda}_j^{\text{IPW}}(t) - \Lambda_j(t) = \mathbb{P}_{\mathcal{J}} \int_0^t \frac{I(A = j) \pi_j(Z)^{-1} dL_j(s)}{E_{\mathcal{J}}\left(R_j(s)\right)} \tag{C.5}$$

By applying Lemma C.3 with the influence function (C.5) to transform the cumulative hazard to the survival function, we have:

$$\phi_j^{(1)}(A, Z, L_j, Y_j)(s) = S_j(t) \int_0^t \frac{I(A = j)\pi_j(Z)^{-1} dL_j(s)}{P_{\mathcal{J}}(R_j(s)) \{1 - \Delta\Lambda_j(s)\}}.$$

To estimate the asymptotic variance of $\phi_j^{(1)}$, we use the facts that $L_j(s)$ is a zero-mean Martingale and that the function $F(s) = I(A = j)\pi_j(Z)^{-1} P_{\mathcal{J}}(R_j(s))^{-1} \{1 - \Delta\Lambda_j(s)\}^{-1}$ is predictable, and bounded. Using Fleming and Harrington (2011) we have weak convergence (by Theorem 5.1.1). Note that the weak convergence result will apply to the future sections with the influence functions that account for weighting in this first term as well, because the weight functions are always bounded and predictable.

Also by Fleming and Harrington (2011), the variance of the Martingale transformation is given by:

$$\begin{aligned}
&\mathrm{Var}_{\mathcal{J}}(\phi_j^{(1)}) \\
&= S_j(t)^2 \mathrm{Var}_{\mathcal{J}} \left(\int_0^t F(s) dL_j(s) \right) \\
&= S_j(t)^2 \int_0^t E_{\mathcal{J}} \left(F(s)^2 R_j(s) \right) \{1 - \Delta\Lambda_j(s)\} \, d\Lambda_j(u) \\
&= S_j(t)^2 \int_0^t \frac{1}{P_{\mathcal{J}}(R_j(s))^2 \{1 - \Delta\Lambda_j(s)\}^2} E_{\mathcal{J}} \left(I(A_i = j)\pi_j(Z)^{-2} R_j(s) \right) \{1 - \Delta\Lambda_j(s)\} \, d\Lambda_j(u) \\
&= S_j(t)^2 \int_0^t \frac{d\Lambda_j(s)}{\{1 - \Delta\Lambda_j(s)\} P_{\mathcal{J}}(R_j(s))} \cdot \frac{E_{\mathcal{J}} \left(I(A_i = j)\pi_j(Z)^{-2} R_j(s) \right)}{P_{\mathcal{J}}(R_j(s))}.
\end{aligned}$$

This asymptotic variance is very similar to the typical variance of the Kaplan-Meier estimator: the first term is exactly the typical asymptotic variance. The second term is new because it contains information about the weights; however, if all $\pi_j(Z) = 1$, then the second term is 1 and we are left with the original KM asymptotic variance. However, we cannot estimate each component of the asymptotic variance naively, because we need to take into account consistency of the estimate of each of the quantities to the correct quantity in the ECE

population. E.g., to consistently estimate $P_{\mathcal{J}}(R_j(s))$, we need to use the weighted estimator $n(\mathcal{I}_{\mathcal{J}})^{-1}\sum_{i\in\mathcal{I}_{\mathcal{J}}} I(Y_i \geq s)I(A_i = j)\pi_j(Z_i)^{-1}$. We will construct these consistent plug-ins for each of the components of the asymptotic variance above. Let $\mathcal{T}_{\mathcal{J}}^{(j)} = \{Y_i : \delta_i = 1, A_i = j\}_{i\in\mathcal{I}_{\mathcal{J}}}$ be the set of unique failure times for those within the ECE sample and assigned to treatment j. Let

$$\tilde{d}_{\mathcal{J}}^{(j)}(t) = \sum_{i\in\mathcal{I}_{\mathcal{J}}} \delta_i I(Y_i = t)I(A_i = j)\pi_j(Z_i)^{-1}$$
$$\tilde{n}_{\mathcal{J}}^{(j)}(t) = \sum_{i\in\mathcal{I}_{\mathcal{J}}} I(Y_i \geq t)I(A_i = j)\pi_j(Z_i)^{-1}$$

be the weighted event count and weighted risk sets at time t. We use the $\tilde{d}$ and $\tilde{n}$ notation to distinguish these weighted counts from the raw event count and risk set sizes without weighting. Then a consistent estimator for each of the quantities is given by:

$$d\hat{\Lambda}_j(s) \equiv \Delta\hat{\Lambda}_j(s) = \frac{\tilde{d}_{\mathcal{J}}^{(j)}(s)}{\tilde{n}_{\mathcal{J}}^{(j)}(s)}$$
$$\widehat{P}_{\mathcal{J}}(R_j(s)) = n(\mathcal{I}_{\mathcal{J}})^{-1}\tilde{n}_{\mathcal{J}}^{(j)}(s)$$
$$\widehat{E}_{\mathcal{J}}\left(\pi_j(Z)^{-1}R_j(s)\right) = n(\mathcal{I}_{\mathcal{J}})^{-1}\sum_{i\in\mathcal{I}_{\mathcal{J}}} I(Y_i \geq s)I(A_i = j)\pi_j(Z_i)^{-2}$$

Therefore, a consistent variance estimator (after dividing by $n(\mathcal{I}_{\mathcal{J}})$) is given by:

$$\widehat{\mathrm{Var}}(\widehat{S}_j^{\mathrm{IPW}(1)}(t))$$
$$= \left[\hat{S}_j^{\mathrm{IPW}(1)}(t)\right]^2 \sum_{s\in\mathcal{T}_{\mathcal{J}}^{(j)}:s\leq t} \left[\frac{\tilde{d}_{\mathcal{J}}^{(j)}(s)}{\tilde{n}_{\mathcal{J}}^{(j)}(s)\left\{\tilde{n}_{\mathcal{J}}^{(j)}(s) - \tilde{d}_{\mathcal{J}}^{(j)}(s)\right\}} \cdot \frac{\sum_{i\in\mathcal{I}_{\mathcal{J}}} \pi_j(Z_i)^{-2}I(A_i = j)I(Y_i \geq s)}{\tilde{n}_{\mathcal{J}}^{(j)}(s)}\right].$$

where $\hat{S}_j^{\mathrm{IPW}(1)}(t)$ is the survival function estimated based on known weights. If the ECE population were the entire trial, this variance estimator is equivalent to the variance estimator for the AKME with known weights proposed in Xie and Liu (2005).

Inverse Probability Weighting: Theorem 3.1 (ii)

To derive the influence function with estimated weights, we have the additional term (b) in equation (C.1). For term (b), which is only relevant when the propensities are estimated (otherwise it is zero) we can condition on H which gives conditional independence, i.e., $A \perp L_j(s)|H$, and then apply Lemma C.8 that has the influence function for $\hat{\pi}_j(H)^{-1}$. Note that below, the expectation is taken over (A, H), and the $\hat{\pi}_j$ is treated as fixed. We have:

$$\begin{aligned}
(b) &= E_{\mathcal{J}}\left\{\left[\frac{I(A=j)}{\hat{\pi}_j(H)} - \frac{I(A=j)}{\pi_j(H)}\right] dL_j(s)\right\} \\
&= E_{\mathcal{J}}\left\{\left[\frac{\pi_j(H)}{\hat{\pi}_j(H)} - \frac{\pi_j(H)}{\pi_j(H)}\right] E\left(dL_j(s)|H\right)\right\} \\
&= \sum_{h=1}^{L} p(h)\pi_j(h) E\left(dL_j(s)|H\right) \phi_{j,h}^{A}(A_i, H_i) + o_P(1/\sqrt{n^*}) \\
&= -\sum_{h=1}^{L} \pi_j(h)^{-1} e_j(h)(s) I(H=h)\left[I(A=j) - \pi_j(h)\right] + o_P(1/\sqrt{n^*})
\end{aligned}$$

where $p(h) = P_{\mathcal{J}}(H=h)$ and $e_j(h)(s) = E_{\mathcal{J}}\left\{dL_j(s)|H=h\right\}$. Therefore, the influence function for the IPW estimator with estimated weights is given by (after applying the functional delta method to transform to the survival scale):

$$\phi_j^{(2)}(W_i)(t) = \phi_j^{(1)}(W_i)(t) - \underbrace{S_j(t)\int_0^t \frac{\sum_{h=1}^{L} \pi_j(h)^{-1} e_j(h)(s) I(H_i=h)\left[I(A_i=j) - \pi_j(h)\right]}{P_{\mathcal{J}}\left(R_j(s)\right)\left(1 - \Delta\Lambda_j(s)\right)}}_{\xi_j(H_i, A_i)(t)}.$$

To compute the variance of this influence function, we have:

$$\mathrm{Var}(\phi_j^{(2)}(W_i)(t)) = \mathrm{Var}(\phi_j^{(1)}(W_i)(t)) + \mathrm{Var}(\xi_j(H_i, A_i)(t)) - 2\mathrm{Cov}(\phi_j^{(1)}(W_i)(t), \xi_j(H_i, A_i))$$

We have already computed the first term on the RHS. Now we will compute the last two terms on the RHS. Recall that the only random component of ξ_j is the variables (H_i, A_i), which are not functions of t. As a result, it is a non-stochastic integral, and the random

terms can be separated from the integral.

$$\begin{aligned}
\text{Var}(\xi_j(H_i, A_i)(t)) &= [S_j(t)]^2 \,\text{Var}\left(\sum_{h=1}^{L} b_h(A,H)c_h(t)\right) \\
b_h(A,H) &= \pi_j(h)^{-1} I(H=h)\left[I(A=j) - \pi_j(h)\right] \\
c_h(t) &= \int_0^t e_j(h)(s) P_{\mathcal{J}}\left(R_j(s)\right)^{-1}\left(1 - \Delta\Lambda_j(s)\right)^{-1}.
\end{aligned}$$

Therefore, using the law of total variance, and noting that $c_h(t)$ is a constant rather than a random variable, we have the following expression:

$$\begin{aligned}
\text{Var}(\xi_j(H_i, A_i)(t)) &= [S_j(t)]^2 \left\{\sum_{h=1}^{L} c_h(t)^2 E\left\{\text{Var}\left(b_h(A,H)|H\right)\right\} + \text{Var}\left\{\sum_{h=1}^{L} c_h(t) E\left(b_h(A,H)|H\right)\right\}\right\} \\
&= [S_j(t)]^2 \left\{\sum_{h=1}^{L} c_h(t)^2 E\left\{\text{Var}\left(b_h(A,H)|H\right)\right\} + \text{Var}\left\{\sum_{h=1}^{L} c_h(t) E\left(b_h(A,H)|H\right)\right\}\right\}.
\end{aligned}$$

Since $\text{Var}(b_h(A,H)|H) = I(H=h)(1-\pi_j(h))\pi_j(h)^{-1}$ and $E(b_h(A,H)|H) = 0$, we have

$$\text{Var}(\xi_j(H_i, A_i)(t)) = [S_j(t)]^2 \sum_{h=1}^{L} c_h(t)^2 p(h)\left[\frac{1-\pi_j(h)}{\pi_j(h)}\right].$$

For the covariance term, we have the following:

$$\begin{aligned}
\text{Cov}(\phi_j^{(1)}(W_i)(t), \xi_j(H_i, A_i)) &= [S_j(t)]^2 \sum_{h=1}^{L} E\left\{\int_0^t \frac{I(A=j)\pi_j(H)^{-1} dL_j(s)}{P_{\mathcal{J}}(R_j(s))\left[1-\Delta\Lambda_j(s)\right]} \cdot c_h(t) b_h(A,H)\right\} \\
&= [S_j(t)]^2 \sum_{h=1}^{L} p(h) c_h(t)\left[\frac{1-\pi_j(h)}{\pi_j(h)}\right] c_h(t)
\end{aligned}$$

where we have obtained the last line through iterated expectation first conditioning on H. This covariance term is the same as the variance term for ξ_j. Therefore, the asymptotic

variance is given by:

$$\text{Var}_{\mathcal{J}}(\phi_j^{(2)}) = \text{Var}_{\mathcal{J}}(\phi_j^{(1)}) - [S_j(t)]^2 \sum_{h=1}^{L} c_h(t)^2 p(h) \left[\frac{1-\pi_j(h)}{\pi_j(h)}\right].$$

In their asymptotic forms, we can see that $\text{Var}_{\mathcal{J}}(\phi_j^{(2)}) \leq \text{Var}_{\mathcal{J}}(\phi_j^{(1)})$, which means that we gain precision from estimating the propensity weights. We will have more efficiency gain if the strata-specific cumulative hazard functions are notably different than the marginal cumulative hazard functions, which we can see based on the alternative form of $c_h(t)$ below.

To estimate the variance of the first term on the RHS, we can use the same strategy as we used for known weights, but replace all of the known weights with estimated weights, to obtain the event and at-risk sizes $\hat{d}_{\mathcal{J}}^{(j)}(t) = \sum_{i \in \mathcal{I}_{\mathcal{J}}} \delta_i I(Y_i = t) I(A_i = j) \hat{\pi}_j(H_i)^{-1}$, and $\hat{n}_{\mathcal{J}}^{(j)}(t) = \sum_{i \in \mathcal{I}_{\mathcal{J}}} I(Y_i \geq t) I(A_i = j) \hat{\pi}_j(H_i)^{-1}$. For the second term, we have some simple quantities like $p(h)$ and $\pi_j(h)$, along with more complicated quantities like $c_h(t)$. To estimate $c_h(t)$, we can re-write it as:

$$\begin{aligned}
&c_h(t) \\
&= \int_0^t \frac{E(dL_j(s)|H=h)}{P_{\mathcal{J}}(R_j(s))\left[1-\Delta\Lambda_j(s)\right]} \\
&= \int_0^t \frac{E(dN_j(s)|H=h) - E(Y_j(s)|H=h)d\Lambda_{j,h}(s) + E(Y_j(s)|H=h)d\Lambda_{j,h}(s) - E(Y_j(s)|H=h)d\Lambda_j(s)}{P_{\mathcal{J}}(R_j(s))\left[1-\Delta\Lambda_j(s)\right]} \\
&= \int_0^t \frac{E_{\mathcal{J}}(Y_j(s)|H=h)\left[d\Lambda_{j,h}(s) - d\Lambda_j(s)\right]}{P_{\mathcal{J}}(R_j(s))\left[1-\Delta\Lambda_j(s)\right]}
\end{aligned}$$

where $\Lambda_{j,h}(s)$ is the cumulative hazard for treatment j within the strata defined by h. We have used the fact that if we use the conditional hazard on h, then we have $E(dN_j(s) - Y_j(s)d\Lambda_{j,h}(s)|H=h) = 0$ because it is now a proper martingale term since we have the correct compensator. So rather than having the hazard in the numerator, we have the difference between the strata-specific and marginal hazards, weighted by the risk set size for strata h. Let $d_{\mathcal{J}}^{(j,h)}(t) = \sum_{i \in \mathcal{I}_{\mathcal{J}}^{(h)}} \delta_i I(Y_i = t) I(A_i = j)$ and $n_{\mathcal{J}}^{(j,h)}(t) = \sum_{i \in \mathcal{I}_{\mathcal{J}}^{(h)}} I(Y_i \geq t) I(A_i = j)$ be the

strata-specific event and at-risk sizes. We define them unweighted because since they are in the same strata, they all have the same weight, so we will pull this out of the expressions. The strata-specific terms can be estimated using:

$$d\hat{\Lambda}_{j,h}(s) \equiv \Delta\hat{\Lambda}_{j,h}(s) = d_{\mathcal{J}}^{(j,h)}(s)/n_{\mathcal{J}}^{(j,h)}(s)$$
$$\hat{E}_{\mathcal{J}}(Y_j(s)|H=h) = n(\mathcal{I}_{\mathcal{J}})^{-1}p(h)^{-1}\hat{\pi}_j(h)^{-1}n_{\mathcal{J}}^{(j,h)}(s).$$

Therefore, a consistent estimate of $c_h(t)$ is:

$$\hat{c}_h(t) = \frac{1}{\hat{\pi}_j(h)\hat{p}(h)} \sum_{s\in\mathcal{T}_{\mathcal{J}}^{(j)}:s\leq t} \left[\left(\frac{d_{\mathcal{J}}^{(j,h)}(s)}{n_{\mathcal{J}}^{(j,h)}(s)} - \frac{\hat{d}_{\mathcal{J}}^{(j)}(s)}{\hat{n}_{\mathcal{J}}^{(j)}(s)} \right) \left(\frac{n_{\mathcal{J}}^{(j,h)}(s)}{\hat{n}_{\mathcal{J}}^{(j)}(s) - \hat{d}_{\mathcal{J}}^{(j)}(s)} \right) \right],$$

and the full estimate of the asymptotic variance $\widehat{\Sigma}_{(2)}^{\mathrm{IPW}}(t)$ is given by:

$$\begin{aligned}
&\widehat{\Sigma}_{(2)}^{\mathrm{IPW}}(t) \left[\hat{S}_j^{\mathrm{IPW}(2)}(t) \right]^{-2} \\
&= n(\mathcal{I}_{\mathcal{J}}) \sum_{s\in\mathcal{T}_{\mathcal{J}}^{(j)}:s\leq t} \left[\frac{\hat{d}_{\mathcal{J}}^{(j)}(s)}{\hat{n}_{\mathcal{J}}^{(j)}(s)\left\{\hat{n}_{\mathcal{J}}^{(j)}(s) - \hat{d}_{\mathcal{J}}^{(j)}(s)\right\}} \cdot \frac{\sum_{i\in\mathcal{I}_{\mathcal{J}}} \hat{\pi}_j(H_i)^{-2} I(A_i=j) I(Y_i \geq s)}{\hat{n}_{\mathcal{J}}^{(j)}(s)} \right] \\
&\quad - \sum_{h=1}^{L} \hat{c}_h(t)^2 \hat{p}(h) \hat{\pi}_j(h)^{-1} \left[1 - \hat{\pi}_j(h)\right] \\
&= n(\mathcal{I}_{\mathcal{J}}) \sum_{s\in\mathcal{T}_{\mathcal{J}}^{(j)}:s\leq t} \left[\frac{\hat{d}_{\mathcal{J}}^{(j)}(s)}{\hat{n}_{\mathcal{J}}^{(j)}(s)\left\{\hat{n}_{\mathcal{J}}^{(j)}(s) - \hat{d}_{\mathcal{J}}^{(j)}(s)\right\}} \cdot \frac{\sum_{i\in\mathcal{I}_{\mathcal{J}}} \hat{\pi}_j(H_i)^{-2} I(A_i=j) I(Y_i \geq s)}{\hat{n}_{\mathcal{J}}^{(j)}(s)} \right] \\
&\quad - \sum_{h=1}^{L} \left(\frac{1-\hat{\pi}_j(h)}{\hat{\pi}_j(h)} \right) \hat{p}(h)^{-1} \hat{\pi}_j(h)^{-2} \left[\sum_{s\in\mathcal{T}_{\mathcal{J}}^{(j)}:s\leq t} \left(\frac{d_{\mathcal{J}}^{(j,h)}(s)}{n_{\mathcal{J}}^{(j,h)}(s)} - \frac{\hat{d}_{\mathcal{J}}^{(j)}(s)}{\hat{n}_{\mathcal{J}}^{(j)}(s)} \right) \left(\frac{n_{\mathcal{J}}^{(j,h)}(s)}{\hat{n}_{\mathcal{J}}^{(j)}(s) - \hat{d}_{\mathcal{J}}^{(j)}(s)} \right) \right]^2
\end{aligned}$$

where $n(\mathcal{I}_{\mathcal{J}}^{(j,h)}) = \sum_{i\in\mathcal{I}_{\mathcal{J}}^{(h)}} I(A_i = j)$ and $\hat{S}_j^{\mathrm{IPW}(2)}(t)$ is the survival function estimated based on estimated propensity weights. To obtain a variance estimator, we divide by $n(\mathcal{I}_{\mathcal{J}})$:

$$\widehat{\mathrm{Var}}(\widehat{S}_j^{\mathrm{IPW}(2)}(t)) = n(\mathcal{I}_{\mathcal{J}})^{-1}\widehat{\Sigma}_{(2)}^{\mathrm{IPW}}(t).$$

Double Weighting with Known Treatment Propensities: Theorem 3.1 (iii)

Using the same decomposition, we have that:

$$(a) = \mathbb{P}_{\mathcal{J}}\phi_a^*(W_i)(s)$$
$$\phi_a^*(W_i) : s \to I(A_i = j)\pi_j(Z_i)^{-1}p_{j,V_i}^C(s)^{-1}dL_j(s)$$

where $p_{j,V}^C(s) = P_{\mathcal{J}}(C^{(j)} \geq s|V)$, and we have used the $*$ notation to distinguish between $\phi_a(W_i)$ used when we have independent censoring in parts (i) and (ii). The difference between this first term and that in the previous proofs is that the risk set on the denominator is weighted. Our results that follow differ from other double-weighted Kaplan-Meier functions in the literature, such as in Schaubel and Wei (2011), because we make different assumptions, and because we allow for a stratified Cox model. One of these key differences is that we assume $A \perp V|Z$ (or H), i.e., the variables that confer independent censoring are conditionally independent of treatment assignment, given the vector Z that specifies treatment assignment probabilities. In Schaubel and Wei (2011), their censoring variable may be a time-dependent $V(t)$, in which case, this conditional independence does not hold. Having this conditional independence allows us to simplify the influence function, as we will show below, and in the proof of Theorem 3.1 (iv). After transforming to the survival scale applying Lemma C.3), the contribution to the influence function from part (a) is given by:

$$\phi_j^{(1)}(W)(t) = S_j(t)\int_0^t \frac{I(A = j)\pi_j(Z)^{-1}p_{j,V}^C(s)^{-1}dL_j(s)}{P_{\mathcal{J}}(p_{j,V}^C(s)R_j(s))\left\{1 - \Delta\Lambda_j(s)\right\}}.$$

For term (b), we have:

$$(b) = E_{\mathcal{J}}\left\{I(A = j)dL_j(s)\left[\frac{1}{\pi_j(Z)\hat{p}_j^C(V)(s)} - \frac{1}{\pi_j(Z)p_j^C(V)(s)}\right]\right\}$$

where $p_j^C(V)(s)$ is the limit of $\hat{p}_j^C(V)(s)$, which we know is $P_{\mathcal{J}}(C^{(j)} \geq s|V_i)$ since it is a (uniformly) consistent estimator. Using the conditional independence of $A \perp V|Z$ from

Assumption 3.2, we can remove the $I(A = j)/\pi_z(Z)$ from above, which leaves us with the following component of the influence function for part (b), which we define as $d\gamma^{(b)}$:

$$
\begin{aligned}
&d\gamma^{(b)}(W_i)(s) \\
&= \sum_{v=1}^{S} P_{\mathcal{J}}(V^S = v) E_{\mathcal{J}}\left\{ dL_j(s)\phi^C_{j,v,i}(V^P)(s) \middle| V^S = v \right\} + o_P(1/\sqrt{n^*}) \\
&= -\sum_{v=1}^{S} \frac{I(A_i = j) I(V_i^S = v)}{P_{\mathcal{J}}(A = j | V^S = v)} \left[e^{(1)}_{j,v}(s)\xi^{(1)}_{j,v}(\delta_i, Y_i, V_i^P) + \int_0^s e^{(2)}_{j,v}(s,u)\xi^{(2)}_{j,v}(\delta_i, Y_i, V_i^P)(u) \right]
\end{aligned}
$$

where $\xi^{(1)}_{j,v}(\cdot)$ and $\xi^{(2)}_{j,v}(\cdot)(t)$ are defined in Lemma C.5, and

$$
\begin{aligned}
e^{(1)}_{j,v}(s) &= E_{\mathcal{J}(v)} \left\{ p^C_{j,v}(V^P)(s)^{-1} dL_j(s) \left[\int_0^s \frac{\left[V^P - c^{(1)}_{j,v}(u) \right] d\Lambda^C_{j,v}(u)}{1 - \Delta\Lambda^C_{j,v}(u)} \right] \right\} \Omega^{-1}_{j,v} \\
e^{(2)}_{j,v}(s,u) &= E_{\mathcal{J}(v)} \left\{ p^C_{j,v}(V^P)(s)^{-1} dL_j(s) \left[\frac{\exp(\beta^T_{j,v} V^P)}{\left[1 - \Delta\Lambda^C_{j,v}(u)\right] r^{(0)}_{j,v}(u)} \right] \right\}
\end{aligned}
$$

where $\Omega_{j,v}$ and $c^{(1)}_{j,v}(u)$ are defined in Lemma C.5, and $E_{\mathcal{J}(v)}(\cdot) = E_{\mathcal{J}}(\cdot | V^S = s)$. After transforming to the survival scale applying Lemma C.3), the full influence function is given by:

$$
\phi^{(3)}_j(W)(t) = \phi^{(1)}_j(W)(t) + S_j(t) \int_0^t \frac{d\gamma^{(b)}(W)(s)}{P_{\mathcal{J}}(p^C_{j,V}(s) R_j(s)) \left\{1 - \Delta\Lambda_j(s)\right\}}.
$$

This form of the influence function is very complicated due to the presence of $\gamma^{(b)}$. Therefore, when doing variance estimation with a non-empty V^P, we recommend using the bootstrap procedure. For variance estimation at time t when $V^P = \emptyset$, the influence function is much simplier, and we use the variance of the empirical influence function. With $V^P = \emptyset$, $d\gamma^{(b)}$

simplifies as follows (see Lemma C.5):

$$
\begin{aligned}
&d\gamma^{(b)}(W)(s) \\
&= -\frac{I(A_i = j)}{\pi_j(V_i) p^C_{j,V_i}(s)} E_{\mathcal{J}}\left\{dL_j(s) | V = V_i\right\} \int_0^s \frac{dL^C_{i,j}(u)}{E_{\mathcal{J}}(R_j(u)|V = V_i)\left[1 - \Delta\Lambda^C_{j,V_i}(u)\right]} \\
&= -\frac{I(A_i = j)}{\pi_j(V_i) p^C_{j,V_i}(s)} E_{\mathcal{J}}\left\{R_j(s) | V = V_i\right\} \left[d\Lambda_{j,V_i}(s) - d\Lambda_j(s)\right] \underbrace{\int_0^s \frac{dL^C_{i,j}(u)}{E_{\mathcal{J}}(R_j(u)|V = V_i)\left[1 - \Delta\Lambda^C_{j,V_i}(u)\right]}}_{M_i(u)}
\end{aligned}
$$

where $dL^C_{i,j,v}(t) = dN^C_{i,j}(t) - Y_{i,j}(t) d\Lambda^C_{j,V_i}(t)$, with $\Lambda^C_{j,V_i}(t)$ the cumulative hazard function for censoring among the ECE population for strata V_i in arm j.

To estimate the variance of $\phi^{(3)}_j(W)$ at time t, we use the empirical variance of the estimated influence function, where we have plugged in estimated quantities within the influence function:

$$
\widehat{\Sigma}^{\text{IPW}}_{(3)}(t) = \widehat{\text{Var}}_{\mathcal{J}}(\widehat{\phi}^{(3)}_j(W_i))
$$

where $\widehat{\text{Var}}_{\mathcal{J}}$ is the empirical variance within the samples from the ECE population. As we estimate the influence function, it is important to check that the estimated influence function is $O_P(1)$. When we construct the final variance estimator, we will divide the estimate of the asymptotic variance estimate by $n(\mathcal{I}_{\mathcal{J}})$. To estimate the influence function, we define the following quantities. Define the weighted risk set and event sizes at time t to be $\widehat{n}^{(j)}_{\mathcal{J}}(t) = \sum_{i \in \mathcal{I}_{\mathcal{J}}} I(Y_i \geq t) I(A_i = j) \hat{w}^{-1}_{i,j}$ and $\widehat{d}^{(j)}_{\mathcal{J}}(t) = \sum_{i \in \mathcal{I}_{\mathcal{J}}} \delta_i I(Y_i = t) I(A_i = j) \hat{w}^{-1}_{i,j}$. Based on these weighted risk and event counts, we denote

$$
\widehat{\sigma}^2_{\widehat{\Lambda}_j}(V)(s) = \sum_{s \in \mathcal{T}^{(j)}_{\mathcal{J}}} \left[\widehat{p}^C_j(V)(s)\right]^{-1} \frac{d\widehat{\Lambda}_j(s)}{\widehat{S}_j(s)\left[1 - d\widehat{\Lambda}_j(s)\right]}
$$

using the choice of weights in Theorem 3.1 (iii) to estimate $\widehat{d}$ and $\widehat{n}$. Then we can estimate

the first component of the influence function as:

$$\widehat{\phi}_j^{(1)}(W_i)(t) = \widehat{S}_j^{(3)}(t) I(A_i = j) \left\{ \frac{\delta_i I(Y_i \leq s) \cdot \hat{w}_{i,j}^{-1}}{\widehat{S}_j(Y_i)\left[1 - d\widehat{\Lambda}_j(Y_i)\right]} - \pi_j(Z_i)^{-1} \cdot \widehat{\sigma}^2_{\widehat{\Lambda}_j}(V_i)(\min(t, Y_i)) \right\}.$$

For the second component, we will first estimate $d\widehat{\gamma}^{(b)}(s)$. Define $n_{\mathcal{J}}^{(j,v)}(t) = \sum_{i \in \mathcal{I}_{\mathcal{J}}^{(v)}} I(A_i = j) I(Y_i \geq t)$ and $c_{\mathcal{J}}^{(j,v)}(t) = \sum_{i \in \mathcal{I}_{\mathcal{J}}^{(v)}} I(A_i = j) I(Y_i = t)(1 - \delta_i)$ as the treatment and strata-specific at-risk and censoring counts for time t. Define $\mathcal{F}_{\mathcal{J}}^{(j,v)} = \{Y_i : \delta_i = 0, A_i = j, V_i = v\}$ as the set of censoring times within strata v and treatment arm j, and let $\widehat{\sigma}^2_{\widehat{\Lambda}^C_{j,v}}(s) = \sum_{s \in \mathcal{F}_{\mathcal{J}}^{(j,v)}} c_{\mathcal{J}}^{(j,v)}(s) / \left[n_{\mathcal{J}}^{(j,v)}(s)(n_{\mathcal{J}}^{(j,v)}(s) - c_{\mathcal{J}}^{(j,v)}(s)) \right]$ is the estimated variance of the Nelson-Aalen estimator for censoring based on strata v. Then term $M_i(s)$ in $d\gamma^{(b)}(s)$ can be estimated using the following:

$$\begin{aligned}
\widehat{M}_i(s) &= \int_0^s \frac{n(\mathcal{I}_{\mathcal{J}}^{(V_i)})\pi_j(V_i)}{\left[n_{\mathcal{J}}^{(j,V_i)}(u) - c_{\mathcal{J}}^{(j,V_i)}(u) \right]} \left[(1 - \delta_i) I(Y_{i,j} = u) - I(Y_{i,j} \geq u) \frac{c_{\mathcal{J}}^{(j,V_i)}(u)}{n_{\mathcal{J}}^{(j,V_i)}(u)} \right] \\
&= n(\mathcal{I}_{\mathcal{J}}^{(V_i)}) \widehat{\pi}_j(V_i) \left[\frac{(1 - \delta_i) I(Y_i \leq s)}{n_{\mathcal{J}}^{(j,V_i)}(Y_i) - c_{\mathcal{J}}^{(j,V_i)}(Y_i)} - \widehat{\sigma}^2_{\widehat{\Lambda}^C_{j,V_i}}(\min(s, Y_i)) \right].
\end{aligned}$$

The $n(\mathcal{I}_{\mathcal{J}}^{(V_i)})\widehat{\pi}_j(V_i)$ cancels with the denominator during estimation of the $E_{\mathcal{J}}\{R_j(s)|V = V_i\}$ term, so we are left with

$$\begin{aligned}
&d\widehat{\gamma}^{(b)}(W_i)(s) \\
&= -\frac{I(A_i = j)}{\widehat{\pi}_j(V_i)\hat{p}^C_{j,V_i}(s)} n_{\mathcal{J}}^{(j,V_i)}(s) \left[\frac{d_{\mathcal{J}}^{(j,V_i)}(s)}{n_{\mathcal{J}}^{(j,V_i)}(s)} - \frac{\widehat{d}_{\mathcal{J}}^{(j)}(s)}{\widehat{n}_{\mathcal{J}}^{(j)}(s)} \right] \left\{ \frac{(1 - \delta_i) I(Y_i \leq s)}{n_{\mathcal{J}}^{(j,V_i)}(Y_i) - c_{\mathcal{J}}^{(j,V_i)}(Y_i)} - \widehat{\sigma}^2_{\widehat{\Lambda}^C_{j,V_i}}(\min(s, Y_i)) \right\}.
\end{aligned}$$

Note that $d\widehat{\gamma}^{(b)}$ has a jump at each failure time within the ECE samples for which $I(A_i = j)$, and at each censoring time within the strata corresponding to V_i, and $I(A_i = j)$, which is the set corresponding to $\mathcal{T}_{\mathcal{J}}^{(j)} \cup \mathcal{F}_{\mathcal{J}}^{(j,V_i)}$. Therefore, the final estimated influence function is

given by:

$$
\begin{aligned}
&\widehat{\phi}^{(3)}(W_i)(t) \\
&= \widehat{\phi}^{(1)}(W_i)(t) + \widehat{S}_j^{(3)}(t) \cdot \int_0^t \frac{d\widehat{\gamma}^{(b)}(W_i)(s)}{\widehat{S}_j(s)\left[1 - d\widehat{\Lambda}_j(s)\right]} \\
&= \widehat{\phi}^{(1)}(W_i)(t) - \widehat{S}_j^{(3)}(t) \cdot \frac{I(A_i = j)}{\widehat{\pi}_j(V_i)} \underbrace{\sum_{s \in \mathcal{T}_{\mathcal{J}}^{(j)} \cup \mathcal{F}_{\mathcal{J}}^{(j,V_i)} : s \leq t} G_{i,j}(s)}_{(\star)}
\end{aligned}
$$

$$
\begin{aligned}
&G_{i,j}(s) \\
&= \underbrace{\frac{\left[\hat{p}_{j,V_i}^C(s)\right]^{-1} n_{\mathcal{J}}^{(j,V_i)}(s)}{\widehat{S}_j(s)\left[1 - d\widehat{\Lambda}_j(s)\right]}\left[d\widehat{\Lambda}_{j,V_i}(s) - d\widehat{\Lambda}_j(s)\right]}_{O_{j,V_i}(s)} \left\{ \underbrace{\frac{(1-\delta_i)}{n_{\mathcal{J}}^{(j,V_i)}(Y_i) - c_{\mathcal{J}}^{(j,V_i)}(Y_i)}}_{r_{i,j}} I(Y_i \leq s) - \widehat{\sigma}^2_{\widehat{\Lambda}^C_{j,V_i}}(\min(s, Y_i)) \right\}.
\end{aligned}
$$

The above $(\star)$ can be simplified significantly for computational purposes by noting that we can take the summation over all observed follow-up times in the dataset (call them $s \in \mathcal{Y}^{(j)}$), and have the following:

$$
\begin{aligned}
&(\star) \\
&= -\left(\sum_{s \in \mathcal{Y} : s \leq t} O_{j,V_i}(s) \widehat{\sigma}^2_{\widehat{\Lambda}^C_{j,V_i}}(s) \right) \\
&\qquad + I(Y_i < t) \left[\left(r_{i,j} - \widehat{\sigma}^2_{\widehat{\Lambda}^C_{j,V_i}}(Y_i) \right) \left(\sum_{s \in \mathcal{Y} : s \leq \min(Y_i, t)} O_{j,V_i}(s) \right) + \left(\sum_{s \in \mathcal{Y} : s \leq \min(Y_i, t)} O_{j,V_i}(s) \widehat{\sigma}^2_{\widehat{\Lambda}^C_{j,V_i}}(s) \right) \right].
\end{aligned}
$$

Evaluating the bottom part only at $I(Y_i < t)$ in the expression above is crucial because $r_{i,j}$ and $\widehat{\sigma}^2_{\widehat{\Lambda}^C_{j,V_i}}(Y_i)$ are indefinite for some $Y_i \geq t$ if it's at the very end of follow-up. Finally, the estimated variance of the survival function is given by:

$$
\widehat{\mathrm{Var}}(\widehat{S}_j^{\mathrm{IPW}(3)}(t)) = n(\mathcal{I}_{\mathcal{J}})^{-1} \widehat{\mathrm{Var}}_{\mathcal{J}}(\widehat{\phi}_j^{(3)}(W_i)).
$$

Double Weighting with Estimated Treatment Propensities: Theorem 3.1 (iv)

For estimated treatment propensity weights, we have a very similar decomposition as in part (iii). The influence function for part (a) can be written as

$$(a) = \mathbb{P}_{\mathcal{J}}\phi_a^*(W_i)(s)$$
$$\phi_a^*(W_i) : s \to I(A_i = j)\pi_j(Z_i)^{-1}p_{j,V_i}^C(s)^{-1}dL_j(s)$$

where $p_{j,V}^C(s) = P_{\mathcal{J}}(C^{(j)} \geq s|V)$. After transforming to the survival scale applying Lemma C.3), the contribution to the influence function from part (a) is given by:

$$\phi_j^{(1)}(W)(t) = S_j(t)\int_0^t \frac{I(A=j)\pi_j(H)^{-1}p_{j,V}^C(s)^{-1}dL_j(s)}{P_{\mathcal{J}}(p_{j,V}^C(s)R_j(s))\left\{1-\Delta\Lambda_j(s)\right\}}.$$

For term (b), we have:

$$\begin{aligned}(b) &= E_{\mathcal{J}}\left\{I(A=j)dL_j(s)\left[\frac{1}{\hat{\pi}_j(H)\hat{p}_j^C(V)(s)} - \frac{1}{\pi_j(H)p_j^C(V)(s)}\right]\right\} \\ &= E_{\mathcal{J}}\left\{I(A=j)dL_j(s)\left[\frac{1}{\hat{\pi}_j(H)}\left(\frac{1}{\hat{p}_j^C(V)(s)} - \frac{1}{p_j^C(V)(s)}\right) + \frac{1}{p_j^C(V)(s)}\left(\frac{1}{\hat{\pi}_j(H)} - \frac{1}{\pi_j(H)}\right)\right]\right\}\end{aligned}$$

where $p_j^C(V)(s)$ is the limit of $\hat{p}_j^C(V)(s)$, which we know is $P_{\mathcal{J}}(C^{(j)} \geq s|V_i)$ since it is a (uniformly) consistent estimator. Using the conditional independence of $A \perp V|H$ from Definition 3.3 and Assumption 3.2:

$$\begin{aligned}(b) = &\underbrace{\sum_{h=1}^{L} P_{\mathcal{J}}(H=h)\left(\frac{\pi_j(h)}{\hat{\pi}_j(h)}\right)E_{\mathcal{J}}\left\{dL_j(s)\left(\frac{1}{\hat{p}_j^C(V)(s)} - \frac{1}{p_j^C(V)(s)}\right)\middle| H=h\right\}}_{d\gamma^{(b1)}} \\ &+ \underbrace{\sum_{h=1}^{L} P_{\mathcal{J}}(H=h)\left(\frac{\pi_j(h)}{\hat{\pi}_j(h)} - \frac{\pi_j(h)}{\pi_j(h)}\right)E_{\mathcal{J}}\left\{\frac{dL_j(s)}{p_j^C(V)(s)}\middle| H=h\right\}}_{d\gamma^{(b2)}}.\end{aligned}$$

Since $\widehat{\pi}_j(h) \to \pi_j(h)$ for every level h, and since $\hat{p}_j^C(V)(s)^{-1} - p_j^C(V)(s)^{-1}$ converges in distribution at rate $1/\sqrt{n^*}$, we can replace $\pi_j(h)/\hat{\pi}_j(h)$ with 1 in the first term (since the remainder is asymptotically negligible). The component of the influence function for part (b) is then:

$$d\gamma^{(b1)}(W_i)(s) = \sum_{h=1}^{L} P_{\mathcal{J}}(H=h) \sum_{v=1}^{S} P_{\mathcal{J}}(V^S = v|H=h) E_{\mathcal{J}}\left\{ dL_j(s)\phi_{j,v,i}^C(V^P)(s) \middle| H=h, V^S=v \right\}.$$

where the expectation is taken over $(dL_j(s), V^P)$, rather than the i index of the influence function $\phi_{j,v,i}^C$. Plugging in the results of Lemma C.5, we have that:

$$\begin{aligned} & d\gamma^{(b1)}(W_i)(s) \\ & = -\sum_{v=1}^{S}\sum_{h=1}^{L} \frac{P_{\mathcal{J}}(H=h|V^S=v)}{P_{\mathcal{J}}(A=j|V^S=v)} I(A_i=j) I(V_i^S=v) \Bigg[e_{j,v}^{(1)}(s)\xi_{j,v}^{(1)}(\delta_i, Y_i, V_i^P) \\ & \quad + \int_0^s e_{j,v}^{(2)}(s,u)\xi_{j,v}^{(2)}(\delta_i, Y_i, V_i^P)(u) \Bigg] \end{aligned}$$

where $\xi_{j,v}^{(1)}(\cdot)$ and $\xi_{j,v}^{(2)}(\cdot)(t)$ are defined in Lemma C.5, and

$$\begin{aligned} e_{j,v}^{(1)}(s) &= E_{\mathcal{J}(h,v)}\left\{ p_{j,v}^C(V^P)(s)^{-1} dL_j(s) \left[\int_0^s \frac{\left[V^P - c_{j,v}^{(1)}(u)\right] d\Lambda_{0j,v}^C(s)}{1 - \Delta\Lambda_{j,v}^C(s)} \right] \Omega_{j,v}^{-1} \right\} \\ e_{j,v}^{(2)}(s,u) &= E_{\mathcal{J}(h,v)}\left\{ p_{j,v}^C(V^P)(s)^{-1} dL_j(s) \left[\frac{\exp(\beta_{j,v}^T V^P)}{\left[1 - \Delta\Lambda_{j,v}^C(u)\right] r_{j,v}^{(0)}(u)} \right] \right\} \end{aligned}$$

where $\Omega_{j,v}$ and $c_{j,v}^{(1)}(u)$ are defined in Lemma C.5, and $E_{\mathcal{J}(h,v)}(\cdot) = E_{\mathcal{J}}(\cdot|H=h, V^S=s)$. We can rewrite the portion with $\xi_{j,v}^{(2)}$, since it is a function of all of the jumps within the censoring process, as:

$$\int_0^s e_{j,v}^{(2)}(s,u)\xi_{j,v}^{(2)}(\delta_i, Y_i, V_i^P)(u) = (1-\delta_i) e_{j,v}^{(2)}(s, Y_i) + \exp(\beta_{j,v}^T V_i^P) \sum_{u \in \mathcal{F}_{j,v}^C : u \le Y_i} e_{j,v}^{(2)}(s,u)$$

where $\mathcal{F}^C_{j,v} = \{Y_i : \delta_i = 0, V_i^S = v, A_i = j, I_{\mathcal{J}}(Z_i) = 1\}$ is the set of censoring times within the ECE population for strata v and treatment group j. For term (ii), we use the influence function of $\widehat{\pi}_j(h)$ that we obtained in Lemma C.8:

$$\begin{aligned} d\gamma^{(b2)}(W_i)(s) &= \sum_{h=1}^{L} P_{\mathcal{J}}(H=h)\pi_j(h)\phi^A_{j,h}(W_i)E_{\mathcal{J}}\left[\frac{dL_j(s)}{p^C_j(V)(s)}\middle| H=h\right] \\ &= -\sum_{h=1}^{L}\frac{I(H=h)\left[I(A=j)-\pi_j(h)\right]}{\pi_j(h)}E_{\mathcal{J}}\left[\frac{dL_j(s)}{p^C_j(V)(s)}\middle| H=h\right] \end{aligned}$$

Therefore, the full influence function is given by:

$$\phi^{(4)}_j(W)(t) = \phi^{(1)}_j(W)(t) + S_j(t)\int_0^t \frac{d\gamma^{(b1)}(W)(s) + d\gamma^{(b1)}(W)(s)}{P_{\mathcal{J}}(p^C_{j,V}(s)R_j(s))\left\{1-\Delta\Lambda_j(s)\right\}}.$$

We now focus on variance estimation when $V^P = \emptyset$. To estimate the $\phi^{(1)}_j(W)(t)$, we use the same strategy as in part (iii), with the difference being that the propensity weights are estimated (including in the calculation of the $\widehat{n}^{(j)}_{\mathcal{J}}$ and $\widehat{d}^{(j)}_{\mathcal{J}}$ terms that also influence $\widehat{\sigma}^2_{\widehat{\Lambda}_j}(t)$), and that we use the estimated survival function based on the estimated treatment propensity and inverse probability of censoring weights $\widehat{S}^{(4)}_j(t)$. Term $d\gamma^{(b1)}$ is similar to $d\gamma^{(b)}$ from part (iii), with the difference being that the conditional expectation of the Martingale term is taken with respect to the joint levels of (V, H):

$$\begin{aligned} &d\gamma^{(b1)}(W)(s) \\ &= -\frac{I(A_i=j)}{\pi_j(V_i)p^C_{j,V_i}(s)}\sum_{h=1}^{L} P_{\mathcal{J}}(H=h|V=V_i)E_{\mathcal{J}}\left\{dL_j(s)|V=V_i,H=h\right\}M_i(s) \\ &= -\frac{I(A_i=j)}{\pi_j(V_i)p^C_{j,V_i}(s)}\sum_{h=1}^{L} P_{\mathcal{J}}(H=h|V=V_i)E_{\mathcal{J}}\left\{R_j(s)|V=V_i,H=h\right\}\left[d\Lambda_{j,h,V_i}(s) - d\Lambda_j(s)\right]M_i(s) \end{aligned}$$

where $M_i(s)$ term as in part (iii), and $\Lambda_{j,h,v}(s)$ is the cumulative hazard function for treatment j, propensity strata h, and censoring strata v.

Let $n^{(j,h,v)}(s)$ and $d^{(j,h,v)}(s)$ be the risk set size and event count, respectively, for

treatment group j, treatment strata h, and censoring strata v at time s. To estimate $E_{\mathcal{J}}\left\{R_j(s)|V=V_i,H=h\right\}$, we can use $n(I_{\mathcal{J}})^{-1}\pi_j(H_i)^{-1}P_{\mathcal{J}}(H=h|V_i=v)P_{\mathcal{J}}(V=V_i)n^{(j,h,V_i)}(s)$. Several of these terms will cancel with $\widehat{M}_i(s)$. Then, we have the following estimate of $d\gamma^{(b1)}(s)$:

$$\begin{aligned}
d\widehat{\gamma}^{(b1)}(s) &= -\frac{I(A_i=j)}{\widehat{\pi}_j(V_i)\widehat{p}^C_{j,V_i}(s)}\sum_{h=1}^{L} n(I_{\mathcal{J}})^{-1}\widehat{\pi}_j(H_i)^{-1}n^{(j,h,V_i)}(s)\left[\frac{d^{(j,h,V_i)}_{\mathcal{J}}(s)}{n^{(j,h,V_i)}_{\mathcal{J}}(s)}-\frac{\widehat{d}^{(j)}_{\mathcal{J}}(s)}{\widehat{n}^{(j)}_{\mathcal{J}}(s)}\right]\widehat{M}_i(s)\\
&= -\frac{I(A_i=j)}{\widehat{\pi}_j(V_i)\widehat{p}^C_{j,V_i}(s)}\sum_{h=1}^{L} n^{(j,h,V_i)}(s)\left[\frac{d^{(j,h,V_i)}_{\mathcal{J}}(s)}{n^{(j,h,V_i)}_{\mathcal{J}}(s)}-\frac{\widehat{d}^{(j)}_{\mathcal{J}}(s)}{\widehat{n}^{(j)}_{\mathcal{J}}(s)}\right]\\
&\quad\times\left[\frac{(1-\delta_i)I(Y_i\le s)}{n^{(j,V_i)}_{\mathcal{J}}(Y_i)-c^{(j,V_i)}_{\mathcal{J}}(Y_i)}-\widehat{\sigma}^2_{\widehat{\Lambda}^C_{j,V_i}}(\min(s,Y_i))\right].
\end{aligned}$$

For $d\gamma^{(b2)}$ we first simplify it using $V^P=\emptyset$:

$$\begin{aligned}
&d\gamma^{(b2)}(W_i)(s)\\
&= -\sum_{v=1}^{S}\frac{I(A_i=j)-\pi_j(H_i)}{\pi_j(H_i)p^C_{j,v}(s)}P_{\mathcal{J}}(V=v|H=H_i)E_{\mathcal{J}}\left[dL_j(s)|H=H_i,V=v\right]\\
&= -\sum_{v=1}^{S}\frac{I(A_i=j)-\pi_j(H_i)}{\pi_j(H_i)p^C_{j,v}(s)}P_{\mathcal{J}}(V=v|H=H_i)E_{\mathcal{J}}\left\{R_j(s)|V=v,H=H_i\right\}\left[d\Lambda_{j,H_i,v}(s)-d\Lambda_j(s)\right].
\end{aligned}$$

To estimate this portion, we can use a similar strategy to $d\gamma^{(b1)}$, but notably do not include the censoring influence function $M_i(s)$:

$$\begin{aligned}
&d\widehat{\gamma}^{(b2)}(W_i)(s)\\
&= -\sum_{v=1}^{S}\frac{I(A_i=j)-\widehat{\pi}_j(H_i)}{\widehat{\pi}_j(H_i)\widehat{p}^C_{j,v}(s)}n(I_{\mathcal{J}})^{-1}\widehat{\pi}_j(H_i)^{-1}\widehat{P}_{\mathcal{J}}(H=H_i)^{-1}n^{(j,H_i,v)}(s)\left[d\Lambda_{j,H_i,v}(s)-d\Lambda_j(s)\right]\\
&= -n(I_{\mathcal{J}})^{-1}\sum_{v=1}^{S}\frac{I(A_i=j)-\widehat{\pi}_j(H_i)}{\widehat{\pi}^2_j(H_i)\widehat{p}^C_{j,v}(s)\widehat{p}(H_i)}n^{(j,H_i,v)}(s)\left[d\widehat{\Lambda}_{j,H_i,v}(s)-d\widehat{\Lambda}_j(s)\right].
\end{aligned}$$

Therefore, the final estimated influence function is:

$$\begin{aligned}\widehat{\phi}^{(4)}(W_i)(t) &= \widehat{\phi}^{(1)}(W_i)(t) + \widehat{S}_j^{(4)}(t) \cdot \int_0^t \frac{d\widehat{\gamma}^{(b1)}(W_i)(s) + d\widehat{\gamma}^{(b2)}(W_i)(s)}{\widehat{S}_j(s)\left[1 - d\widehat{\Lambda}_j(s)\right]} \\ &= \widehat{\phi}^{(1)}(W_i)(t) - \widehat{S}_j^{(4)}(t) \cdot \frac{I(A_i = j)}{\widehat{\pi}_j(V_i)} \sum_{s \in \mathcal{T}_{\mathcal{J}}^{(j)} \cup \mathcal{F}_{\mathcal{J}}^{(j,V_i)} : s \leq t} \left(\sum_{h=1}^{L} G_{i,j}^{(h)}(s) \right) \\ &\quad - \widehat{S}_j^{(4)}(t) \cdot n(I_{\mathcal{J}})^{-1} \left[\frac{I(A_i = j) - \widehat{\pi}_j(H_i)}{\widehat{\pi}_j^2 \widehat{p}(H_i)} \right] \sum_{s \in \mathcal{T}_{\mathcal{J}}^{(j)} : s \leq t} \left(\sum_{v=1}^{S} K_{i,j}^{(v)} \right) \\ K_{i,j}^{(v)} &= \left[\hat{p}_{j,v}^{C}(s)\right]^{-1} \frac{n_{\mathcal{J}}^{(j,H_i,v)}(s)}{\widehat{S}_j(s)\left[1 - d\widehat{\Lambda}_j(s)\right]} \left[d\widehat{\Lambda}_{j,H_i,v}(s) - d\widehat{\Lambda}_j(s) \right],\end{aligned}$$

where $G^{(h)}$ is the same as G in part (iii), but where the $O_{j,V_i}(s)$ is replaced by $O_{j,h,V_i}(s)$ below, and the $r_{i,j}$ is replaced by the $r_{i,j,h}$ below:

$$\begin{aligned}O_{j,h,V_i}(s) &= \frac{\left[\hat{p}_{j,V_i}^{C}(s)\right]^{-1} n_{\mathcal{J}}^{(j,h,V_i)}(s)}{\widehat{S}_j(s)\left[1 - d\widehat{\Lambda}_j(s)\right]} \left[d\widehat{\Lambda}_{j,h,V_i}(s) - d\widehat{\Lambda}_j(s) \right] \\ r_{i,j,h} &= \frac{(1 - \delta_i)}{n_{\mathcal{J}}^{(j,h,V_i)}(Y_i) - c_{\mathcal{J}}^{(j,h,V_i)}(Y_i)}.\end{aligned}$$

Finally, the estimated variance of the survival function is given by:

$$\widehat{\mathrm{Var}}(\widehat{S}_j^{\mathrm{IPW}(4)}(t)) = n(\mathcal{I}_{\mathcal{J}})^{-1} \widehat{\mathrm{Var}}_{\mathcal{J}}(\widehat{\phi}_j^{(4)}(W_i)).$$

C.2.3 Proof of Theorem 3.2

Lemma C.8. *With H from Definition 3.3, we have the following result:*

$$E_{\mathcal{J}}(I(A = j)|H = h) = \pi_j(h).$$

Proof. The proof relies on Definition 3.3 that $\pi_j(Z) = \pi_j(H)$:

$$\begin{aligned}
E_{\mathcal{J}}(I(A=j)|H=h) &= E(I(A=j)|H=h, I_{\mathcal{J}}(Z)=1) \\
&= E(I(A=j)I_{\mathcal{J}}(Z)|H=h)P(I_{\mathcal{J}}(Z)=1|H=h)^{-1} \\
&= E(E(I(A=j)I_{\mathcal{J}}(Z)|H=h, Z=z)|H=h)P(I_{\mathcal{J}}(Z)=1|H=h)^{-1} \\
&= E(I_{\mathcal{J}}(Z)E(I(A=j)|H=h, Z=z)|H=h)P(I_{\mathcal{J}}(Z)=1|H=h)^{-1} \\
&= E(I_{\mathcal{J}}(Z)\pi_j(z)|H=h)P(I_{\mathcal{J}}(Z)=1|H=h)^{-1} \\
&= \pi_j(h)E(I_{\mathcal{J}}(Z)|H=h)P(I_{\mathcal{J}}(Z)=1|H=h)^{-1} \\
&= \pi_j(h).
\end{aligned}$$

□

Proof of Theorem 3.2. We aim to derive the asymptotic behavior of the post-stratification estimator, given by (3.6), and which can be expressed as:

$$\hat{S}_j^{\text{PSS}}(t) = \frac{1}{n\left(\mathcal{I}_{\mathcal{J}}\right)} \sum_{h=1}^{L} n\left(\mathcal{I}_{\mathcal{J}}^{(h)}\right) \hat{S}_{j,h}(t)$$

where $\hat{S}_{j,h}(t) = \prod_{t_k \leq t}\left[1 - \frac{\sum_{i \in \mathcal{I}_{\mathcal{J}}^{(h)}: T_i = t_k} I(A_i=j)\delta_i}{\sum_{i \in \mathcal{I}_{\mathcal{J}}^{(h)}: T_i \geq t_k} I(A_i=j)}\right]$. First, we derive the asymptotic distribution of $\hat{S}_{j,h}(t)$, the strata-specific survival function. Then, since each strata contains independent observations, we can simply use the delta method to form a weighted average of the strata-specific survival functions. Using the same logic as in the proof of Theorem 3.1, we have the following expansion:

$$\hat{\Lambda}_{j,h}(t) - \Lambda_{j,h}(t) = \int_0^t \frac{\mathbb{P}_{\mathcal{J}} I(H=h)I(A=j)dN_j(s)}{\mathbb{P}_{\mathcal{J}} I(H=h)I(A=j)Y_j(s)} - \int_0^t d\Lambda_j(s)$$

where $\Lambda_{j,h}(s) = \frac{P_{\mathcal{J}} I(H=h)N_j(s)}{P_{\mathcal{J}} I(H=h)R_j(s)}$, which, dividing the top and bottom by $P_{\mathcal{J}}(H=h)$, can also be written as $E(N_j(s)|H=h, I_{\mathcal{J}}(Z)=1)/E(R_j(s)|H=h, I_{\mathcal{J}}(Z)=1)$. In other words,

it's the cumulative hazard in the ECE population, but conditional on having stratification variable $H = h$. Using the same expansion as in Theorem 3.1, we find that the influence function is given by:

$$\hat{\Lambda}_{j,h}(t) - \Lambda_{j,h}(t) = \int_0^t \frac{\mathbb{P}_{\mathcal{J}} I(A = j) I(H = h) dL_j(s)}{P_{\mathcal{J}} I(A = j) I(H = h) R_j(s)} + o_P(1/\sqrt{n^*})$$

with $dL_{j,h}(s) = dN_j(s) - R_j(s) d\Lambda_{j,h}(s)$. Because $N_j(s)$ and $R_j(s)$ are independent of A conditional on H (by Definition 3.3), the population quantity for the numerator, $P_{\mathcal{J}} I(A = j) I(H = h) dL_{j,h}(s) = 0$, so we can subtract it to show that the influence function is given by:

$$\begin{aligned}
\phi_h(H, A, N_j, Y_j)(t) &= \int_0^t \frac{I(A = j) I(H = h) dL_{j,h}(s)}{P_{\mathcal{J}} I(A = j) I(H = h) R_j(s)} \\
&= \int_0^t \frac{I(A = j) I(H = h) dL_{j,h}(s)}{E_{\mathcal{J}}(R_j(s) | H = h) E_{\mathcal{J}}(I(A = j) | H = h) P_{\mathcal{J}}(H = h)} \\
&= \int_0^t \frac{I(A = j) I(H = h) dL_{j,h}(s)}{E_{\mathcal{J}}(R_j(s) | H = h) \pi_j(h) P_{\mathcal{J}}(H = h)}
\end{aligned}$$

where we have used Lemma C.8 to go from the second to third lines. Applying Lemma C.3, we have that:

$$\begin{aligned}
&\hat{S}_{j,h}(t) - S_{j,h}(t) \\
&= \frac{S_{j,h}(t)}{\pi_j(h) P_{\mathcal{J}}(H = h)} \mathbb{P}_{\mathcal{J}} I(A = j) I(H = h) \int_0^t \frac{dL_{j,h}(s)}{E_{\mathcal{J}}(R_j(s) | H = h) \left(1 - \Delta\Lambda_{j,h}(s)\right)} + o_P(1/\sqrt{n^*})
\end{aligned}$$

where $\Delta\Lambda_{j,h}(t) = \Lambda_{j,h}(t) - \Lambda_{j,h}(t-)$, the point mass in the hazard at t. To obtain $\hat{S}_j$, we can

do a simple linear transformation:

$$
\begin{aligned}
&\hat{S}_j^{\mathrm{PSS}}(t) - S_j(t) \\
&= \sum_{h=1}^{L} \mathbb{P}_{\mathcal{J}}(H=h)\hat{S}_{j,h}(t) - \sum_{h=1}^{L} P_{\mathcal{J}}(H=h)S_{j,h}(t) \\
&= \sum_{h=1}^{L} P_{\mathcal{J}}(H=h)\left\{\hat{S}_{j,h}(t) - S_{j,h}(t)\right\} + \sum_{h=1}^{L} S_{j,h}(t)\left\{\mathbb{P}_{\mathcal{J}}(H=h) - P_{\mathcal{J}}(H=h)\right\} + o_P(1/\sqrt{n^*})
\end{aligned}
$$

since $\mathbb{P}_{\mathcal{J}}(H=h) - P_{\mathcal{J}}(H=h) = o_P(1)$ and since $\hat{S}_{j,h}(t) - S_{j,h}(t)$ is asymptotically normal (by Lemma C.3) at the $O_P(1/\sqrt{n^*})$ rate. Note that $\mathbb{P}_{\mathcal{J}}(H=h) - P_{\mathcal{J}}(H=h) = \mathbb{P}_{\mathcal{J}}\left(I(H=h) - P_{\mathcal{J}}(H=h)\right)$, so we have

$$
\begin{aligned}
&\hat{S}_j^{\mathrm{PSS}}(t) - S_j(t) \\
&= \mathbb{P}_{\mathcal{J}}\phi_j^{\mathrm{PSS}}(t) + o_P(1/\sqrt{n^*}) \\
\phi_j^{\mathrm{PSS}} &= \sum_{h=1}^{L} S_{j,h}(t)\left\{\pi_j(h)^{-1}\int_0^t \frac{I(A=j)I(H=h)dL_{j,h}(s)}{E_{\mathcal{J}}(R_j(s)|H=h)\left(1-\Delta\Lambda_{j,h}(s)\right)} + I(H=h) - P_{\mathcal{J}}(H=h)\right\} \\
&= \underbrace{\sum_{h=1}^{L} S_{j,h}(t)\left\{\pi_j(h)^{-1}\int_0^t \frac{I(A=j)I(H=h)dL_{j,h}(s)}{E_{\mathcal{J}}(R_j(s)|H=h)\left(1-\Delta\Lambda_{j,h}(s)\right)}\right\}}_{(I)} + \underbrace{\left\{\sum_{h=1}^{L} I(H=h)S_{j,h}(t) - S_j(t)\right\}}_{(II)}.
\end{aligned}
$$

Term (I) and term (II) both have mean zero. Then to find the asymptotic variance of the PSS estimator evaluated at some t, we have $E_{\mathcal{J}}(\phi_j^{\mathrm{PSS}}(t)\phi_j^{\mathrm{PSS}}(t)) = E_{\mathcal{J}}((I)^2) + E_{\mathcal{J}}((I)(II)) +$

$E_{\mathcal{J}}((II)^2)$. The cross-terms in (I) and in (I) times (II) are zero,

$$E_{\mathcal{J}}((I)^2) = \sum_{h=1}^{L} S_{j,h}(t)^2 \pi_j(h)^{-1} P_{\mathcal{J}}(H=h) E_{\mathcal{J}} \left\{ \int_0^t \frac{dL_{j,h}(s)}{E_{\mathcal{J}}(R_j(s)|H=h)\,(1-\Delta\Lambda_{j,h}(s))} \middle| H=h \right\}^2$$

$$E_{\mathcal{J}}((I)(II)) = 0$$

$$\begin{aligned} E_{\mathcal{J}}((II)^2) &= \sum_{h=1}^{L} S_{j,h}(t)^2 P_{\mathcal{J}}(H=h) - S_j^2(t) \\ &= \sum_{h=1}^{L} S_{j,h}(t) P_{\mathcal{J}}(H=h) \left\{ S_{j,h}(t) - S_j(t) \right\} \end{aligned}$$

Since $L_{j,h}$ is a martingale, and since $f(t) := E_{\mathcal{J}}(R_j(s)|H=h)\,(1-\Delta\Lambda_{j,h}(s))$ is a deterministic function (and thus it is trivially predictable), we can apply the results of Theorem 2.6.2 in Fleming and Harrington (2011), which is also done in Wang et al. (2021) to calculate the variance of the stochastic integral above. It simplifies to:

$$\begin{aligned} \mathrm{Var}_{\mathcal{J}} \left(\int_0^t f(s)^{-1} dL_{j,h}(s) \right) &= \int_0^t f(s)^{-2} E_{\mathcal{J}}(R_j(s)|H=h) \left\{ 1-\Delta\Lambda_{j,h}(s) \right\} d\Lambda_{j,h}(s) \\ &= \int_0^t \frac{d\Lambda_{j,h}(s)}{E_{\mathcal{J}}(R_j(s)|H=h) \left\{ 1-\Delta\Lambda_{j,h}(s) \right\}}. \end{aligned}$$

Therefore, the final asymptotic variance is given by:

$$\Sigma^{\mathrm{PSS}}(t) = \sum_{h=1}^{L} S_{j,h}(t) P_{\mathcal{J}}(H=h) \left\{ \frac{S_{j,h}(t)}{\pi_j(h)} \int_0^t \frac{d\Lambda_{j,h}(s)}{E_{\mathcal{J}}(R_j(s)|H=h) \left\{ 1-\Delta\Lambda_{j,h}(s) \right\}} + S_{j,h}(t) - S_j(t) \right\}.$$

Let $\mathcal{T}_{\mathcal{J}}^{(j,h)} = \{Y_i : \delta_i = 1, A_i = j\}_{i \in \mathcal{I}_{\mathcal{J}}^{(h)}}$ be the set of unique failure times for those within the ECE sample and assigned to treatment j and strata h. Let $d_{\mathcal{J}}^{(j,h)}(t) = \sum_{i \in \mathcal{I}_{\mathcal{J}}^{(h)}} \delta_i I(Y_i = t) I(A_i = j)$ and $n_{\mathcal{J}}^{(j,h)}(t) = \sum_{i \in \mathcal{I}_{\mathcal{J}}^{(h)}} I(Y_i \geq t) I(A_i = j)$. Then a consistent estimator of the

asymptotic variance is given by:

$$
\begin{aligned}
&\widehat{\Sigma}^{\mathrm{PSS}}(t) \\
&= \sum_{h=1}^{L} \hat{p}_{\mathcal{J}}(h) \hat{S}_{j,h}(t) \left[\hat{S}_{j,h}(t) - \hat{S}_j^{\mathrm{PSS}}(t) \right] \\
&\quad + \sum_{h=1}^{L} \frac{\hat{S}_{j,h}(t)^2 \hat{p}_{\mathcal{J}}(h)}{\hat{\pi}_j(h)} \cdot n(\mathcal{I}_{\mathcal{J}}) \cdot \hat{p}_{\mathcal{J}}(h) \cdot \hat{\pi}_j(h) \sum_{s \in \mathcal{T}_{\mathcal{J}}^{(j,h)}: s \le t} \frac{d_{\mathcal{J}}^{(j,h)}(s)}{n_{\mathcal{J}}^{(j,h)}(s) \left[n_{\mathcal{J}}^{(j,h)}(s) - d_{\mathcal{J}}^{(j,h)}(s) \right]} \\
&= \sum_{h=1}^{L} \hat{p}_{\mathcal{J}}(h) \hat{S}_{j,h}(t) \left[\hat{S}_{j,h}(t) - \hat{S}_j^{\mathrm{PSS}}(t) \right] \\
&\quad + n(\mathcal{I}_{\mathcal{J}}) \sum_{h=1}^{L} \hat{S}_{j,h}(t)^2 \left[\hat{p}_{\mathcal{J}}(h) \right]^2 \sum_{s \in \mathcal{T}_{\mathcal{J}}^{(j,h)}: s \le t} \frac{d_{\mathcal{J}}^{(j,h)}(s)}{n_{\mathcal{J}}^{(j,h)}(s) \left[n_{\mathcal{J}}^{(j,h)}(s) - d_{\mathcal{J}}^{(j,h)}(s) \right]},
\end{aligned}
$$

where $\hat{S}_{j,h}(t)$ is the Kaplan-Meier estimator in (3.2) using only individuals within the strata h ($H_i = h$), $\hat{p}_{\mathcal{J}}(h) = n\left(\mathcal{I}_{\mathcal{J}}^{(h)}\right) n(\mathcal{I}_{\mathcal{J}})^{-1}$, $\hat{\pi}_j(h) = \sum_{i=1}^{n} I(A_i = j) I(H_i = h) / \sum_{i=1}^{n} I(H_i = h)$. And consistent variance estimator is given by:

$$
\widehat{\mathrm{Var}}(\widehat{S}_j^{\mathrm{PSS}}(t)) = n(\mathcal{I}_{\mathcal{J}})^{-1} \widehat{\Sigma}^{\mathrm{PSS}}(t).
$$

□