Win Ratio Analysis for Composite Outcomes in Observational Cardiovascular Studies: An Application to the Multi-Ethnic Study of Atherosclerosis and Simulation Study

Robert Trujillo

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

Master of Science

University of Washington

2026

Committee:

Robyn McClelland

Spencer Hansen

Program Authorized to Offer Degree:
Department of Biostatistics

University of Washington

Abstract

Win Ratio Analysis for Composite Outcomes in Observational Cardiovascular Studies: An Application to the Multi-Ethnic Study of Atherosclerosis and Simulation Study

Robert Trujillo

Chair of the Supervisory Committee:
Robyn McClelland
Department of Biostatistics

The win ratio is a method for analyzing composite hierarchical outcomes that respect a pre-specified clinical priority among component events. Despite growing adoption in cardiovascular trials, its use in observational studies remains virtually unexplored. This thesis evaluates the win ratio and several of its extensions for use with observational data through two investigations. In a simulation study, we compare the hierarchical analysis of the win ratio to the time-to-first-event analysis of the Cox model under varying levels of unmeasured confounding, censoring, and differential component effects. We tested the performance of various inverse-probability weighted win ratio estimators with regards to bias, coverage, type I error, and power. We also show that the win ratio and Cox model reach different conclusions depending on where the treatment effect is concentrated in the hierarchical outcome. In the applied analysis, we compare cardiovascular outcomes across sex and racial/ethnic subgroups in the Multi-Ethnic Study of Atherosclerosis (MESA) cohort using the win ratio and related methods. We construct hierarchical composites outcome endpoints of increasing complexity that capture a broad spectrum of cardiovascular disease, illustrating an example where incorporating baseline coronary artery calcium as a tie-breaking component reverses the direction of the win ratio effect. We provide one of the first applications of the win ratio to demographic comparisons in a large observational cohort study.

TABLE OF CONTENTS

LIST OF FIGURES

GLOSSARY

MESA: Multi-Ethnic Study of Atherosclerosis

TFE: Time-to-First-Event

HR: Hazard Ratio

WR: Win Ratio

WO: Win Odds

NB: Net Benefit

IPTW: Inverse-Probability of Treatment Weighting

IPCW: Inverse-Probability of Censoring Weighting

CVD: Cardiovascular Disease

MI: Myocardial Infarction

RCA: Resuscitated Cardiac Arrest

TIA: Transient Ischemic Attack

CAC: Coronary Artery Calcium

CABG: Coronary Atery Bypass Graft

EGC: Electrocardiogram

ACKNOWLEDGMENTS

I would like to first thank my advisor and reading committee chair, Robyn. I've learned an enormous amount from her over the past three years, and my time with her as an analyst for MESA has allowed me to build valuable skills and has shaped who I am as an applied statistician. Her patience, guidance, and wisdom have carried me through the ups and downs of graduate school and helped me grow. She also introduced me to win statistics and encouraged me to present my work at a conference.

I'm also grateful to my other reading committee member, Spencer, who helped me get up to speed with the MESA data, sharpen my coding, think through statistical problems more carefully, and become a better statistician. I would also like to thank Craig for his guidance and support, especially during my conference poster presentation. I'm also thankful for my supportive labmates, Quinn, Jingyi, and Will, who I learned a great deal from and who made lab meetings something I looked forward to. I am grateful to ARCS for helping fund my graduate education and for supporting me through the difficult parts of this journey. I also want to thank the Biostatistics Department, especially Minh, for their support during some of the harder moments of my time in graduate school.

I also want to thank the people who made graduate school feel lighter and more meaningful. To the friends I made along the way, Nolan, Awan, Grant, Charlotte, Ellen, Nina, and many more: thank you for being the highlight of my time in the program and for your support both on and off campus. I also want to thank my friends Richard and Sheldine, and their pets Stella, Koko, and Merlin, for providing me refuge from school stress when needed. I'm also very grateful to my parents Robert and Blanca, and siblings Alejandro and Mariana, for their constant encouragement, affirmation, and love. Lastly, I am deeply grateful to my girlfriend Nancy for her unwavering love and support through every high and low of graduate school. I truly couldn't have made it through without her.

DEDICATION

To my family.

Chapter 1

INTRODUCTION

1.1 Composite Outcomes in Cardiovascular Studies

Composite endpoints that combine fatal events (e.g. cardiovascular death) with one or more non-fatal events (e.g. myocardial infarction, heart failure hospitalization) are commonly used as the primary outcome measure in randomized clinical trials in cardiovascular studies [1, 2]. The main advantage of using composite outcomes is that they increase the number of primary outcome events, which subsequently increases statistical power and reduces the required sample size [1,2]. Another benefit occurs when multiple outcomes may be clinically relevant. Combining these outcomes into a single composite avoids the need to select a single primary endpoint, and can provide a pragmatic approach to the problem of competing risks by avoiding the need to treat fatal events as a censoring event for non-fatal outcomes. Additionally, composite outcomes help to circumvent problems of multiplicity and increase risk of type I error that arise when multiple co-primary outcomes are assessed independently [2,3]. The standard approach to analyzing composite outcomes is to define the time-to-event of the composite as the time-to-first component event, that is, the time to the event of the component that happens the soonest. In this time-to-first-event (TFE) approach a Cox proportional hazards model is typically used for analysis, yielding a hazard ratio (HR) as the measure of treatment efficacy along with the accompanying confidence interval and p-value [2].

1.2 The Limits of Time-to-First-Event Analysis

The use of composite outcomes presents its own set of challenges. Under the TFE approach, once an individual experiences any component event then all other subsequent events are ignored. For example, a patient who is hospitalized and later dies contributes only the hospitalization to the analysis of the composite event [2, 4]. Components considered more

important by clinicians and patients, such as mortality, tend to have lower event rates and smaller treatment effects than components considered less important [2, 5]. Moreover, fatal events by definition can only occur after, or instead of, non-fatal events in the same individual. This creates an asymmetry where non-fatal events tend to dominate the attribution of the composite events, even if fatal and non-fatal events occur at similar overall frequencies in the study population [2, 5].

Another issue with composite endpoints is that they ignore any clinical hierarchy of importance among components. For example, under a TFE analysis a Cox model will give the same statistical weight to both cardiovascular death and non-fatal MI hospitalization, treating the component events as interchangeable [4, 6]. The resulting hazard ratio is a quantity that averages over component-specific effects. When treatment effects differ by component, this hazard ratio may be difficult to interpret clinically [4, 7]. Ferreira-González et al. (2007) found that in over half of the cardiovascular trials they reviewed, component-specific treatment effects varied substantially, with larger effects on less clinically important components [5]. In the extreme case, a treatment could reduce hospitalizations but increase mortality, yet still appear beneficial under a TFE analysis because hospitalizations dominate the composite event count [6, 7]. These limitations have motivated the development of alternative approaches to composite endpoint analysis that respect a clinical hierarchy of importance among the components.

1.3 The Win Ratio for Composite Outcomes in Clinical Trials

The win ratio was introduced by Pocock et al. (2012) as a method for analyzing composite endpoints that respects a pre-specified clinical hierarchy of importance among the components [6]. The method builds on earlier work by Finkelstein and Schoenfeld (1999) and Buyse (2010), who proposed joint tests on multiple outcomes using generalized pairwise comparisons [8, 9]. As opposed to the TFE approach using a Cox model, the win ratio approach takes into account the most important event in a composite outcome before considering subsequent events down the hierarchy. In the all-against-all paradigm, every subject in the treatment group is compared against every subject in the control group. For each pair, subjects are first compared on the most important outcome (e.g. time-to-death),

and the subject that experiences the event the latest, or not at all, is considered the winner. If a winner can't be determined (typically due to censoring) then the comparison proceeds to the next component (e.g. time-to-hospitalization). When components are exhausted without determining a winner, then the comparison is considered a tie [4,6]. The win ratio is then defined as the ratio of wins to losses for the treatment group, with values greater than 1 indicating a treatment benefit. Complementary quantities such as the net benefit (NB) and win odds (WO) provide absolute and odds-scale measures of the same underlying comparison, respectively, and together with the win ratio belong to a greater family of methods called win statistics [10].

Win statistics have been gaining popularity as both primary and secondary endpoints for cardiovascular clinical trials [4, 11]. For example, in the ATTR-ACT trial the win ratio served as the primary endpoint supporting the FDA approval of tafamidis for transthyretin amyloid cardiomyopathy [12]. In the EMPULSE trial, empagliflozin significantly improved clinical outcomes in patients hospitalized with acute heart failure through a win ratio analysis using a 4-component composite hierarchical endpoint [13].

Concurrently, there has been a substantial amount of theoretical development for win statistics methods over the past decade [10, 14–26]. Of note, a series of papers by Dong and Colleagues have extended the methodology to address concerns that commonly arise in real clinical trials such as non-proportional hazards, independent and dependent censoring, and stratified analyses [10, 16, 18–20]. The win ratio has been extended to include inverse probability weighting techniques that address bias from censoring and baseline covariate imbalance, and recent work by Mao and Wang (2021) includes the development of a covariate-adjusted regression framework. [19, 21, 25]

1.4 The Win Ratio for Composite Outcomes in Observational Studies

The rapid development of the win ratio literature has almost exclusively focused on their use in the context of randomized clinical trials. In this setting, randomization ensures that comparisons between treatment and control groups are unconfounded. That is, any systemic difference in outcomes between the groups can be attributed to the treatment. The use of the win ratio in observational studies is virtually unexplored and presents its

own additional challenges, particularly the presence of confounding due to non-random treatment assignment [23, 25].

Very few examples of the use of win statistics in observational settings exist. Chiaruttini et al. (2024) applied win statistics in an observational cancer study comparing treatment strategies for locally advanced rectal adenocarcinoma, using propensity score matching and stratification to address confounding [27]. Schmidt et al. (2024) performed a composite hierarchical analysis of death and heart failure hospitalization rates among patients with or without transthyretin amyloid cardiomyopathy using data from the SCAN-MP (Screening for Cardiac Amyloidosis With Nuclear Imaging in Minority Populations) study [28].

Despite the scarcity of real-world applications in observational studies, recent developments of the win ratio methodology lend themselves to addressing issues of bias from confounding. Dong et al. (2018) proposed a stratified win ratio that can partially address confounding through stratification when the number of strata is small [16]. Wang et al. (2025) proposed inverse probability of treatment weighting (IPTW) adjusted win ratios to balance baseline covariates. Zhang et al. (2022) placed the IPTW win ratio within a formal causal inference framework, and addressed clustered observational data using calibrated weights [23]. Mao and Wang (2021) developed proportional win-fractions regression, which allows covariate adjustment through a regression framework rather than weighting [21].

This thesis is motivated by the lack of real world applications of the win ratio in observational settings and by several questions that still remain open: 1) How does the IPTW-adjusted win ratio perform under unmeasured confounding? 2) How does the win ratio compare to the Cox model in observational settings with confounding where treatment effects may differ across fatal and non-fatal components? 3) Can the win ratio provide clinically useful insights when applied to demographic subgroup comparisons (e.g. race/ethnicity, sex) in observational cohort studies, where the "treatment" is a non-modifiable exposure? These questions are addressed through two complementary approaches. Firstly, we conduct a simulation study that evaluates the performance of the unadjusted win ratio, IPTW-adjusted win ratio, and Cox proportional hazards model under varying degrees of unmeasured confounding and censoring. Secondly, we illustrate the utility and careful application of win ratio methods, including inverse probability weighting techniques and proportional

win-fractions regression, to the Multi-Ethnic Study of Atherosclerosis (MESA) cohort for comparing cardiovascular outcomes across demographic subgroups in an observational setting [29].

Chapter 2

WIN STATISTICS METHODS

2.1 Win Ratio, Win Odds, and Net Benefit

In this chapter, we present the statistical framework for the win statistics methods used in both the simulation study and applied analysis. Following Luo et al. (2015), we formalize the all-against-all pairwise comparisons that define the win ratio using kernel functions [14]. For a treatment (T) subject i and a control (C) subject j, define their observed composite outcome by Y_i and Y_j. For example, suppose our composite endpoint consists of observed time to death $Y_{D,i}$ and hospitalization $Y_{H,i}$ with corresponding event indicators $\delta_{D,i}$ and $\delta_{H,i}$. The full observed composite outcome for subject i (and similarly for subject j) would be given by the vector $Y_i = (Y_{D,i}, Y_{H,i}, \delta_{D,i}, \delta_{H,i})$.

We define the respective win and loss kernels by:

$$\phi_1(Y_i, Y_j) = \mathbb{I}\{Y_i \text{ wins over } Y_j\}, \quad \phi_2(Y_i, Y_j) = \mathbb{I}\{Y_j \text{ wins over } Y_i\} \tag{2.1}$$

where $\mathbb{I}(\cdot)$ is the indicator function that is resolved by the within-pair comparison between subject i and j following the hierarchical procedure described in Section 1.3. The most clinically important component is compared first, and the comparison proceeds to the next component only if the higher priority component does not determine a winner. When a component is a continuous or ordinal measure rather than a time-to-event outcome, the investigators define what constitutes a win or loss [4]. If no component resolves the comparison, the pair is tied. Because these kernels are evaluated on the observed data, the frequency of ties depends on the censoring distribution. Heavier censoring produces more tied pairs since fewer comparisons can be resolved at any level of the hierarchy.

The population win and loss probabilities are defined by the expected value of the win and loss kernels:

$$\pi_T = E[\phi_1(Y_i, Y_j)], \quad \pi_C = E[\phi_2(Y_i, Y_j)] \tag{2.2}$$

The population win ratio is given by the ratio of the win and loss probabilities:

$$WR = \pi_T / \pi_C \tag{2.3}$$

where $WR > 1$ indicates treatment benefit, and a null value of $WR = 1$ indicates no treatment benefit. The sample win proportions can each be defined as U-statistics:

$$\widehat{\pi}_T = \frac{1}{N_T N_C} \sum_{i=1}^{N_T} \sum_{j=1}^{N_C} \phi_1(Y_i, Y_j), \quad \widehat{\pi}_C = \frac{1}{N_T N_C} \sum_{i=1}^{N_T} \sum_{j=1}^{N_C} \phi_2(Y_i, Y_j) \tag{2.4}$$

where N_T and N_C are the number of individuals in the sample for the treatment and control groups, respectively. The sample win ratio is defined as the ratio of these two U-statistics:

$$\widehat{WR} = \frac{\widehat{\pi}_T}{\widehat{\pi}_C} \tag{2.5}$$

This U-statistics representation has served the basis for many of the methodological developments of the win ratio in the literature, providing a robust framework for large sample inference and variance estimation [14, 15], hypothesis testing [14, 30], a causal inference framework [17, 23], inverse probability weighting [19, 20, 25], and semiparametric regression [21].

Two additional win statistics can provide perspectives on treatment efficacy that complement the win ratio [10]. The net benefit

$$NB = \pi_T - \pi_C \tag{2.6}$$

represents an absolute difference in the probability of winning between the treatment and control groups and is estimated with the sample NB:

$$\widehat{NB} = \widehat{\pi}_T - \widehat{\pi}_C \tag{2.7}$$

The NB is on the probability scale and ranges from -1 to 1 with a null value of 0, though in practice the effective range is narrower when the probability of a tie isn't negligible. The win odds (WO) is defined as

$$WO = \frac{\pi_T + 0.5 \cdot \pi_{tie}}{\pi_C + 0.5 \cdot \pi_{tie}} \tag{2.8}$$

where $\pi_{tie} = 1 - \pi_T - \pi_C$. The WO accounts for the proportion of ties by granting half a win to each group to all tied pairs, and can be estimated with the sample WO:

$$\widehat{WO} = \frac{\widehat{\pi}_T + 0.5 \cdot \widehat{\pi}_{tie}}{\widehat{\pi}_C + 0.5 \cdot \widehat{\pi}_{tie}} \tag{2.9}$$

where $\widehat{\pi}_{tie} = 1 - \widehat{\pi}_T - \widehat{\pi}_C$. The WO lies on the odds scale and is less sensitive to heavy censoring. Because the WR, NB, and WO are all derived from the same win probabilities, they all test the identical null hypothesis $\pi_T = \pi_C$, and produce approximately equal z-values, yielding very similar p-values and statistical power [10]. Population and sample quantities for all three win statistics are summarized in Table 2.1.

2.2 The Win Ratio Estimand

2.2.1 The Win Ratio's Dependence on Follow-Up Time and Censoring

The U-statistics formulation (2.3) of the population win ratio provides a robust framework for estimation, but defining an estimand for the win ratio that is generalizable and doesn't depend on censoring introduces other methodological challenges. The population win and loss probabilities (2.2) are defined in terms of expectations of kernel functions over the observed time-to-event variables, not the true latent time-to-event variables. Since the observed event times are bound by censoring time, they are dependent on both the censoring distribution and follow-up time. As articulated by Mao (2024), this dependence can conflate the true underlying treatment effect with the censoring pattern, so that two studies evaluating the same treatment could yield different win ratios simply because they have different follow-up durations or censoring rates [24].

We make this dependence on censoring and follow-up time mathematically explicit with an example: consider the simple case of a single time-to-event outcome evaluated up to a follow-up horizon φ. Let $S_T(t)$ and $S_C(t)$ denote the survival functions for the treatment and control groups, $\lambda_T(t)$ and $\lambda_C(t)$ their hazard functions, and $Q_T(t)$ and $Q_C(t)$ the survival functions for their respective independent censoring distributions. Following Dong et al. (2020), the probability that the treatment group wins by time φ is [19]:

$$\pi_T(\varphi) = \int_0^{\varphi} S_T(t)\, Q_T(t)\, S_C(t)\, Q_C(t)\, \lambda_C(t)\, dt \tag{2.10}$$

Similarly, the probability that the control group wins by time φ is

$$\pi_C(\varphi) = \int_0^{\varphi} S_C(t)\, Q_C(t)\, S_T(t)\, Q_T(t)\, \lambda_T(t)\, dt \tag{2.11}$$

The true win ratio estimand evaluated up to time φ is the ratio of these probabilities:

$$WR(\varphi) = \frac{\pi_T(\varphi)}{\pi_C(\varphi)} \tag{2.12}$$

illustrating that the win ratio estimand is a function of both the follow-up time φ and censoring distributions $Q_T(t)$ and $Q_C(t)$.

2.2.2 The Win Ratio Over Time

In practice, the impact of this dependence becomes evident when evaluating the win ratio dynamically over time. By computing the win ratio at successive time points throughout the study, investigators can observe how the treatment effect evolves as the follow-up time lengthens [18]. Typically, early in the follow-up period, non-fatal events drive the majority of the definitive pairwise comparisons. As later times, fatal events increasingly occur and take precedence in the hierarchy, resolving pairs that would have otherwise tied, or superseding earlier non-fatal events in the comparison [4, 18]. If the overall win ratio estimate varies substantially as follow-up time increases, it indicates that the estimand is heavily influenced by the shifting composition of events over time rather than reflecting a stable treatment effect [18]. Conversely, a win ratio that is approximately constant over time provides evidence consistent with the proportional win-fractions assumption (Section 2.4) and suggests that the choice of follow-up duration may not substantially affect the conclusion [21]. Dong et al. (2020) showed that plotting the win ratio over time can reveal long-term treatment benefits on fatal events that may be masked by the standard time-to-first-event analysis when non-fatal events dominate early follow-up [18, 31]. In our applied analysis (Chapter 4), we compute the WR, WO, and NB over time to assess the stability of the effect across follow-up durations.

2.2.3 Addressing Dependence on Censoring and Follow-Up Time

To address the censoring dependence of the win ratio estimand, two structural approaches have been proposed [24]. The first is to define the estimand as the win ratio at a pre-specified time φ, where all subjects are compared over the same time frame. Wang, Li, and Qu (2025) developed estimation methods for this restricted time win ratio using a multiple imputation approach [26]. The second approach involves semiparametric modeling using the proportional win-fractions (PW) regression model developed by Mao and Wang (2021) [21]. This approach yields regression coefficients that are invariant to the censoring distribution under a proportionality assumption similar to that of the Cox model, and is discussed further in Section 2.4. Both approaches redefine or constrain the estimand. A third complementary strategy is to target the original estimand but adjust for censoring through reweighting, as in the IPCW approach discussed in Section 2.3.

2.2.4 The Win Ratio Under a Bivariate Lehmann Family

An exception in which the estimand dependence problem resolves was established by Oakes (2016), who showed that under a Bivariate Lehmann family of distributions the win ratio does not depend on the follow-up time or censoring distribution [32]. Given joint survival functions $S_T(\cdot)$ and $S_C(\cdot)$ for a treatment and control group under two time-to-event outcomes in semi-competing risk, then if

$$S_T(t_1, t_2) = [S_C(t_1, t_2)]^\gamma \tag{2.13}$$

for $\gamma > 0$, the bivariate distribution of the time-to-event variables lies in the Lehmann family [33]. One special case of this family occurs under a proportional hazards structure with one time-to-event outcome. For example, if the treatment group hazard is proportional to the control group hazard such that $\lambda_T(t) = \theta \cdot \lambda_C(t)$ for some constant $\theta > 0$, then substituting this relationship into the integral in (2.12) yields the relationship $\pi_T(\varphi) = \theta \cdot \pi_C(\varphi)$. Then the integrals in (2.3) cancel to yield:

$$WR(\varphi) = 1/\theta \tag{2.14}$$

regardless of censoring distributions or follow-up time φ.

This result motivates the choice of data-generating models in our simulation study (Chapter 3) where we use a bivariate exponential marginal distribution with a Gumbel-Hougaard copula to model dependence, which falls in the Bivariate Lehmann family [32,33]. This allows us to define a "true" win ratio that does not depend on the follow-up horizon and serves as an unambiguous benchmark for evaluating the bias of various estimators. We discuss this choice and its implications further in Chapter 3.

The result in (2.14) also connects the win ratio to other familiar measures of treatment effect. Under proportional hazards for a single time-to-event outcome, the constant θ is the hazard ratio from a Cox model, so $WR = 1/HR$ [32, 34]. Similarly, for a simple binary or ordinal outcome, the win ratio reduces to the odds ratio [34]. These identities clarify what the win ratio targets in univariate scenarios, but they fundamentally break down for composite outcomes with multiple components where the win ratio and hazard ratio estimate different quantities, primarily because TFE analysis ignores terminal events that occur after initial non-fatal events [7,18]. This distinction motivates the comparisons made between the win ratio and Cox model in our simulation study (Chapter 3) and applied analysis (Chapter 4).

2.3 Inverse-Probability Weighted Win Ratio

The framework introduced in Sections 2.1 and 2.2 was established under the assumption of randomized treatment assignment and independent censoring. In observational settings, these assumptions are often violated since both treatment exposure and the censoring mechanism may depend on baseline covariates. Under this setting, an unadjusted win ratio will reflect covariate imbalance rather than a true difference between treatment groups. In this section, we describe two inverse-probability weighting approaches that modify the pairwise comparison kernels (2.1): inverse-probability-of-treatment weighting (IPTW), which adjusts for confounding, and inverse-probability-of-censoring weighting (IPCW), which adjusts for censoring bias. Both can be applied individually or in combination.

2.3.1 IPTW: Adjusting for Confounding

To address an exposure or treatment assignment that may depend on measured baseline covariates, Wang et al. (2025) proposed the IPTW-adjusted win ratio [25]. By weighting each subject's contribution to the pairwise comparison by the inverse of their estimated probability of belonging to their treatment group, the method balances covariates between treatment groups and creates a synthetic, standardized sample [25].

Let $A_i \in \{0, 1\}$ denote the treatment group assignment and X_i the vector of measured baseline covariates for subject i. The propensity score is defined as

$$e(X_i) = P(A_i = 1 \mid X_i) \tag{2.15}$$

and is typically estimated via logistic regression of A on X. For each subject, a weight ω_i is computed as a function of $\hat{e}(X_i)$ and A_i.

Wang et al. (2025) proposed three weighting schemes. The first, IPTW-ATE, weights treated subject i and control subject j by:

$$\omega_i = \frac{1}{\hat{e}(X_i)}, \quad \omega_j = \frac{1}{1 - \hat{e}(X_j)} \tag{2.16}$$

creating a pseudo-population in which covariates are balanced across groups. The resulting win ratio estimates the average treatment effect (ATE) in the full study population. Similarly, the stabilized IPTW-ATE (SIPTW-ATE) weights subjects by

$$\omega_i = \frac{N_T}{N_T + N_C} \cdot \frac{1}{\hat{e}(X_i)}, \quad \omega_j = \frac{N_C}{N_T + N_C} \cdot \frac{1}{1 - \hat{e}(X_j)} \tag{2.17}$$

reducing variance from extreme weights by multiplying by the marginal probability of receiving treatment while preserving the ATE interpretation. The IPTW-ATT scheme weights subjects by:

$$\omega_i = 1, \quad \omega_j = \frac{\hat{e}(X_j)}{1 - \hat{e}(X_j)} \tag{2.18}$$

reweighting the control group to resemble the treated group on measured covariates, resulting in a win ratio that estimates the average treatment effect in the treated (ATT) [25].

The IPTW-adjusted win proportions are computed by applying these weights directly

to the pairwise comparison kernels (2.1). The weighted win proportions are:

$$\widehat{\pi}_T^{\text{IPTW}} = \frac{\sum_{i=1}^{N_T}\sum_{j=1}^{N_C} \omega_i\,\omega_j\,\phi_1(Y_i, Y_j)}{\sum_{i=1}^{N_T}\sum_{j=1}^{N_C} \omega_i\,\omega_j}, \quad \widehat{\pi}_C^{\text{IPTW}} = \frac{\sum_{i=1}^{N_T}\sum_{j=1}^{N_C} \omega_i\,\omega_j\,\phi_2(Y_i, Y_j)}{\sum_{i=1}^{N_T}\sum_{j=1}^{N_C} \omega_i\,\omega_j} \tag{2.19}$$

and the IPTW-adjusted win ratio is

$$\widehat{WR}^{\text{IPTW}} = \frac{\widehat{\pi}_T^{\text{IPTW}}}{\widehat{\pi}_C^{\text{IPTW}}} \tag{2.20}$$

In practice, propensity scores near 0 or 1 can produce extreme weights that inflate variance. The stabilized weights in (2.17) can mitigate this, though if extreme weights persist then trimming the weights at a specified quantile is commonly used to stabilize the estimate even further at the cost of a slight increase in bias.

The interpretation of the IPTW-adjusted win ratio depends on the nature of the exposure A_i. For modifiable exposures, Zhang (2022) formally established that the inverse-probability weighted U-statistic successfully targets the true causal win ratio provided the four standard causal assumptions hold: consistency, the stable unit treatment value assumption (SUTVA), positivity ($0 < e(X_i) < 1$), and strong ignorability (no unmeasured confounding) [23]. We formally define this true causal win ratio in Chapter 3, where it serves as a benchmark for evaluating estimator performance.

However, applying this strict potential outcomes framework to non-modifiable demographic characteristics, such as race or sex, is methodologically problematic, as defining a counterfactual outcome for an immutable trait violates the core causal tenet of "no causation without manipulation" [35]. In these settings, IPTW-adjusted win ratio is more rigorously interpreted through the lens of epidemiological standardization. That is, by creating a balanced pseudo-population, the IPTW approach answers the standardized association question: *what would the relative disparity in the composite hierarchical outcome be if the compared groups shared identical baseline covariate profiles*? This approach relaxes the causal assumptions to only require positivity and exchangeability, and serves as the methodological foundation for our applied analysis of the MESA cohort (Chapter 4).

Regardless of whether the ultimate goal is causal inference or evaluating associations, the validity of the IPTW-adjusted win ratio rests entirely on the assumption of no unmeasured confounding, which is inherently untestable in observational data. If important confounders are omitted from X_i, the weights will fail to fully balance the comparison groups, and the resulting IPTW-adjusted win ratio will retain residual bias, motivating our efforts to study unmeasured confounding in our simulation study (Chapter 3).

2.3.2 IPCW: Adjusting for Censoring

The win ratio is also susceptible to bias from the censoring distribution. As discussed in Section 2.2.4, if the data-generating model belongs to the Bivariate Lehmann family (e.g. by satisfying the proportional hazards assumption) the win ratio estimand (2.14) is invariant to independent censoring [32]. However, even under this estimand invariance, the sample win proportions (2.4) are computed from observed data and underestimate the censoring-free win probabilities π_T and π_C since heavy censoring increases the proportion of tied pairs. Moreover, when the proportional hazards assumption is violated, or when the censoring distribution depends on covariates, the estimand itself may also shift [18, 19].

To address bias from independent censoring, Dong et al. (2020) proposed an IPCW win ratio that re-weights each resolved pairwise comparison to compensate for censored pairs [19]. Let $\widehat{Q}_T(\cdot)$ and $\widehat{Q}_C(\cdot)$ denote the Kaplan-Meier estimates of the censoring survival functions $Q_T(\cdot)$ and $Q_C(\cdot)$ introduced in Section 2.2. For a comparison in which treatment subject i wins over control subject j, the IPCW weight is:

$$\omega_{ij}^{\text{IPCW}} = \frac{1}{\widehat{Q}_T(Y_j)\,\widehat{Q}_C(Y_j)} \tag{2.21}$$

evaluated at the observed event time of the losing subject j. The weight is defined analogously when the control subject wins, with the denominator evaluated at Y_i. Intuitively, each resolved comparison is inflated by the inverse probability that both members of the pair remained uncensored long enough for the comparison to be determined. The weighted

win proportions are:

$$\widehat{\pi}_T^{\text{IPCW}} = \frac{1}{N_T N_C} \sum_{i=1}^{N_T} \sum_{j=1}^{N_C} \omega_{ij}^{\text{IPCW}} \, \phi_1(Y_i, Y_j), \quad \widehat{\pi}_C^{\text{IPCW}} = \frac{1}{N_T N_C} \sum_{i=1}^{N_T} \sum_{j=1}^{N_C} \omega_{ji}^{\text{IPCW}} \, \phi_2(Y_i, Y_j) \tag{2.22}$$

and the IPCW-adjusted win ratio is:

$$\widehat{WR}^{\,\text{IPCW}} = \frac{\widehat{\pi}_T^{\text{IPCW}}}{\widehat{\pi}_C^{\text{IPCW}}} \tag{2.23}$$

Under independent censoring, $\widehat{WR}^{\,\text{IPCW}}$ consistently estimates $WR = \pi_T/\pi_C$ [19].

To address bias from dependent censoring, Dong et al. (2021) extended this approach to a covariate-adjusted IPCW (CovIPCW) win ratio [20]. This approach uses a Cox model adjusted by covariates, rather than a marginal Kaplan-Meier estimator, for the estimated censoring survival functions in (2.21). By assigning higher weights to patients who remain uncensored but share similar baseline characteristics with those who dropped out, the method recovers the unbiased estimate. When both confounding and informative censoring are present, as is common in observational settings, IPCW weights can be applied jointly with the IPTW weights to adjust for both sources of bias simultaneously.

2.4 Proportional Win-Fractions Regression

While the IPTW approach addresses confounding by modeling the treatment assignment mechanism, confounding can also be addressed directly through outcome modeling. Mao and Wang (2021) proposed a class of semiparametric proportional win-fractions (PW) regression models that extend the win ratio from a two-sample comparison to a covariate-adjusted regression framework [21]. Under the PW model, the covariate-specific win and loss fractions are assumed to be proportional over time.

Define the time-dependent win functions:

$$\phi_1(Y_i, Y_j)(t) = \mathbb{I}\{Y_i \text{ wins over } Y_j \text{ by time } t\}, \quad \phi_2(Y_i, Y_j)(t) = \mathbb{I}\{Y_j \text{ wins over } Y_i \text{ by time } t\} \tag{2.24}$$

which conceptually identical to the pairwise comparison kernels 2.1, but are explicitly evaluated using only the event and censoring history observed up to time t. For subjects with

measured baseline covariate vectors X_i and X_j, the PW model specifies:

$$\frac{E\left[\phi_1(Y_i, Y_j)(t) \mid X_i, X_j\right]}{E\left[\phi_2(Y_i, Y_j)(t) \mid X_i, X_j\right]} = \exp\left\{\beta^T (X_i - X_j)\right\} \tag{2.25}$$

for all $t \in [0, \varphi]$, where β is the vector of regression coefficients and φ is the maximum follow-up time [21]. The regression coefficient β_k can be interpreted as the conditional log win ratio associated with a one-unit increase in the k-th covariate, holding all other covariates constant. This model's validity rests on the proportional win-fractions assumption [21]. Analogous to using a Schoenfeld residual plot for a Cox model, model checking is performed by a visual inspection of a standardized cumulative residual process over time. Under the proportional win-fractions assumption, this "score" process fluctuates around a mean-zero horizontal line like a Brownian bridge.

A key property of the PW model is that the regression coefficient β is invariant to the censoring distribution under the proportionality assumption [21], resolving the estimand ambiguity discussed in Section 2.2 through model structure rather than time restriction or IPCW adjustment. The PW model also includes the Cox proportional hazards model as a special case. In particular, when analyzing a single time-to-event endpoint and the proportional hazards assumption holds, the PW regression coefficient equals the inverse hazard ratio [21].

It is important to note that the PW regression approach estimates a conditional win ratio, whereas the IPTW approach estimates a marginal win ratio. They estimate subtly different quantities that don't necessarily have to agree, particularly because the win ratio, like the hazard ratio and odds ratio, is non-collapsible [36]. This distinction is relevant when interpreting the results of our applied analysis (Chapter 4), where we report both IPTW-adjusted and PW regression win ratios.

Table 2.1: Win statistics, related quantities, and their definitions.

Quantity	**Definition**
Population Quantities	
π_T	Probability that an individual from the treatment group wins
π_C	Probability that an individual from the control group wins
$\pi_{tie} = 1 - \pi_T - \pi_C$	Probability that two individuals from the treatment and control groups tie
$WR = \frac{\pi_T}{\pi_C}$	Win ratio (population)
$NB = \pi_T - \pi_C$	Net benefit (population)
$WO = \frac{\pi_T + 0.5 \cdot \pi_{tie}}{\pi_C + 0.5 \cdot \pi_{tie}}$	Win odds (population)
Sample Quantities	
N_T	Number of individuals in the treatment group sample
N_C	Number of individuals in the control group sample
$N = N_T + N_C$	Total number of individuals in the sample
n_T	Number of wins for the treatment group in the all-against-all comparisons
n_C	Number of wins for the control group in the all-against-all comparisons
n_{tie}	Number of ties in the all-against-all comparisons
$\widehat{\pi}_T = \frac{n_T}{N_T N_C}$	Win proportion for the treatment group (number of wins for the treatment group per number of comparisons made between groups)
$\widehat{\pi}_C = \frac{n_C}{N_T N_C}$	Win proportion for the control group
$\widehat{\pi}_{tie} = 1 - \widehat{\pi}_T - \widehat{\pi}_C$	Proportion of ties
$\widehat{WR} = \frac{\widehat{\pi}_T}{\widehat{\pi}_C}$	Win ratio (sample)
$\widehat{NB} = \widehat{\pi}_T - \widehat{\pi}_C$	Net benefit (sample)
$\widehat{WO} = \frac{\widehat{\pi}_T + 0.5 \cdot \widehat{\pi}_{tie}}{\widehat{\pi}_C + 0.5 \cdot \widehat{\pi}_{tie}}$	Win odds (sample)

Chapter 3

SIMULATION STUDIES

3.1 Motivation and Aims

In observational settings, the validity of inverse-probability weighting methods relies on the critical assumption of no unmeasured confounding. In practice, this assumption is untestable, yet no existing work has systematically evaluated how the IPTW-WR performs as a function of unmeasured confounding strength [23, 25]. Understanding the magnitude and direction of bias in this setting, and how the bias compares to the unadjusted win ratio, is crucial for applied researchers considering the IPTW-WR in observational settings where unmeasured confounding is plausible. Moreover, it remains unknown how, and whether, combined IPCW and IPTW methods improve win ratio performance beyond either adjustment alone, especially under confounding [19, 20]. Furthermore, while the theoretical distinctions between the win ratio and Cox model approaches to composite outcomes have been well documented, their comparative performance under complex observational data structures requires empirical investigation. In particular, when treatment effects differ across components, the two approaches may yield substantively different conclusions. Characterizing when and how this divergence arises is important for informing the choice of analytic method in applied studies.

To address these gaps, we perform two complementary simulation studies. Study 1 evaluates the empirical bias, confidence interval coverage, and statistical power of the win ratio and its IPW extensions in observational settings when the assumption of no unmeasured confounding is violated. Study 2 compares the win ratio and Cox model under differential treatment effects across the fatal and non-fatal components of the composite outcome. Both simulation studies build on the work of Zhang et. al (2022), who established a causal inference framework for the win ratio in observational settings and conducted simulation studies using semi-competing risk event data with exponential marginal event times and

Gumbel-Hougaard copula dependence. We extend this work in three ways. First, we calibrate the baseline hazard and censoring mechanisms directly to MESA data, grounding the simulation in a realistic observational study setting. Second, we introduce unmeasured confounding and systematically vary it across multiple levels to characterize the resulting bias. Lastly, we introduce a novel concordance analysis to characterize when the win ratio and Cox models diverge in their apparent conclusions about treatment benefit.

3.2 Data-Generating Model

3.2.1 Covariates and Treatment Assignment

For N independent subjects, we simulate a hierarchical composite outcome with two time-to-event components, one fatal (T_D) and one non-fatal (T_H), along with a binary treatment indicator A, a measured covariate X, and an unmeasured covariate U. We assume the covariates are independent and identically distributed:

$$X_1 \sim \text{Normal}(0, 1), \quad U \sim \text{Normal}(0, 1) \tag{3.1}$$

Treatment assignment follows a logistic model that depends on both the measured and unmeasured covariates:

$$P(A = 1 \mid X, U) = \frac{1}{1 + \exp\{-(\alpha_0 + \alpha_X X + \alpha_U U)\}} \tag{3.2}$$

where $\alpha_0 = 0$ ensures approximately balanced treatment groups, α_X is the fixed effect of the measured confounder on treatment assignment, and α_U is the effect of the unmeasured confounder on treatment assignment, which is varied across simulation scenarios. When $\alpha_U = 0$, treatment assignment depends only on the measured covariate X, and the assumption of no unmeasured confounding holds. When $\alpha_U > 0$, the unmeasured confounder U affects both treatment assignment and the event hazards, violating this assumption and introducing confounding bias.

3.2.2 Event Time Distributions

We assume exponential marginal distributions for the latent time-to-event components:

$$T_D \sim \text{Exponential}(\Lambda_D), \quad T_H \sim \text{Exponential}(\Lambda_H) \tag{3.3}$$

with hazard rates given by proportional hazards models:

$$\Lambda_D = \lambda_D \exp\{-(\eta_D A + \beta_{X,D} X + \beta_{U,D} U)\} \tag{3.4}$$

$$\Lambda_H = \lambda_H \exp\{-(\eta_H A + \beta_{X,H} X + \beta_{U,H} U)\} \tag{3.5}$$

where λ_D and λ_H are baseline hazard rates, η_D and η_H are the treatment effects on the fatal and non-fatal events respectively, $\beta_{X,D}$ and $\beta_{X,H}$ are the effects of the measured covariate, and $\beta_{U,D}$ and $\beta_{U,H}$ are the effects of the unmeasured covariate [37]. Following Zhang et al. (2022), we parameterize the hazard models so that positive coefficients correspond to a reduction in event hazards [23]. The marginal survival functions are then given by:

$$S_D(t_D \mid A, X, U) = \exp(-\Lambda_D \cdot t_D), \quad S_H(t_H \mid A, X, U) = \exp(-\Lambda_H \cdot t_H) \tag{3.6}$$

The choice of exponential marginals is deliberate. Under exponential distributions, the proportional hazards assumption holds and the win ratio is time-invariant, as shown by Oakes (2016) [32]. This means the WR estimand does not depend on the follow-up horizon, and a "true" win ratio can be defined unambiguously (Section 3.2.6). This time-invariance eliminates the censoring-dependence of the WR estimand discussed by Mao (2024), allowing us to isolate the effect of confounding bias from the estimand issue [24]. The practical consequence is that any observed bias in the simulation can be attributed to confounding, not to the ambiguity of the estimand.

3.2.3 Dependence Structure

We model the joint survival function for (T_D, T_H) with the Gumbel-Hougaard copula, following previous win ratio simulation studies [14, 23, 32]:

$$S_{D,H}(t_D, t_H \mid A, X_1, U) = \exp\left\{-\left[(\Lambda_D t_D)^\theta + (\Lambda_H t_H)^\theta\right]^{1/\theta}\right\} \tag{3.7}$$

where $\theta \geq 1$ controls the association between T_D and T_H, with $\theta = 1$ denoting independence. Kendall's tau is given by $\tau_K = 1 - \theta^{-1}$ [33].

In practice, the latent event times T_D and T_H are generated by first drawing a bivariate vector (V_D, V_H) from the Gumbel-Hougaard copula, where V_D and V_H are marginally standard uniform, and then applying the probability integral transform:

$$T_D = \frac{-\log(V_D)}{\Lambda_D}, \quad T_H = \frac{-\log(V_H)}{\Lambda_H} \tag{3.8}$$

3.2.4 Censoring Mechanism

We introduce a random censoring time C, assumed independent of (T_D, T_H, X, U) conditional on treatment:

$$C \sim \text{Exponential}(\lambda_C) \tag{3.9}$$

where λ_C is the censoring rate, varied across scenarios.

We additionally impose administrative censoring at a fixed time horizon φ in all scenarios, corresponding to the approximate follow-up duration in MESA. The observed time-to-fatal-event and its event indicator are:

$$Y_D = \min(T_D, C, \varphi), \quad \delta_D = \mathbb{I}\{T_D \leq \min(C, \varphi)\} \tag{3.10}$$

and the observed time-to-non-fatal-event, incorporating the semi-competing risk structure where death censors the non-fatal event, are:

$$Y_H = \min(T_H, T_D, C, \varphi), \quad \delta_H = \mathbb{I}\{T_H \leq \min(T_D, C, \varphi)\} \tag{3.11}$$

The observed dataset for each simulated sample is:

$$\mathcal{O} = \{(A_i, X_i, Y_{D,i}, \delta_{D,i}, Y_{H,i}, \delta_{H,i}) : i = 1, \ldots, N\} \tag{3.12}$$

Note that the unmeasured covariate U is used in data generation but is not included in $\mathcal{O}$, reflecting the observational setting where not all confounders are measured. When random censoring is absent ($\lambda_C = 0$), a small amount of administrative censoring remains due to φ.

3.2.5 Calibration Using MESA Data

To ensure that the simulation reflects realistic epidemiological parameters and event trajectories, the baseline hazard rates and dependence structures are calibrated using data from the Multi-Ethnic Study of Atherosclerosis (MESA). MESA is a diverse, population-based cohort study of 6814 adults aged 45-84 years who were free of clinical cardiovascular disease at baseline and followed for incident events [29]. These calibrated parameters serve as starting points for the simulation data-generating mechanism. The formal analysis of the MESA cohort is presented separately in Chapter 4.

We estimated the baseline hazard rates from the MESA cohort by fitting simple exponential regression models to time-to-fatal-CVD and time-to-non-fatal-CVD distributions. This yielded baseline rates of $\lambda_D = 0.00047$ per year for fatal events and $\lambda_H = 0.00051$ per year for non-fatal events. Because the MESA cohort was selected to be free of clinical cardiovascular disease at baseline, these rates reflect a relatively healthy population of adults [29]. Using these raw rates in a simulation has the risk of generating too few events to adequately evaluate estimator performance and statistical power. Therefore, to approximate a higher-risk population while strictly preserving the MESA-calibrated ratio of non-fatal to fatal events, we scaled both baseline rates by a factor of 10.

To calibrate the Gumbel-Hougaard copula dependence parameter, we estimated Kendall's tau between the time-to-death and time-to-non-fatal-event among the $n = 183$ MESA participants who experienced both events during follow-up. This yielded an estimated Kendall's tau of $\tau_K = 0.585$, which translated to a copula parameter of $\theta = 2.41$. While we acknowledge that estimating the correlation conditionally on observing both events introduces an inherent selection bias, this estimate provides a highly plausible, data-driven starting value for the joint survival function. Finally, the baseline random censoring rate was estimated from the MESA cohort as $\lambda_C = 0.058$ per year.

3.2.6 The True Causal Win Ratio

To systematically quantify the empirical bias of the unadjusted and weighted estimators in our study, we require a benchmark against which to compare. Under the potential outcomes

framework established by Zhang and Jeong (2021), the "true" win ratio WR_{true} is given by:

$$WR_{\text{true}} = \frac{P(Y^{(1)} \text{ wins over } Y^{(0)})}{P(Y^{(0)} \text{ wins over } Y^{(1)})} \tag{3.13}$$

where $Y^{(1)}$ and $Y^{(0)}$ are the potential composite outcomes under treatment and control, respectively, and the comparison follows the hierarchical procedure described in Section 1.3 [17, 22, 23]. We derive the closed form of WR_{true} used in this simulation study in Appendix C.

3.3 Simulation Framework

3.3.1 Study 1: Win Ratio Under Unmeasured Confounding

The first simulation study is designed to evaluate the performance of the win ratio estimators when the assumption of no unmeasured confounding is violated. We employ a factorial design that crosses three levels of unmeasured confounding, three levels of treatment effect size, and two censoring levels. The strength of unmeasured confounding is controlled through the coefficient α_U governing the effect of U on treatment assignment and through the coefficients $(\beta_{U,D}, \beta_{U,H})$ governing its effect on the fatal and non-fatal event hazards. We set these to three levels: none ($\alpha_U = 0$, $\beta_{U,D} = 0$, $\beta_{U,H} = 0$), where the no confounding assumption holds, moderate ($\alpha_U = 0.5$, $\beta_{U,D} = 0.25$, $\beta_{U,H} = 0.20$); and strong ($\alpha_U = 1.0$, $\beta_{U,D} = 0.50$, $\beta_{U,H} = 0.40$).

Treatment effects sizes are set equal across both components ($\eta_H = \eta_D$), and are set to 0 (null), 0.20 (medium), and 0.35 (large). In the proportional hazards models 3.4 and 3.5, these values represent the treatment having an effect of reducing the log-hazard for each component event. Moreover, setting the effects equal to each other ensures that any differences across methods arise from confounding and censoring rather than from differential component effects, which are investigated separately in Study 2. Finally, we evaluate two censoring paradigms: administrative censoring only ($\lambda_C = 0$, with $\varphi = 20$ years) and moderate independent censoring ($\lambda_C = 0.058$, with $\varphi = 20$ years).

This factorial design yields $3 \times 3 \times 2 = 18$ scenarios. All other data-generating parameters are fixed across Study 1: the sample size $N = 1{,}000$, the effect of measured confounder X

on treatment assignment $\alpha_X = 0.5$, and the MESA-calibrated baseline hazard rates, copula dependence parameter, and administrative censoring described in Section 3.2.5. The true causal win ratio WR_{true}, described analytically in Section 3.2.6, is computed via numerical integration and reported for each scenario.

For each scenario, we generated $B = 250$ independent datasets. Within each dataset, we compute and compare five methods: the unadjusted win ratio, the IPCW-adjusted win ratio, the IPTW-adjusted win ratio, the combined IPTW+IPCW win ratio, and a Cox proportional hazards model. For IPTW adjustment, the propensity score is estimated via logistic regression of A on X and the weights are stabilized and trimmed at the 99th percentile [25]. The Cox model adjusts for X as a covariate.

3.3.2 Study 2: Win Ratio vs. Cox Under Differential Treatment Effects

The second simulation study investigates how the apportionment of the treatment effect across the fatal and non-fatal components influences the relative performance of the win ratio compared to the Cox model. The key motivation is that the WR, by comparing death before non-fatal events in the hierarchy, may detect mortality benefits that the Cox TFE analysis misses when non-fatal events dominate early follow-up. To isolate the effect of treatment apportionment from that of confounding or censoring, we fixed the total treatment signal at $\eta_H + \eta_D = 0.45$ and distributed it across three treatment paradigms. The first is a death-driven effect ($\eta_D = 0.40$, $\eta_H = 0.05$), where treatment primarily reduces mortality. The second is a balanced effect ($\eta_D = 0.225$, $\eta_H = 0.225$), serves as a reference condition. The third is a non-fatal-driven effect ($\eta_D = 0.05$, $\eta_H = 0.40$), where treatment primarily reduces non-fatal events. The constant sum ensures that any differences between methods reflect the apportionment of the treatment signal rather than its total magnitude. Because the TFE analysis is sensitive to the relative frequency of the component events, we crossed these three treatment patterns with three ratios of the non-fatal to fatal baseline event rate ($\lambda_H/\lambda_D \in \{2, 4, 6\}$), implemented by setting $\lambda_H = r \times \lambda_D$ with $\lambda_D = 0.0047$ fixed. Higher ratios correspond to settings where non-fatal events increasingly dominate the composite.

This factorial design that crosses treatment signal allocation and event rates yields $3 \times$

$3 = 9$ scenarios, and for each scenario we generated $B = 250$ independent datasets. To strictly isolate the methodological handling of the composite hierarchy from the effects of unmeasured confounding, we set $\alpha_U = 0$ throughout Study 2. Confounding is limited entirely to the measured covariate X with $\alpha_X = 0.5$, and the IPTW adjustment fully accounts for it. Administrative censoring is set at $\varphi = 20$ years with no random censoring ($\lambda_C = 0$). The same five methods from Study 1 are applied to each replication.

3.3.3 Performance Metrics

Point Estimates and Standard Errors

In each scenario, and for each win ratio method, we report the mean log-transformed point estimates across B replications:

$$\bar{L} = \frac{1}{B} \sum_{b=1}^{B} \log(\widehat{WR}_b) \tag{3.14}$$

We characterize precision through two complementary standard error estimates. The empirical standard error measures the Monte Carlo variability of the estimator across replications:

$$\widehat{\mathrm{SE}}_{\mathrm{emp}} = \sqrt{\frac{1}{B} \sum_{b=1}^{B} \left(\log(\widehat{WR}_b) - \bar{L} \right)^2} \tag{3.15}$$

For a single replication b, the WINS package [38] constructs a Wald-type 95% confidence interval for the estimated log-WR, from which a standard error can be recovered:

$$\widehat{\mathrm{SE}}_b = \frac{\log \widehat{WR}_{U,b} - \log \widehat{WR}_{L,b}}{2 \times 1.96} \tag{3.16}$$

We report the estimated standard error as the median of $\widehat{\mathrm{SE}}_b$ across all B replications:

$$\widehat{\mathrm{SE}}_{\mathrm{est}} = \mathrm{median}_{b \in \{1,\ldots,B\}} \left\{ \widehat{\mathrm{SE}}_b \right\} \tag{3.17}$$

When $\widehat{\mathrm{SE}}_{\mathrm{est}} \approx \widehat{\mathrm{SE}}_{\mathrm{emp}}$, the variance estimator is well-calibrated. Systematic divergence between the two indicates overestimation or underestimation of variability, which becomes especially relevant when the estimator doesn't satisfy the condition of no unmeasured confounding.

For the Cox proportional hazards model, we report the mean of the estimated log inverse hazard ratio $\log(1/\widehat{HR}_b)$ across B replications

$$\bar{L}_{\text{Cox}} = -\frac{1}{B}\sum_{b=1}^{B}\log(\widehat{HR}_b) \tag{3.18}$$

along with the corresponding empirical and estimated standard errors. The Cox model is included throughout as a descriptive comparator rather than a direct one.

Bias

For each win ratio method, bias is computed on the log scale as

$$\text{Bias} = \bar{L} - \log(WR_{\text{true}}), \tag{3.19}$$

and denotes any over or under estimation of treatment effect on the log scale [23, 25]. Bias is not reported for the Cox estimate.

Coverage and Type I Error Rate

Empirical coverage is the proportion of estimated 95% confidence intervals that contain WR_{true}:

$$\text{Coverage} = \frac{1}{B}\sum_{b=1}^{B}\mathbb{I}\left\{\widehat{WR}_{L,b} \leq WR_{\text{true}} \leq \widehat{WR}_{U,b}\right\}. \tag{3.20}$$

Coverage is reported for win ratio methods only.

For each method and scenario, we additionally report the proportion of replications in which the two-sided p-value falls below 0.05. This quantity estimates the type I error rate under the null ($WR_{\text{true}} = 1$) and estimates power under non-null scenarios ($WR_{\text{true}} \neq 1$) [25, 39].

Median Attributable Win Proportions & Net Discordance

In Study 2 we report two additional metrics. First, we report the median proportion of pairwise wins attributed to each endpoint, computed from the component-specific win counts returned by the WINS package [38]. Treatment wins resolved from the death component reflect the hierarchical priority of fatal outcomes, while wins resolved by non-fatal event

component contribute only when the fatal comparison is uninformative. The Cox model will not take this hierarchy into account, and thus these attributable win proportions can give insight into why the WR and Cox models might diverge across treatment allocation regimes.

Second, we quantify the divergence between the two methods through a concordance classification. For each replication b, the significance at $\alpha = 0.05$ of the win ratio and adjusted Cox model at $\alpha = 0.05$ is classified into one of four categories: both significant, WR only significant, Cox only significant, or neither significant. We can then tally which category each replication falls into across the B replications, and define the Net discordance as the estimated difference in the observed proportions:

$$\text{Net Discordance} = \frac{\#\text{WR only significant}}{B} - \frac{\#\text{Cox only significant}}{B} \tag{3.21}$$

Positive net discordance indicates that the WR detects the treatment signal more frequently than the Cox model, and vice versa. This metric directly addresses the question of when the hierarchical composite approach outperforms the time-to-first-event approach in identifying a treatment benefit.

3.4 Results

We present the results of the two simulation studies described in Section 3.3. All point estimates and standard errors are reported on the log scale. Each scenario was evaluated over $B = 250$ replications with $N = 1000$ subjects in each replication.

3.4.1 Study 1

Table 3.1 summarizes the 18 simulation scenarios for study 1. The four WR methods (unadjusted, IPCW, IPTW, IPTW+IPCW) and the covariate-adjusted Cox model were applied to each replication. The full results are reported in Tables 3.2–3.4.

Null Scenarios

Table 3.2 presents performance under the null hypothesis ($\eta_H = \eta_D = 0$, $WR_{\text{true}} = 1$). When no unmeasured confounding was present and censoring was administrative only, the

IPTW-adjusted WR and the covariate-adjusted Cox model maintained estimated type I error rates around the 5% level with proportion of p-values < 0.05 at 4.4% and 4.8%, respectively. The unadjusted WR had a proportion of p-values < 0.05 at 7.2% and exhibited a positive mean log estimate ($\bar{L} = 0.080$) despite $\log(WR_{\text{true}}) = 0$. This overestimation occurred because the measured covariate X is associated with both treatment assignment ($\alpha_X = 0.5$) and event hazards ($\beta_{X,D} = 0.20$, $\beta_{X,H} = 0.15$), which the unadjusted WR does not account for. The IPTW estimator, which adjusted for X, had a near zero mean log estimate ($\bar{L} = -0.002$), confirming that the overestimation in the unadjusted WR reflects confounding from X. Standard error calibration was fair in this setting: the unadjusted WR had $\widehat{\text{SE}}_{\text{emp}} = 0.188$ and $\widehat{\text{SE}}_{\text{est}} = 0.179$, and the IPTW-WR had $\widehat{\text{SE}}_{\text{emp}} = 0.193$ and $\widehat{\text{SE}}_{\text{est}} = 0.185$. As expected, the IPCW adjustment had no effect under administrative-only censoring, producing results identical to the corresponding non-IPCW method.

As unmeasured confounding increased, estimated type I error rose sharply (Appendix Figure 3.1). Under moderate confounding ($\alpha_U = 0.5$) with administrative censoring, the unadjusted WR had a proportion of p-values < 0.05 at 15.2% and the IPTW-WR at 10.0%. Under strong confounding ($\alpha_U = 1.0$), the unadjusted estimated type I error rate reached 70.4% and the IPTW-WR 55.6%. The IPTW adjustment reduced the inflated type I error at every confounding level but could not eliminate it, because the propensity score model was misspecified by only including X and not accounting for the confounder U. Under strong confounding, the estimated SE also underestimated the empirical SE(e.g., for the unadjusted WR: $\widehat{\text{SE}}_{\text{est}} = 0.174$ vs. $\widehat{\text{SE}}_{\text{emp}} = 0.187$; Appendix Figure 3.3), indicating that the variance estimate is miscalibrated when the no-unmeasured-confounding assumption is violated.

Moderate censoring ($\lambda_C = 0.058$) increased standard errors by roughly 30% relative to administrative censoring, and all methods remained near the 5% rejection level when there was no unmeasured confounding. Under strong confounding with moderate censoring, the IPTW+IPCW combination had a smaller type I error relative to IPTW alone (30.4% vs. 39.2%).

Table 3.1: Study 1 simulation parameter summary. All 18 scenarios are defined by the factorial crossing of three unmeasured confounding levels, three treatment effect sizes, and two censoring levels. Fixed parameters: $N = 1{,}000$, $\alpha_X = 0.5$, $\beta_{X,D} = 0.20$, $\beta_{X,H} = 0.15$, $\theta = 2.41$, $\lambda_D = 0.0047$, $\lambda_H = 0.0051$, $\varphi = 20$.

Confounding	α_U	$\beta_{U,D}$	$\beta_{U,H}$	Effect	η	Censoring	λ_C	WR_{true}
None	0.0	0.00	0.00	Null	0.00	Admin. only	0.000	1.0000
				Null	0.00	Moderate	0.058	1.0000
				Medium	0.20	Admin. only	0.000	1.2214
				Medium	0.20	Moderate	0.058	1.2214
				Large	0.35	Admin. only	0.000	1.4191
				Large	0.35	Moderate	0.058	1.4191
Moderate	0.5	0.25	0.20	Null	0.00	Admin. only	0.000	1.0000
				Null	0.00	Moderate	0.058	1.0000
				Medium	0.20	Admin. only	0.000	1.2214
				Medium	0.20	Moderate	0.058	1.2214
				Large	0.35	Admin. only	0.000	1.4191
				Large	0.35	Moderate	0.058	1.4191
Strong	1.0	0.50	0.40	Null	0.00	Admin. only	0.000	1.0000
				Null	0.00	Moderate	0.058	1.0000
				Medium	0.20	Admin. only	0.000	1.2214
				Medium	0.20	Moderate	0.058	1.2214
				Large	0.35	Admin. only	0.000	1.4191
				Large	0.35	Moderate	0.058	1.4191

Non-Null Scenarios

Performance under the medium ($WR_{\text{true}} = 1.22$) and large ($WR_{\text{true}} = 1.42$) treatment effects are presented in Table 3.3 under administrative censoring only, and in Table 3.4

under both administrative censoring and moderate independent censoring. Bias increased monotonically with unmeasured confounding strength for all WR methods (Figure 3.1). When no unmeasured confounding was present and censoring was administrative only, the IPTW estimator was approximately unbiased (bias = 0.023 for the medium effect, 0.033 for the large effect), while the unadjusted WR showed positive bias (0.103 and 0.114), reflecting the confounding from the measured covariate X. Under strong confounding, the IPTW estimator reduced bias by approximately 0.05 log units relative to the unadjusted WR, but substantial residual bias remained. For the medium effect with administrative censoring, the unadjusted bias was 0.438 compared to 0.381 for IPTW (Table 3.3). These residual biases exceed the true log win ratio itself ($\log(1.22) = 0.200$). The direction of bias was positive across all scenarios.

Coverage of the 95% confidence intervals declined sharply with increasing confounding (Figure 3.2). When no unmeasured confounding was present, the IPTW-WR achieved coverage near the 95% level (93.6% for the medium effect, 95.6% for the large effect, under administrative censoring; Table 3.3). Under strong confounding, coverage fell to 32.4% for the unadjusted WR and 43.2% for IPTW in the medium-effect, administrative-censoring scenario. This failure reflects both the bias of the point estimate and the underestimation of variability documented in the null-scenario results above.

The proportion of replications with $P < 0.05$ increased with confounding strength under non-null scenarios (Appendix Figure 3.5). Under no confounding with administrative censoring, the IPTW-WR had lower rejection rates than the unadjusted WR (20.0% vs. 37.2% for the medium effect; 51.6% vs. 64.8% for the large effect), because adjusting for X reduced the apparent effect size down toward WR_{true}. Under strong confounding, all methods exceeded 85% rejection for the large effect, reflecting overinflated point estimates rather than improved power. The covariate-adjusted Cox model followed the IPTW-WR closely in rejection rate throughout.

Moderate censoring increased variability and reduced power relative to administrative-only censoring, but the relative outcomes among methods remained the same. Under moderate censoring with strong confounding, the IPTW+IPCW combination improved coverage by approximately 10 percentage points relative to IPTW alone for the medium effect (74.4%

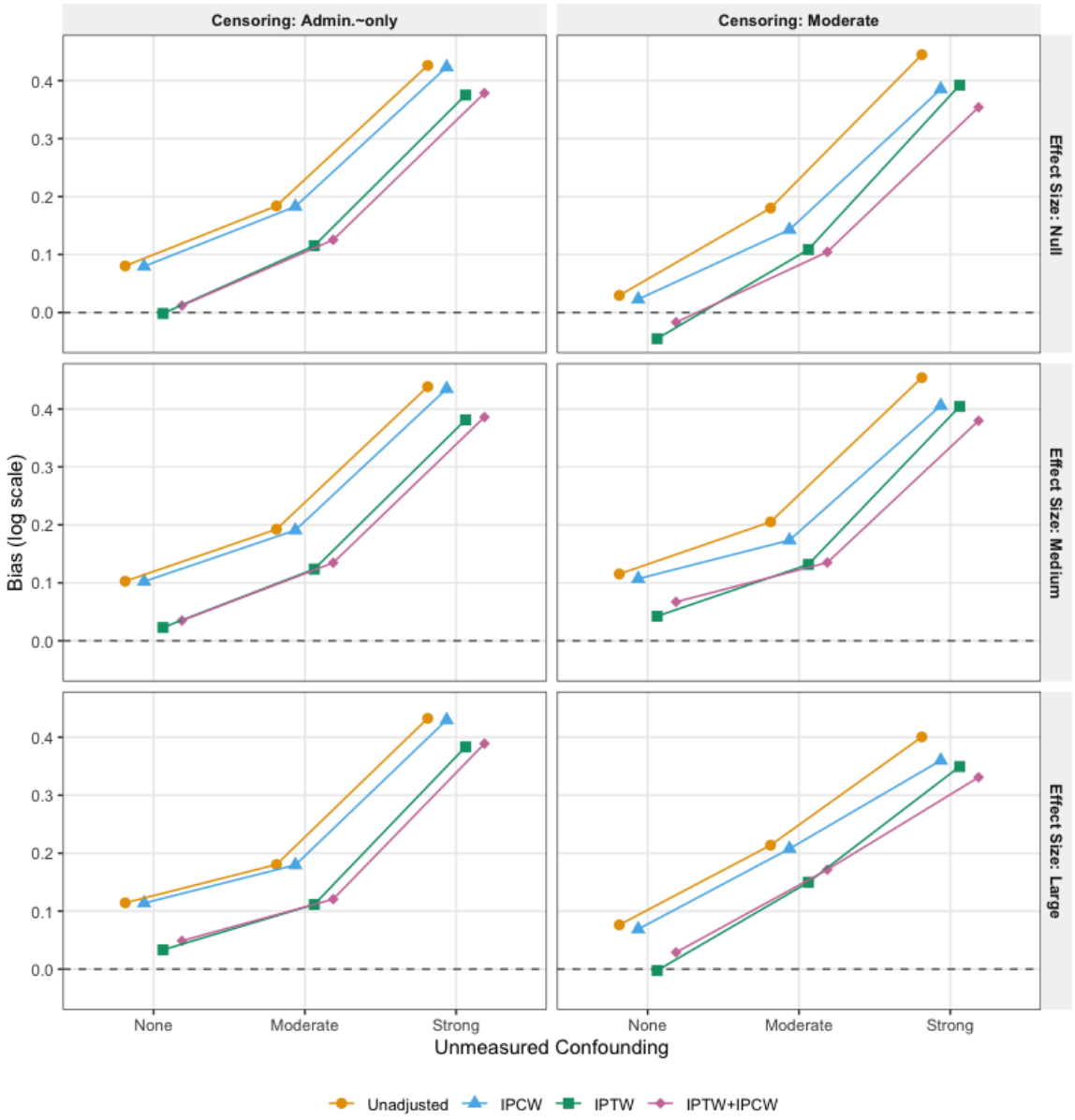

Figure 3.1: (Study 1) Log-scale bias by unmeasured confounding level across all scenarios. Each panel corresponds to one combination of treatment effect size and censoring level.

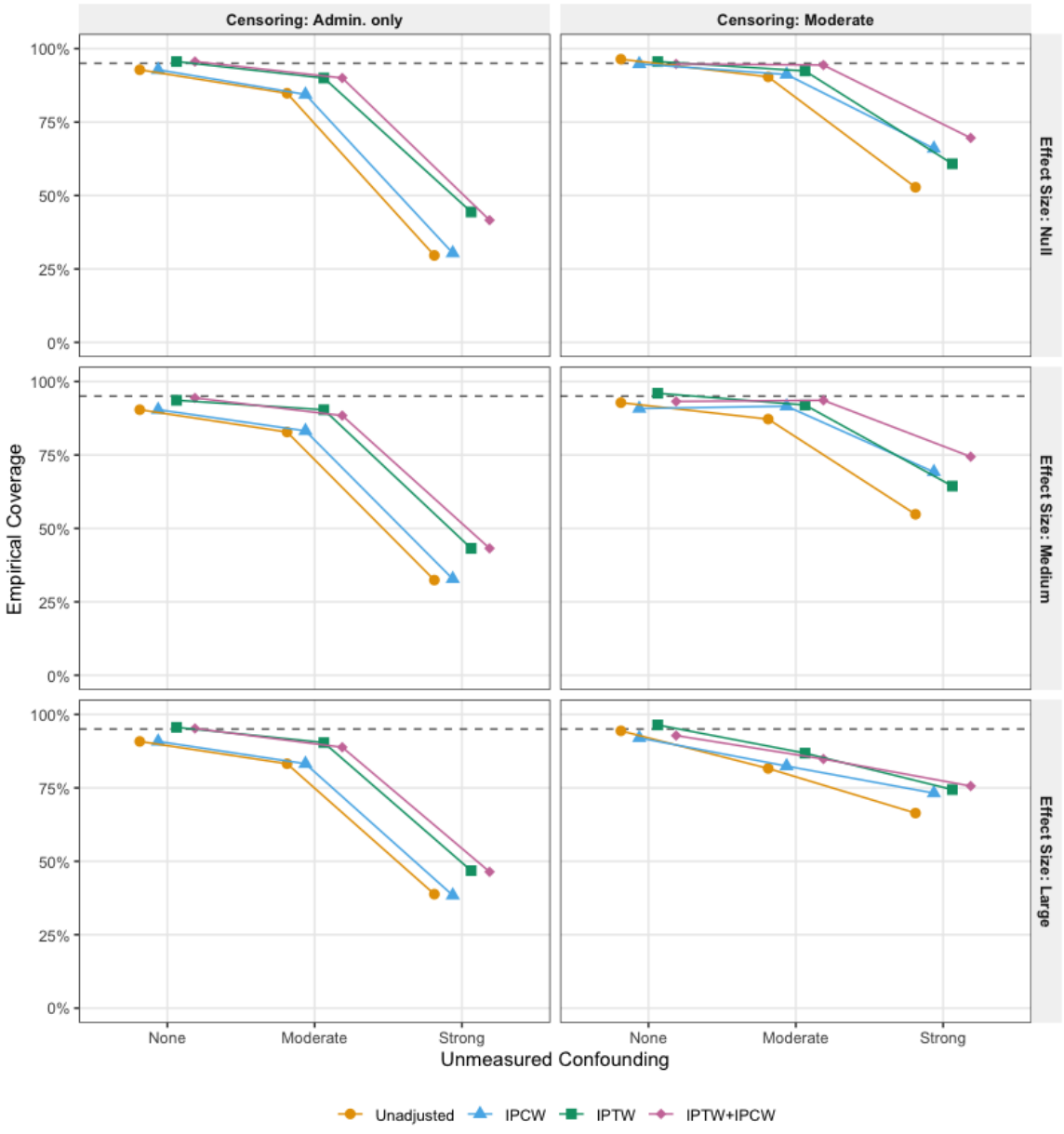

Figure 3.2: (Study 1) Empirical coverage of the 95% confidence interval by unmeasured confounding level across all scenarios. The dashed line marks the 95% level.

vs. 64.4%; Table 3.4). The tie proportion also increased from approximately 64% to 72% in null scenarios with no confounding (Appendix Figure 3.6), reflecting the reduction in resolved comparisons.

3.4.2 Study 2

Table 3.5 presents point estimates, power, and component-wise win proportions for the IPTW-adjusted WR and covariate-adjusted Cox model across the nine scenarios of Study 2. Because the baseline fatal hazard λ_D was fixed, the death-component win and loss proportions remained stable across all scenarios, with treatment wins hovering around 8.7–8.9%, and control wins ranging from 5.9% to 8.3% (Figure 3.3). As the event ratio $r = \lambda_H/\lambda_D$ increased from 2 to 6, the non-fatal component contributed a progressively larger share of resolved comparisons and the tie proportion decreased correspondingly (67.7% at $r = 2$ to 32.9% at $r = 6$ under the death-driven pattern).

Table 3.2: (Study 1) Performance under the null hypothesis ($\eta_H = \eta_D = 0$, $WR_{\text{true}} = 1$). All quantities are reported on the log scale. For win ratio methods, $\hat{\theta} = \widehat{WR}$ and for the Cox model, $\hat{\theta} = 1/\widehat{HR}$. Emp. SE is the empirical standard deviation of $\log \hat{\theta}$ across replications. Est. SE is the median of the standard error estimates. Coverage is the proportion of 95% confidence intervals containing WR_{true}.

Confounding	Censoring	Method	Mean $\log \hat{\theta}$	Emp. SE	Est. SE	Coverage	$P < 0.05$
None	Admin. only	Unadjusted	0.080	0.188	0.179	92.8%	7.2%
		IPCW	0.080	0.187	0.179	92.8%	7.2%
		IPTW	-0.002	0.193	0.185	95.6%	4.4%
		IPTW+IPCW	0.012	0.191	0.183	95.6%	4.4%
		Cox	-0.006	0.192	0.185	—	4.8%
	Moderate	Unadjusted	0.029	0.240	0.243	96.4%	3.6%
		IPCW	0.023	0.268	0.268	94.8%	5.2%
		IPTW	-0.045	0.251	0.249	95.6%	4.4%
		IPTW+IPCW	-0.017	0.272	0.271	94.8%	5.2%
		Cox	-0.045	0.236	0.240	—	4.8%
Moderate	Admin. only	Unadjusted	0.184	0.178	0.178	84.8%	15.2%
		IPCW	0.183	0.177	0.178	84.4%	15.6%
		IPTW	0.115	0.184	0.183	90.0%	10.0%
		IPTW+IPCW	0.125	0.181	0.182	90.0%	10.0%
		Cox	0.110	0.182	0.184	—	9.6%
	Moderate	Unadjusted	0.180	0.248	0.243	90.4%	9.6%
		IPCW	0.143	0.287	0.270	91.2%	8.8%
		IPTW	0.109	0.250	0.249	92.4%	7.6%
		IPTW+IPCW	0.104	0.288	0.273	94.4%	5.6%
		Cox	0.100	0.239	0.240	—	6.4%
Strong	Admin. only	Unadjusted	0.426	0.187	0.174	29.6%	70.4%
		IPCW	0.423	0.188	0.173	30.4%	69.6%
		IPTW	0.375	0.188	0.177	44.4%	55.6%
		IPTW+IPCW	0.379	0.187	0.176	41.6%	58.4%
		Cox	0.370	0.189	0.180	—	54.0%
	Moderate	Unadjusted	0.445	0.253	0.233	52.8%	47.2%
		IPCW	0.385	0.285	0.260	66.0%	34.0%
		IPTW	0.392	0.258	0.237	60.8%	39.2%
		IPTW+IPCW	0.354	0.285	0.261	69.6%	30.4%
		Cox	0.384	0.243	0.231	—	37.6%

Table 3.3: (Study 1) Performance under non-null treatment effects ($WR_{\text{true}} > 1$) with administrative censoring only. All log-scale quantities follow the same conventions as Table 3.2. The proportion with $P < 0.05$ estimates power.

Effect	Confounding	Method	Mean $\log \hat{\theta}$	Emp. SE	Est. SE	Bias	Coverage	$P < 0.05$
		Unadjusted	0.303	0.187	0.189	0.103	90.4%	37.2%
		IPCW	0.302	0.187	0.189	0.102	90.4%	37.2%
	None	IPTW	0.223	0.197	0.194	0.023	93.6%	20.0%
		IPTW+IPCW	0.235	0.193	0.193	0.035	94.4%	22.8%
		Cox	0.219	0.195	0.196	—	—	18.0%
		Unadjusted	0.392	0.195	0.187	0.192	82.8%	56.0%
		IPCW	0.391	0.195	0.187	0.191	83.2%	57.2%
Medium	Moderate	IPTW	0.324	0.196	0.192	0.124	90.4%	40.0%
		IPTW+IPCW	0.335	0.196	0.190	0.135	88.4%	41.6%
		Cox	0.320	0.195	0.193	—	—	38.0%
		Unadjusted	0.638	0.186	0.181	0.438	32.4%	94.8%
		IPCW	0.635	0.186	0.181	0.435	32.8%	94.0%
	Strong	IPTW	0.581	0.194	0.184	0.381	43.2%	86.4%
		IPTW+IPCW	0.586	0.191	0.183	0.386	43.2%	88.0%
		Cox	0.575	0.193	0.191	—	—	85.6%
		Unadjusted	0.464	0.205	0.195	0.114	90.8%	64.8%
		IPCW	0.464	0.205	0.195	0.114	90.8%	65.6%
	None	IPTW	0.383	0.212	0.202	0.033	95.6%	51.6%
		IPTW+IPCW	0.399	0.209	0.200	0.049	95.2%	53.6%
		Cox	0.381	0.210	0.206	—	—	48.0%
		Unadjusted	0.530	0.201	0.192	0.180	83.2%	76.8%
		IPCW	0.530	0.201	0.192	0.180	83.2%	76.0%
Large	Moderate	IPTW	0.462	0.203	0.196	0.112	90.4%	62.8%
		IPTW+IPCW	0.471	0.201	0.195	0.121	88.8%	65.6%
		Cox	0.458	0.202	0.202	—	—	61.6%
		Unadjusted	0.783	0.209	0.188	0.433	38.8%	97.6%
		IPCW	0.780	0.210	0.188	0.430	38.4%	97.6%
	Strong	IPTW	0.733	0.209	0.191	0.383	46.8%	96.4%
		IPTW+IPCW	0.739	0.208	0.190	0.389	46.4%	96.8%
		Cox	0.727	0.210	0.200	—	—	95.6%

Table 3.4: (Study 1) Performance under non-null treatment effects ($WR_{\text{true}} > 1$) with moderate censoring. All log-scale quantities follow the same conventions as Table 3.2. The proportion with $P < 0.05$ estimates power.

Effect	Confounding	Method	Mean $\log \hat{\theta}$	Emp. SE	Est. SE	Bias	Coverage	$P < 0.05$
Medium	None	Unadjusted	0.315	0.251	0.255	0.115	92.8%	23.6%
		IPCW	0.307	0.297	0.283	0.107	90.8%	21.2%
		IPTW	0.242	0.258	0.262	0.042	96.0%	13.6%
		IPTW+IPCW	0.267	0.299	0.285	0.067	93.2%	14.8%
		Cox	0.237	0.246	0.254	—	—	14.4%
	Moderate	Unadjusted	0.405	0.259	0.253	0.205	87.2%	35.6%
		IPCW	0.374	0.282	0.281	0.174	91.6%	25.6%
		IPTW	0.332	0.262	0.259	0.132	92.0%	26.0%
		IPTW+IPCW	0.335	0.282	0.282	0.135	93.6%	20.4%
		Cox	0.317	0.236	0.251	—	—	19.2%
	Strong	Unadjusted	0.654	0.246	0.245	0.454	54.8%	73.6%
		IPCW	0.606	0.298	0.273	0.406	69.2%	60.8%
		IPTW	0.604	0.252	0.250	0.404	64.4%	66.0%
		IPTW+IPCW	0.580	0.300	0.275	0.380	74.4%	56.8%
		Cox	0.594	0.240	0.249	—	—	66.8%
Large	None	Unadjusted	0.426	0.261	0.264	0.076	94.4%	35.2%
		IPCW	0.419	0.308	0.294	0.069	92.0%	31.6%
		IPTW	0.348	0.276	0.271	-0.002	96.4%	24.4%
		IPTW+IPCW	0.379	0.311	0.297	0.029	92.8%	26.0%
		Cox	0.342	0.266	0.264	—	—	25.6%
	Moderate	Unadjusted	0.564	0.293	0.258	0.214	81.6%	56.0%
		IPCW	0.558	0.324	0.289	0.208	82.4%	46.8%
		IPTW	0.500	0.295	0.265	0.150	86.8%	46.8%
		IPTW+IPCW	0.522	0.326	0.291	0.172	84.8%	40.4%
		Cox	0.497	0.276	0.262	—	—	47.6%
	Strong	Unadjusted	0.751	0.300	0.251	0.401	66.4%	82.0%
		IPCW	0.710	0.325	0.279	0.360	73.2%	68.8%
		IPTW	0.699	0.305	0.255	0.349	74.4%	74.4%
		IPTW+IPCW	0.681	0.327	0.281	0.331	75.6%	63.6%
		Cox	0.697	0.277	0.259	—	—	75.6%

Table 3.5: (Study 2) Point estimates and power for the IPTW-adjusted win ratio and covariate-adjusted Cox model under differential component treatment effects (total treatment effect fixed at $\eta_H + \eta_D = 0.45$, measured confounding only, and administrative censoring only). Both methods adjust for the measured covariate X. For the IPTW-WR, $\hat{\theta} = \widehat{WR}^{\text{IPTW}}$, and for the Cox model, $\hat{\theta} = 1/\widehat{HR}$. Median proportion of pairwise comparisons won or lost on each endpoint for the IPTW-WR are reported. Composite hierarchical outcome consists of one fatal (Death) and one non-fatal (NF) outcome.

							Median Attributable Win Proportions				
Pattern	λ_H/λ_D	Method	Mean $\log\hat{\theta}$	Emp. SE	Est. SE	$P < 0.05$	Death Win	Death Loss	NF Win	NF Loss	Tie
Death-driven	2	IPTW-WR	0.120	0.150	0.155	14.4%	8.8%	5.9%	8.3%	9.3%	67.7%
		Cox	0.097	0.149	0.155	10.8%					
	4	IPTW-WR	0.075	0.111	0.117	7.6%	8.9%	5.9%	18.3%	19.5%	47.1%
		Cox	0.053	0.111	0.117	6.4%					
	6	IPTW-WR	0.064	0.094	0.101	8.8%	8.9%	5.9%	25.6%	26.4%	32.9%
		$\log(1/\widehat{HR})$	0.045	0.093	0.100	6.8%					
Balanced	2	IPTW-WR	0.239	0.155	0.160	32.0%	8.7%	6.9%	8.3%	6.5%	69.6%
		Cox	0.236	0.155	0.161	30.4%					
	4	IPTW-WR	0.227	0.126	0.121	47.2%	8.9%	6.9%	19.0%	15.1%	50.1%
		Cox	0.223	0.127	0.121	45.2%					
	6	IPTW-WR	0.214	0.104	0.104	50.0%	8.8%	6.8%	26.8%	21.7%	35.8%
		$\log(1/\widehat{HR})$	0.213	0.103	0.103	49.6%					
Non-fatal-driven	2	IPTW-WR	0.297	0.167	0.163	43.6%	8.7%	8.3%	8.3%	4.3%	70.5%
		Cox	0.319	0.167	0.165	47.2%					
	4	IPTW-WR	0.359	0.128	0.126	80.0%	8.7%	8.3%	19.3%	11.4%	52.3%
		Cox	0.382	0.125	0.126	85.6%					
	6	IPTW-WR	0.375	0.103	0.108	95.2%	8.7%	8.2%	27.9%	16.8%	38.3%
		$\log(1/\widehat{HR})$	0.397	0.104	0.107	96.8%					

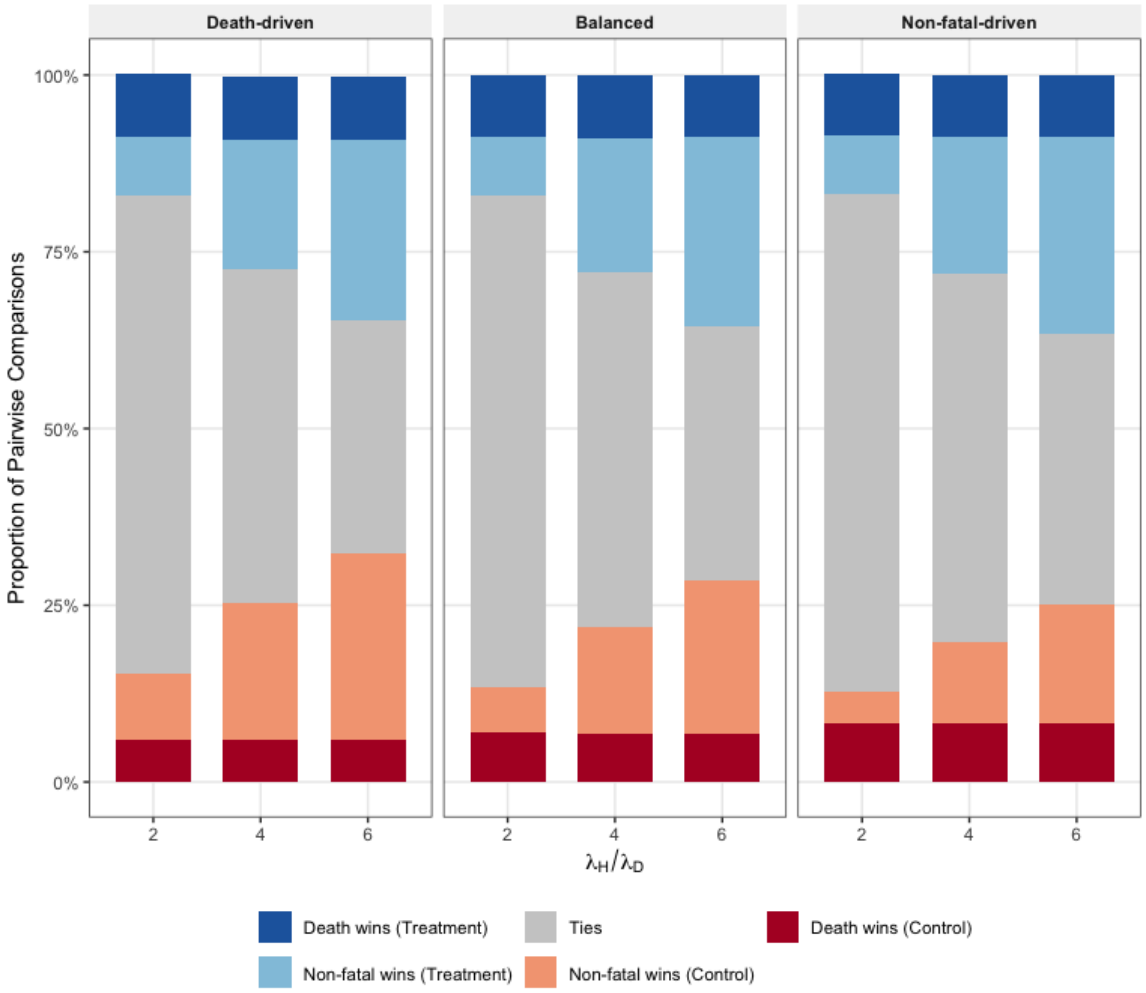

Figure 3.3: (Study 2) Median proportion of pairwise comparisons resolved at each level of the hierarchy for the IPTW-adjusted WR across the nine scenarios.

Under balanced treatment effects ($\eta_D = \eta_H = 0.225$), the IPTW-WR and Cox model produced nearly identical log-scale point estimates and power (Figure 3.4). At $r = 6$, the IPTW-WR achieved 50.0% power and the Cox model 49.6%. Under death-driven treatment effects ($\eta_D = 0.40$, $\eta_H = 0.05$), the IPTW-WR produced slightly larger point estimates than the Cox model ($\bar{L} = 0.120$ vs. 0.097 at $r = 2$). However, both methods had low estimated power at higher event ratios (7.6% and 6.4% at $r = 4$; 8.8% and 6.8% at $r = 6$), only slightly

above the 5% level. Under non-fatal-driven treatment effects ($\eta_D = 0.05$, $\eta_H = 0.40$), the Cox model produced larger log-scale point estimates (0.382 vs. 0.359 at $r = 4$; 0.397 vs. 0.375 at $r = 6$) and achieved higher power (85.6% vs. 80.0% at $r = 4$; 96.8% vs. 95.2% at $r = 6$).

The concordance classification in Table 3.6 summarizes the agreement in apparent treatment benefit between the two methods across replications. Under death-driven effects at $r = 2$, the IPTW-WR detected the treatment signal in 3.6% of replications where the Cox model did not, while the reverse never occurred, yielding a net discordance of +3.6%. Under balanced effects, net discordance was near zero (+0.4% to +2.0%). Under non-fatal-driven effects, at $r = 4$, the Cox model rejected in 5.6% of replications where the IPTW-WR did not, yielding the largest absolute net discordance observed (−5.6%). At $r = 6$, the high overall power of both methods (95.2% and 96.8%) shrank the discordance to −1.6%.

3.5 Discussion

These two simulation studies addressed complementary gaps in the win ratio literature. Study 1 evaluated the IPTW-adjusted win ratio under unmeasured confounding, a scenario not previously studied. Study 2 introduced a concordance classification to characterize when the win ratio and Cox model reach different conclusions about treatment benefit.

The Study 1 results confirmed that the IPTW-adjusted win ratio is approximately unbiased with type I error and coverage near the nominal rates of 5% and 95%, respectively, when the propensity score model is correctly specified. This is consistent with the theoretical results of Zhang (2022) [23] and the simulation findings of Wang et al. (2025) [25]. When unmeasured confounders were introduced, the treatment model was misspecified, leading to an incomplete IPTW adjustment. The subsequent bias increased monotonically with confounding strength, and coverage shrank due to both point-estimate bias and variance underestimation. The covariate-adjusted Cox model exhibited similar bias throughout, confirming that this vulnerability is a general property of covariate-adjustment methods rather than a weakness specific to the win ratio. An important practical consequence is that confounding can inflate apparent power, as rejection rates exceeded 85% under strong confounding not because of improved sensitivity but because of inflated point estimates. The IPCW

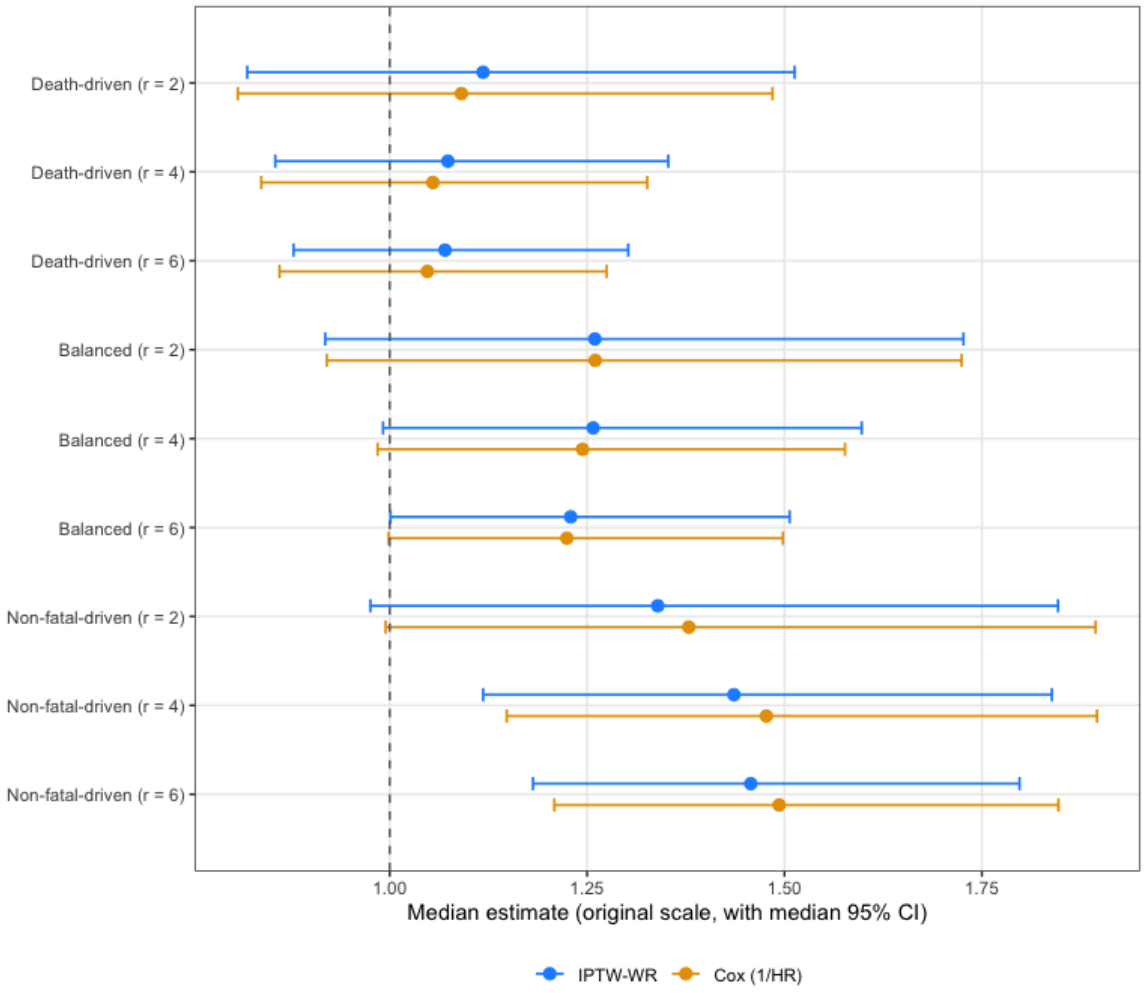

Figure 3.4: (Study 2) Median point estimates (original scale) with median 95% confidence intervals for the IPTW-adjusted WR and covariate-adjusted Cox model ($1/\widehat{HR}$) across the nine Study 2 scenarios.

Table 3.6: (Study 2) Concordance classification between the IPTW-adjusted win ratio and covariate-adjusted Cox model. Both methods adjust for the measured covariate X. For each replication, the significance of each method at $\alpha = 0.05$ is classified into one of four categories: both significant, WR only significant, Cox only significant, or neither significant. Net discordance is defined as $\hat{P}$(WR significant only) − $\hat{P}$(Cox significant only). Positive values indicate that the IPTW-WR detects the treatment benefit more frequently than the Cox model, and vice versa.

Pattern	λ_H/λ_D	Both	WRonly	Cox only	Neither	Net Disc.
Death-driven	2	10.8%	3.6%	0.0%	85.6%	+3.6%
	4	6.0%	1.6%	0.4%	92.0%	+1.2%
	6	6.4%	2.4%	0.4%	90.8%	+2.0%
Balanced	2	29.6%	2.4%	0.8%	67.2%	+1.6%
	4	44.8%	2.4%	0.4%	52.4%	+2.0%
	6	46.0%	4.0%	3.6%	46.4%	+0.4%
Non-fatal-driven	2	42.8%	0.8%	4.4%	52.0%	-3.6%
	4	80.0%	0.0%	5.6%	14.4%	-5.6%
	6	95.2%	0.0%	1.6%	3.2%	-1.6%

adjustment had negligible effect under administrative-only censoring, as expected [19], and provided modest improvements in coverage under moderate random censoring, especially when combined with IPTW [20].

Study 2 revealed that the relative performance of the two methods depends on the apportionment of the treatment effect across components. Under balanced effects, the methods were nearly equivalent. When the treatment primarily reduced mortality, the WR's hierarchical prioritization gave it a modest power advantage. When the treatment primarily reduced non-fatal events, the pattern reversed. Here, the Cox model achieved higher power because the TFE composite was dominated by the component with the strongest signal,

while the WR spent its first-priority comparisons on the near-null death component, attenuating the non-fatal signal. This reversal complements Dong et al. (2020), who showed that the WR can reveal mortality benefits masked by non-fatal events in TFE analysis [18]. Our results demonstrate the converse.

The MESA-calibrated parameters influenced several features of the results. The low baseline hazards produced sparse events and high tie proportions, contributing to the moderate power in Study 1 and low power in the death-driven Study 2 scenarios. The modest MESA-calibrated censoring rate also limited the ability for IPCW to substantially improve performance. A simulation calibrated to a higher-risk population or heavier censoring might yield different conclusions about the practical importance of IPCW. Our work adds to the small literature on win statistics in observational settings by characterizing the consequences of propensity score model misspecification due to unmeasured confounding [22, 23, 27].

Several limitations should be noted. The $B = 250$ replications limit the accuracy of estimated proportions to approximately ± 1.4 percentage points. The exponential marginals guarantee a time-invariant WR_{true}, though such a model may not reflect reality for many observational studies [24, 32]. Under non-exponential distributions, the WR estimand may depend on the follow-up horizon, introducing additional ambiguity. The single measured and unmeasured covariate is a simplification of the high-dimensional settings encountered in practice. Finally, censoring was independent in all scenarios, and we did not implement formal sensitivity analyses for unmeasured confounding. These limitations suggest several directions for future work: evaluation under non-exponential marginals where the WR estimand is time-dependent, expansion to greater number of covariates, and incorporation of dependent censoring using the methods of Dong et al. (2021) [20].

Chapter 4

APPLICATION TO THE MULTI-ETHNIC STUDY OF ATHEROSCLEROSIS

4.1 Motivation and Aims

We now demonstrate the practical utility of the win ratio and related methods to real observational data from the Multi-Ethnic Study of Atherosclerosis (MESA) [29]. MESA is a cohort study of subclinical cardiovascular disease that is well-suited for this purpose. The abundance of fatal and non-fatal cardiovascular outcomes permits the construction of hierarchical composite endpoints of varying complexity, and the multi-ethnic composition of the cohort allows for the novel application of the win ratio to demographic subgroup comparisons in an observational setting. The applied analysis is structured around three aims. The first aim is to evaluate the unadjusted win ratio dynamically over the follow-up period. By computing the win statistics at successive time points, we can investigate whether long-term demographic disparities in fatal events are obscured by early non-fatal events in a conventional TFE analysis [18]. The simulation study demonstrated under controlled conditions that the win ratio analysis detects mortality-driven treatment signals that a TFE analysis may miss, and this applied analysis permits a similar investigation in real data.

The second aim is to explore the incorporation of continuous subclinical measures, namely baseline coronary artery calcium (CAC), as the lowest-priority component in the composite hierarchical outcome. Cardiovascular trials such as EMPULSE have used continuous measures as final components of the composite endpoint to enhance the power of the win ratio to detect a difference by resolving more ties [13]. We adapt this approach to an observational setting, using CAC to break ties among participants who remain event-free throughout follow-up. Also, by constructing composite hierarchical outcomes of increasing complexity, we can assess how the number and type of components affects the win ratio

estimate, the proportion of tied pairs, and the resulting statistical power. Moreover, we illustrate how a broad spectrum of cardiovascular disease can be represented in a single composite hierarchical outcome.

The third aim is to apply both the IPTW-adjusted win ratio and the covariate-adjusted PW regression model to adjust the demographic comparisons for covariates [21, 25]. As established in Chapter 2, these methods target different estimands. By employing both methods, we illustrate how different adjustment paradigms operate in practice and formally quantify disparities in cardiovascular outcomes between demographic groups after controlling for baseline risk factors. The simulation study showed that the IPTW-adjusted win ratio is approximately unbiased when the assumption of no unmeasured confounding holds, but can exhibit substantial residual bias when unmeasured confounders are present. In the MESA analysis, unmeasured confounding cannot be completely ruled out. With this limitation in mind, we can still examine how the adjusted and unadjusted win ratios compare, and interpret the adjusted results accordingly.

This applied analysis is motivated in part by the well-documented racial and ethnic disparities in cardiovascular mortality within the MESA cohort [40]. The win ratio approach offers a complementary perspective on these disparities by incorporating the clinical hierarchy of cardiovascular events and prioritizing fatal over non-fatal outcomes. Throughout this analysis, we conduct the analogous time-to-first event analysis and compare the inverse hazard ratio to the win ratio.

4.2 Data and Methods

4.2.1 The MESA Cohort

The Multi-Ethnic Study of Atherosclerosis (MESA) is a prospective, population-based cohort study of 6814 men and women aged 45 to 84 years at enrollment, recruited from six communities in the United States between 2000 and 2002 [29]. Participants self-identified as White, Black, Hispanic/Latino, or Chinese American and were free of clinical cardiovascular disease at baseline. Participants were followed for incident cardiovascular events and mortality through the calendar year 2021, with follow-up still ongoing, primarily through

telephone interviews at 9 to 12 month intervals, supplemented by clinic visits, medical record abstractions, and death certificate review. All cardiovascular events were adjudicated by a centralized events committee using standardized criteria, and all participant deaths were captured and verified. All baseline covariates were collected at Exam 1 of MESA.

We conducted a complete-case analysis, excluding individuals with missing values for any baseline covariate used in the adjustment models (age, sex, race/ethnicity, body mass index, current smoking status, diabetes, hypertension medication use, total cholesterol, HDL cholesterol, years of education, and baseline coronary artery calcium score). This exclusion yielded a working sample of $N = 6715$ participants.

4.2.2 Hierarchical Composite Outcomes

We constructed three composite hierarchical outcomes of increasing complexity. In each, components are listed in decreasing order of clinical priority. For any pairwise comparison, the most important component is evaluated first, and the comparison proceeds to the next component only if the higher-priority component does not determine a winner, exhausting the entire hierarchy if needed and ending up with a tie if no winner was determined.

The simplest outcome consists of two time-to-event components: (1) cardiovascular death and (2) non-fatal CVD, defined as time-to-first non-fatal myocardial infarction (MI), resuscitated cardiac arrest (RCA), or non-fatal stroke. Cardiovascular death includes death attributed to atherosclerotic coronary heart disease, stroke, other atherosclerotic disease, or other cardiovascular disease.

The second outcome extends the hierarchy to five time-to-event components: (1) cardiovascular death, (2) non-fatal MI, RCA, or stroke, (3) heart failure (HF), (4) transient ischemic attack (TIA) or angina pectoris, and (5) coronary revascularization. This richer hierarchy captures a broader cascade of atherosclerotic cardiovascular disease and creates more opportunities for breaking ties during pairwise comparisons.

The third outcome adds baseline coronary artery calcium (CAC) score as a sixth component at the bottom of the hierarchy, following the five time-to-event components above. CAC is a CT-derived measure of calcified plaque in the coronary arteries that serves as a marker

of subclinical atherosclerosis and is a strong independent predictor of future cardiovascular events [29]. Because CAC is a continuous baseline measure rather than a time-to-event outcome, comparisons at this level use a different rule where the participant with the lower CAC score is considered the winner. This component serves as a final tie-breaker among participants who remain event-free throughout the follow-up period, following the approach of modern cardiovascular trials that incorporate continuous measures at the bottom of the hierarchy to increase the proportion of resolved comparisons [13].

4.2.3 Exposures and Covariates

The primary exposures of interest are sex (Female vs. Male) and race/ethnicity (Black vs. White, Hispanic/Latino vs. White, Chinese vs. White). In each comparison, we use the all-against-all comparison approach where every member of the exposure group is compared with every member of the reference group exactly once following the hierarchical procedure described above. Covariates for adjustment were selected based on the progressive adjustment strategy of Post et al. (2022), who evaluated racial and ethnic differences in all-cause mortality and cardiovascular mortality in MESA using Cox proportional hazards models at subsequent levels of adjustment [40]. We adapted the following sequential adjustment strategy for the IPTW and Cox analyses: a demographic model (age at baseline, plus sex or race/ethnicity as appropriate for the comparison), a demographic plus socioeconomic status (SES) model (adding years of education as a continuous measure), and a fully adjusted model (adding body mass index, current smoking status, diabetes, hypertension medication use, total cholesterol, and HDL cholesterol). These covariates are summarized in Table 4.1.

4.2.4 Statistical Methods

Unadjusted Win Statistics

For each composite outcome and each subgroup comparison, we computed the unadjusted win ratio (WR), net benefit (NB), and win odds (WO) using the all-against-all paradigm. These three statistics all test the same null hypothesis of equal win probabilities between groups and yield approximately equal z-values [10]. We report 95% confidence intervals and

p-values for each [14, 15]. To illustrate how each component of the hierarchy contributes to the all-against-all comparison procedure, we also report component-wise win, loss, and tie proportions across replications. These proportions indicate the fraction of all pairwise comparisons resolved at each level of the hierarchy and the fraction remaining as ties after all components have been evaluated.

Concurrently, we fit a Cox proportional hazards model for the time-to-first-event (TFE) composite corresponding to each outcome and report the inverse hazard ratio for comparability with the win ratio. As discussed in Chapter 2, the WR and 1/HR coincide only when the composite consists of a single time-to-event outcome and the proportional hazards assumption holds [32]. For the hierarchical composite endpoints considered here, the two methods estimate different quantities. The Cox model is reported as a descriptive comparator to illustrate how demographic differences manifest under the conventional TFE analysis relative to the hierarchical win ratio approach. For the most complex outcome that includes CAC, the Cox model is not applicable to the continuous CAC component and is therefore fit only to the five time-to-event components.

Win Statistics Over Time

To assess the stability of the estimated win ratio across the follow-up period, we computed the IPTW-adjusted win ratio at successive time horizons of 5, 10, 15, and 20 years. The IPTW weights for this analysis use the demographic adjustment model (age and sex or race/ethnicity). At each horizon, pairwise comparisons are restricted to events occurring before the specified time, so that the resulting win ratio reflects the cumulative evidence up to that point. Plotting the WR, NB, and WO over time provides evidence on whether the observed demographic disparities shift as the composition of the resolved comparisons evolves with a longer follow-up time [18]. This analysis is conducted for the simple two-component composite outcome only.

Inverse Probability of Treatment Weighting

To adjust for measured confounders, we applied the stabilized inverse-probability-of-treatment-weighted (IPTW) win ratio [25]. As discussed in Section 4.1, the exposure variables of interest in this analysis are non-modifiable demographic characteristics rather than a randomized treatment assignments. The IPTW weights are therefore better understood as standardization weights that balance the distribution of measured covariates across groups.

For each subgroup comparison, the propensity score was estimated via logistic regression of the group indicator on the relevant covariates, with the covariate set determined by the adjustment level: demographic, demographic plus SES, or fully adjusted (Section 4.2.3). Stabilized weights were computed as $w_i = \hat{P}(A = 1)/\hat{e}(X_i)$ for exposed participants and $w_i = \hat{P}(A = 0)/(1 - \hat{e}(X_i))$ for reference participants, where $\hat{e}(X_i)$ is the estimated propensity score, and weights were trimmed at the 1st and 99th percentiles to reduce the influence of extreme values.

We computed four win ratio variants at each adjustment level: the unadjusted WR (repeated for reference), the IPCW-adjusted WR (censoring weights only), the IPTW-adjusted WR (treatment weights only), and the combined IPTW+IPCW-adjusted WR. For the parallel Cox analysis at each adjustment level, we fit a covariate-adjusted Cox proportional hazards model with the same covariates included directly in the model, yielding a conditional hazard ratio. As noted in Chapter 3, the IPTW-adjusted WR estimates a marginal (population-averaged) win ratio while the covariate-adjusted Cox estimates a conditional hazard ratio. These are not directly comparable due to non-collapsibility, but the comparison illustrates how the two analytic approaches characterize the same data.

Proportional Win-Fractions Regression

To complement the IPTW analysis with a regression-based approach, we fit the proportional win-fractions (PW) regression model of Mao and Wang (2021) to the simple composite outcome [21]. As described in Section 2.4, the PW model directly regresses the win and loss probabilities on covariates through a log-linear model, yielding covariate-adjusted win ratios that are interpretable as the multiplicative change in the win ratio per unit change

in each covariate, holding other covariates constant. Unlike the IPTW approach, which estimates a marginal win ratio, the PW model estimates a conditional win ratio analogous to an adjusted hazard ratio.

For each of the four subgroup comparisons, we fit a PW regression model with the subgroup indicator as the primary exposure variable and demographic covariates as adjustment variables. In the Female versus Male comparison, the model included an indicator for female sex, age, and a race/ethnicity variable. In the three race/ethnicity comparisons (Black vs. White, Hispanic vs. White, Chinese vs. White), the model included the respective race/ethnicity indicator, age, and sex. The race/ethnicity models were subset to each respective pair of comparators. We report the estimated regression coefficient ($\hat{\beta}$), its standard error, the exponentiated coefficient as a covariate-adjusted win ratio ($e^{\hat{\beta}}$), the associated 95% Wald confidence interval, and the p-value from the Wald test.

To assess the proportional win-fractions assumption, we also examined the standardized score processes for the primary subgroup indicator in each model. Under the null hypothesis that the model is correctly specified, this process follows an approximate standard Brownian motion and should fluctuate randomly around zero [21]. These diagnostic plots are presented in Appendix B.

4.3 Results

4.3.1 Baseline Characteristics

Table 4.1 summarizes the baseline characteristics of the 6715 MESA participants included in the analysis, stratified by sex and race/ethnicity. The cohort was 53% female, with a median age of 62 years (IQR: 53–70). The racial and ethnic composition was 39% White, 27% Black, 22% Hispanic/Latino, and 12% Chinese American.

Several cardiovascular risk factor distributions varied across demographic groups. Black participants had the highest proportion of hypertension medication users (50%), diabetes (18%), and current smoking (18%), along with the highest median BMI (29.4 kg/m^2). Hispanic/Latino participants had an elevated prevalence of diabetes (18%) and the lowest median educational attainment (12.0 years, IQR: 8.0–13.0). Chinese participants had the

lowest median BMI (23.7 kg/m^2), the lowest smoking prevalence (5.6%), and the lowest hypertension medication prevalence (29%). White participants had the highest median educational attainment (16.0 years, IQR: 13.0–18.0) and the highest proportion with CAC scores $\geq$ 300 (17%).

Half of participants overall had a baseline CAC score of zero. This proportion was markedly higher among females (60%) than males (39%), and higher among Black (56%) and Hispanic (55%) participants than among White (43%) and Chinese (50%) participants. Median follow-up was 19.4 years.

Table 4.1: Baseline characteristics of MESA participants ($N = 6{,}715$) by sex and race/ethnicity. Values are median (Q1, Q3) for continuous variables and n (%) for categorical variables.

	Overall	**By Sex**		**By Race/Ethnicity**			
Variable	$N = 6{,}715$	Female	Male	White	Chinese	Black	Hispanic
		$n = 3{,}546$	$n = 3{,}169$	$n = 2{,}590$	$n = 798$	$n = 1{,}841$	$n = 1{,}486$
Age at Exam 1	62 (53, 70)	62 (53, 70)	62 (53, 70)	63 (54, 71)	62 (53, 70)	62 (53, 70)	61 (53, 69)
Female, n (%)	3,546 (53%)	—	—	1,351 (52%)	412 (52%)	1,015 (55%)	768 (52%)
BMI (kg/m^2)	27.6 (24.5, 31.2)	27.7 (24.3, 32.3)	27.5 (24.7, 30.3)	27.1 (24.2, 30.4)	23.7 (21.7, 25.9)	29.4 (26.1, 33.5)	28.7 (26.0, 31.9)
Current smoker, n (%)	874 (13%)	413 (12%)	461 (15%)	298 (12%)	45 (5.6%)	330 (18%)	201 (14%)
Diabetes, n (%)	842 (13%)	398 (11%)	444 (14%)	155 (6.0%)	103 (13%)	323 (18%)	261 (18%)
HTN medication, n (%)	2,502 (37%)	1,361 (38%)	1,141 (36%)	861 (33%)	229 (29%)	926 (50%)	486 (33%)
Total cholesterol (mg/dL)	192 (170, 215)	197 (176, 220)	186 (165, 209)	194 (173, 217)	191 (171, 210)	188 (166, 211)	195 (172, 220)
HDL cholesterol (mg/dL)	48 (40, 59)	54 (45, 64)	43 (37, 51)	50 (41, 61)	48 (40, 56)	50 (41, 61)	45 (39, 54)
Education (years)	13.0 (12.0, 16.0)	13.0 (12.0, 16.0)	13.0 (12.0, 16.0)	16.0 (13.0, 18.0)	13.0 (11.0, 16.0)	13.0 (12.0, 16.0)	12.0 (8.0, 13.0)
CAC score, n (%)							
0	3,360 (50%)	2,129 (60%)	1,231 (39%)	1,115 (43%)	398 (50%)	1,032 (56%)	815 (55%)
1–99	1,776 (26%)	852 (24%)	924 (29%)	695 (27%)	235 (29%)	456 (25%)	390 (26%)
100–299	747 (11%)	306 (8.6%)	441 (14%)	350 (14%)	97 (12%)	166 (9.0%)	134 (9.0%)
300+	832 (12%)	259 (7.3%)	573 (18%)	430 (17%)	68 (8.5%)	187 (10%)	147 (9.9%)
Follow-up time (years)	19.4 (12.8, 20.4)	19.6 (13.7, 20.4)	19.2 (12.0, 20.2)	19.6 (13.8, 20.5)	19.5 (13.6, 20.4)	18.9 (11.4, 20.1)	19.2 (11.6, 20.3)

4.3.2 Unadjusted Win Statistics Across Outcome Complexities

Table 4.2 presents the unadjusted win ratio and inverse hazard ratio for each subgroup comparison across the three composite outcomes of increasing complexity. The win odds and net benefit were also computed for each comparison. These statistics yielded approximately identical z-values and significance conclusions [10]. We focus on the win ratio and inverse hazard ratio throughout.

For the sex comparison, the win ratio was $\widehat{WR} = 1.478$ (95% CI: 1.314–1.663, $p < 0.001$) for the simple composite, increasing to 1.551 (95% CI: 1.407–1.710, $p < 0.001$) for the 5-component outcome and 1.740 (95% CI: 1.629–1.860, $p < 0.001$) for the 6-component outcome with CAC. The inverse hazard ratio was $1/\widehat{HR} = 1.408$ (95% CI: 1.255–1.580, $p < 0.001$) for the simple composite and 1.517 (95% CI: 1.379–1.670, $p < 0.001$) for the 5-component composite. The monotone increase in the win ratio from 1.478 to 1.740 as endpoint complexity increased reflects the contribution of lower-priority components and the CAC tie-breaker, all of which favored females.

The Black versus White comparison exhibited a reversal in direction as outcome complexity increased. For the simple composite, $\widehat{WR} = 0.852$ (95% CI: 0.738–0.984, $p = 0.029$), indicating that Black participants experienced worse composite cardiovascular outcomes than White participants. This was consistent with $1/\widehat{HR} = 0.845$ (95% CI: 0.735–0.973, $p = 0.019$). For the 5-component outcome, the win ratio was attenuated to 0.945 (95% CI: 0.839–1.065, $p = 0.353$), and was no longer statistically significant. For the CAC outcome, the direction reversed entirely, with $\widehat{WR} = 1.201$ (95% CI: 1.109–1.301, $p < 0.001$). This reversal occurred because Black participants had lower baseline CAC burden (56% with score of 0 versus 43% for White), and the CAC tie-breaker resolved a large fraction of previously event-free pairs in favor of Black participants. The component-wise results (Table 4.3 and Appendix B) highlight the fact that Black participants experienced more fatal cardiovascular events than White participants yet had considerably less subclinical atherosclerosis as measured by CAC.

The Hispanic versus White comparison was not significant for the simple or 5-component outcomes ($\widehat{WR} = 0.945$, $p = 0.474$ and $\widehat{WR} = 0.985$, $p = 0.819$, respectively). However,

the CAC outcome yielded a significant win ratio of 1.228 (95% CI: 1.127–1.337, $p < 0.001$). Hispanic participants had comparable clinical event rates to White participants but lower CAC burden (55% with CAC = 0 versus 43% for White), which drove the win ratio above 1.0 when the tie-breaker was included. The Cox model remained non-significant across all outcome complexities ($1/\widehat{HR} = 0.954$, $p = 0.457$ for the 5-component outcome).

For the Chinese versus White comparison, the win ratio was significantly above 1.0 for all three outcomes: 1.252 ($p = 0.026$), 1.421 ($p < 0.001$), and 1.348 ($p < 0.001$). The Cox model yielded a similar result for the 5-component outcome ($1/\widehat{HR} = 1.409$, $p < 0.001$) but was not significant for the simple composite ($1/\widehat{HR} = 1.210$, $p = 0.070$). This discrepancy reflects the different estimands [32].

Table 4.2: Unadjusted win ratio ($\widehat{WR}$) and inverse hazard ratio ($1/\widehat{HR}$) across three composite outcomes of increasing complexity. Reference groups are: Male (sex), White (race/ethnicity). Bold p-values indicate significance at $\alpha = 0.05$.

		Win Ratio		Inverse Hazard Ratio	
Outcome	Subgroup	$\widehat{WR}$ (95% CI)	p	$1/\widehat{HR}$ (95% CI)	p
Simple	Female	1.478 (1.314, 1.663)	< **0.001**	1.408 (1.255, 1.580)	< **0.001**
(CV Death > Non-fatal CVD)	Black	0.852 (0.738, 0.984)	**0.029**	0.845 (0.735, 0.973)	**0.019**
	Hispanic	0.945 (0.809, 1.104)	0.474	0.891 (0.765, 1.038)	0.138
	Chinese	1.252 (1.027, 1.527)	**0.026**	1.210 (0.985, 1.486)	0.070
5-Component	Female	1.551 (1.407, 1.710)	< **0.001**	1.517 (1.379, 1.670)	< **0.001**
(CV Death > MI/RCA/Stroke	Black	0.945 (0.839, 1.065)	0.353	0.967 (0.860, 1.087)	0.574
> HF > Angina/TIA	Hispanic	0.985 (0.868, 1.118)	0.819	0.954 (0.842, 1.081)	0.457
> Revasc.)	Chinese	1.421 (1.205, 1.676)	< **0.001**	1.409 (1.184, 1.678)	< **0.001**
6-Component + CAC	Female	1.740 (1.629, 1.860)	< **0.001**	—	—
(5-Component > CAC)	Black	1.201 (1.109, 1.301)	< **0.001**	—	—
	Hispanic	1.228 (1.127, 1.337)	< **0.001**	—	—
	Chinese	1.348 (1.214, 1.496)	< **0.001**	—	—

4.3.3 Component-Wise Win and Loss Proportions

To illustrate how each component of the hierarchical outcome contributes to the overall win ratio, we present the component-wise win and loss proportions for the 6-component outcome. Following the visual framework recommended by Pocock et al. (2024), Figure 4.1 shows the proportion of all pairwise comparisons resolved as wins, ties, or losses at each level of the hierarchy for the Female versus Male comparison [4]. Table 4.3 reports the corresponding numerical values. The analogous figures and tables for the remaining three subgroups (Black, Hispanic, Chinese) are presented in Appendix B.

In the Female versus Male comparison, 3,546 × 3,169 = 11,237,274 pairs were evaluated. Cardiovascular death resolved 14.4% of all pairs (8.47% female wins, 5.94% male wins), yielding a win difference of +2.53 percentage points. Non-fatal MI, stroke, or RCA resolved an additional 13.1% of pairs, with a win difference of +2.77 points. Heart failure, angina/TIA, and revascularization resolved smaller fractions of remaining pairs, contributing win differences of +0.93, +1.60, and +0.58 points, respectively. After all five time-to-event components, 61.0% of pairwise comparisons remained tied.

The addition of baseline CAC score as the sixth and final component had a substantial impact. CAC resolved 43.3% of all remaining pairs (28.56% female wins versus 14.75% male wins), with a win difference of +13.81 percentage points. This was larger than the contribution of any single time-to-event event component, and reflects the fact that only 39% of males had a CAC score of zero compared to 60% of females. After the CAC component, only 17.7% of pairs remained tied. The overall win difference was +22.22 percentage points, of which 62% came from CAC alone, illustrates how the inclusion of a continuous baseline measure at the bottom of the hierarchy can substantially reduce the proportion of tied pairs. This is consistent with the approach used in recent cardiovascular trials such as EMPULSE and ATTRibute-CM [4, 13].

4.3.4 Win Statistics Over Time

Figures 4.2–4.5 display the IPTW-adjusted (demographic model) win ratio, net benefit, and win odds at successive time horizons of 5, 10, 15, and 20 years for the simple two-component

Table 4.3: Component-wise win and loss proportions for the 6-component outcome (Female vs. Male, 3,546 × 3,169 = 11,237,274 pairs). Each row shows the percentage of all pairwise comparisons resolved by that component. "Ties" reports the cumulative percentage remaining unresolved. "Win Diff." is the difference between win and loss proportions (positive values favor the exposure group).

Component	Wins (%)	Ties (%)	Losses (%)	Win Diff. (%)
CV Death	8.47	85.60	5.94	+2.53
MI/Stroke/RCA	7.92	72.52	5.15	+2.77
Heart Failure	2.70	68.06	1.77	+0.93
Angina/TIA	3.88	61.90	2.28	+1.60
Revascularization	0.72	61.04	0.14	+0.58
CAC Score	28.56	17.72	14.75	+13.81

composite outcome.

For the Female versus Male comparison (Figure 4.2), the win ratio was the largest at 5 years ($\widehat{WR} \approx 1.7$) and decreased over time, stabilizing near 1.45 at 20 years. The confidence intervals excluded the null throughout, indicating a persistent female advantage. The net benefit increased over time even as the win ratio declined, reflecting the fact that more pairwise comparisons were resolved at longer horizons, increasing both win and loss rates in absolute terms even as the ratio of wins to losses modestly decreased.

For the Black versus White comparison (Figure 4.3), the win ratio was approximately 0.85 at 5 years and remained below 1.0 at all subsequent time points, declining slightly toward 0.80 at 20 years. The confidence interval at 5 years included the null, but by 10 years the disparity became more apparent, and by 20 years the lower confidence bound had dropped further.

For the Hispanic versus White comparison (Figure 4.4), the win ratio remained below 1.0 at all time points (ranging between approximately 0.80 and 0.87), but the confidence intervals included the null throughout.

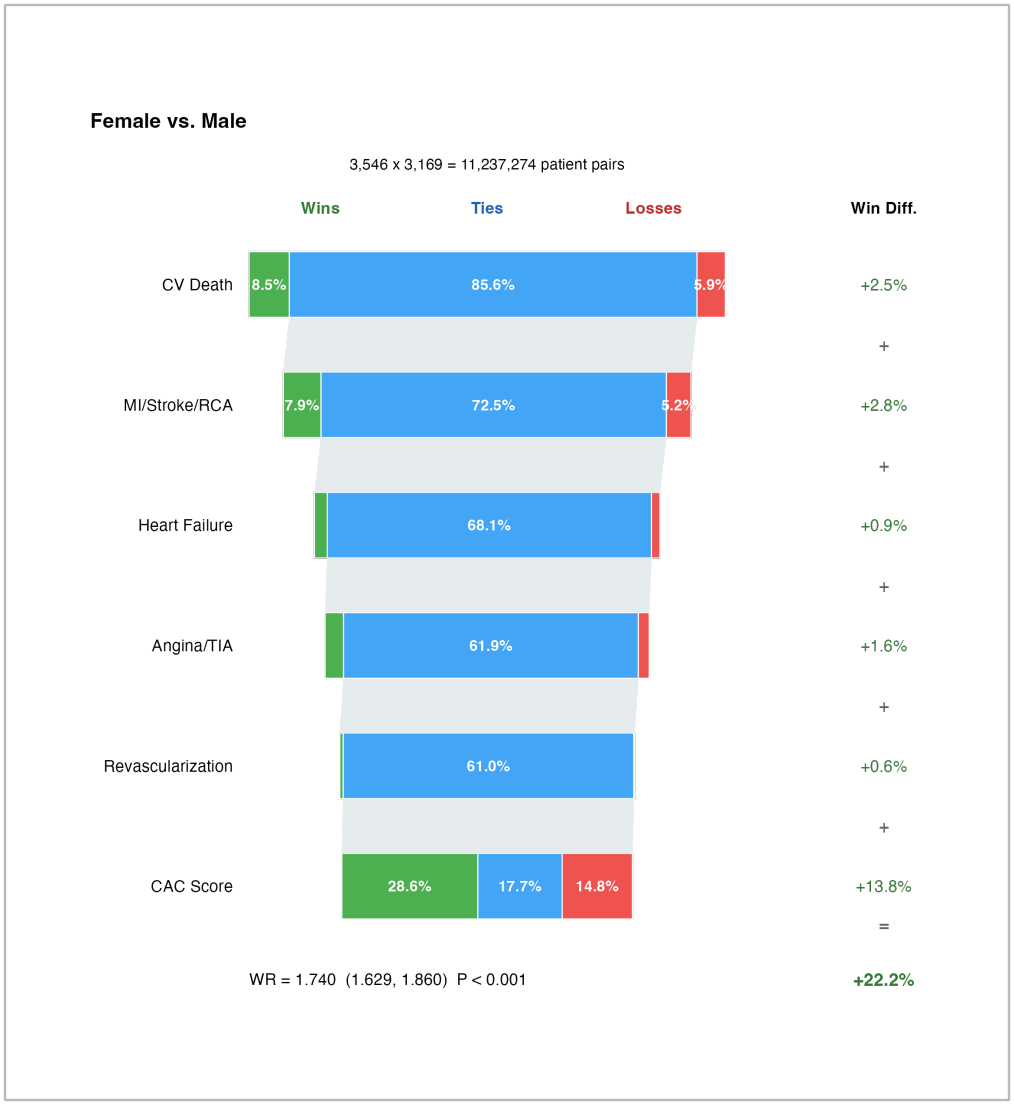

Figure 4.1: Hierarchical cascade of pairwise comparisons for the 6-component outcome (Female vs. Male). Green boxes represent the percentage of the pairwise comparisons won by the exposure group (Female) at each level. Blue boxes represent the percentage remaining as ties. Red boxes represent losses. Connecting regions show how tied pairs flow to the next level of the hierarchy. The right column reports the win difference at each level. The overall win ratio is $\widehat{WR} = 1.740$ (95% CI: 1.629–1.860, $p < 0.001$).

For the Chinese versus White comparison (Figure 4.5), the win ratio favored Chinese participants at all time points, starting at approximately 1.65 at 5 years and declining to approximately 1.2 by 20 years. The confidence intervals were wide at early-on due to the smaller sample size and fewer events but consistently indicated a favorable outcome.

Across all four subgroups, the three win statistics yielded conclusions in alignment with each other at every time point. The declining win ratio magnitude over time is expected: as the follow-up horizon increases, more events are observed and more pairwise comparisons are resolved, increasing both win and loss rates. This tends to shrink the ratio even when absolute differences are maintained or grow, consistent with the theoretical observations of Dong et al. (2020) [18].

4.3.5 IPTW-Adjusted Win Ratios

We applied the IPTW-adjusted win ratio and the covariate-adjusted Cox model at three progressive adjustment levels for each subgroup comparison. At each level, we also computed the IPCW-adjusted and combined IPCW+IPTW-adjusted win ratios. Throughout, the IPCW weights had virtually no effect on any estimate, reflecting the fact that censoring in MESA is primarily administrative. We relegate the full results that include the IPCW adjustment models to Appendix B. Here, we only report the IPTW-adjusted win ratio alongside the adjusted inverse hazard ratio for the simple outcome and the 6-component outcome with CAC in Tables 4.4 and 4.5, respectively. The 5-component outcome results are also provided in the appendix.

Simple Composite Outcome

Table 4.4 presents the IPTW-adjusted win ratio and covariate-adjusted Cox inverse hazard ratio for the simple composite outcome across the three adjustment levels. For the Female versus Male comparison, the IPTW-adjusted win ratio remained significant and stable across all adjustment levels, ranging from 1.454 (full) to 1.513 (demographic + SES). The Cox $1/\widehat{HR}$ exhibited a similar pattern (1.419 to 1.514). The female cardiovascular advantage persisted at subsequent levels covariate adjustment.

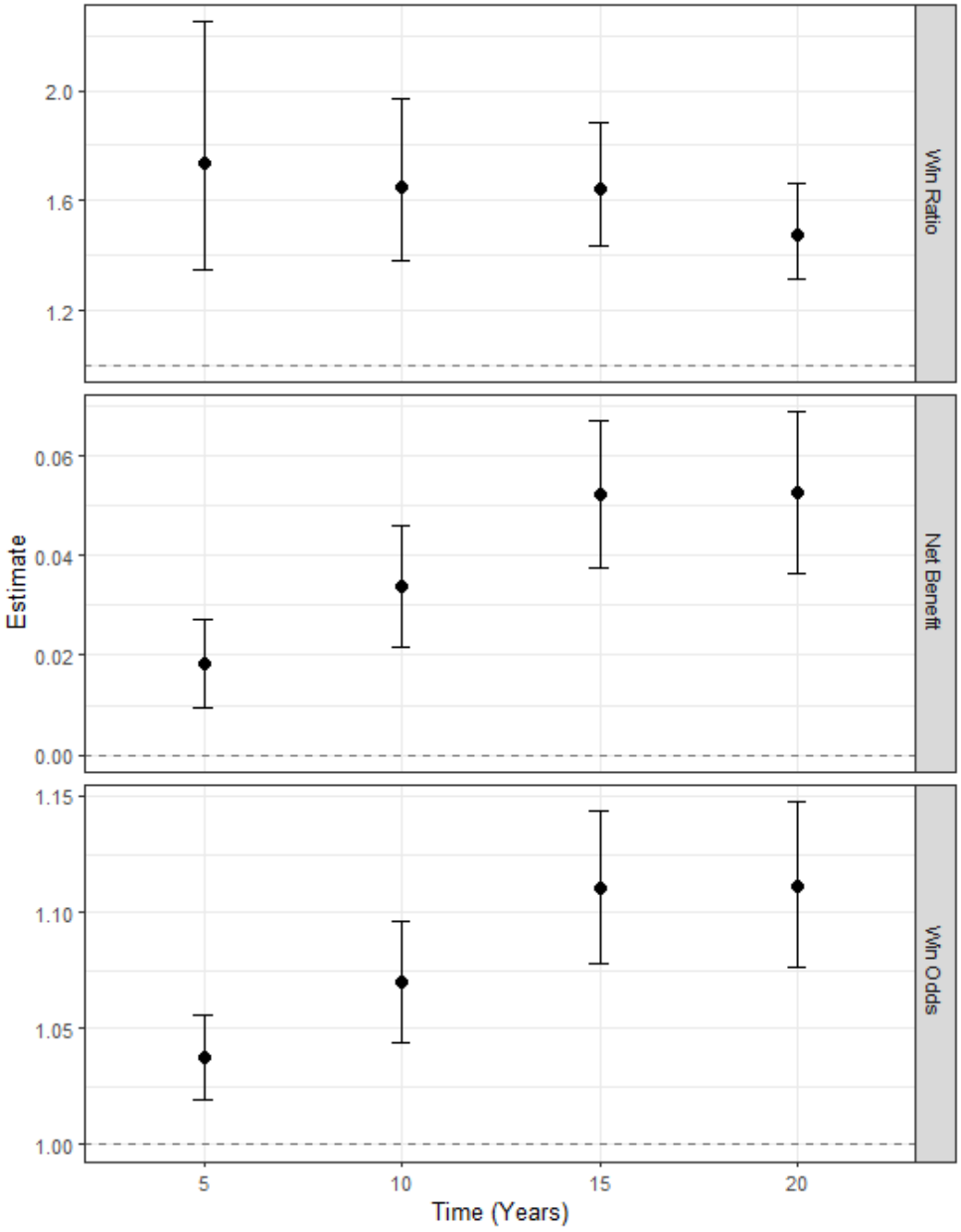

Figure 4.2: IPTW-adjusted win ratio, net benefit, and win odds over time for the simple composite outcome (CV Death > Non-fatal CVD). Female vs. Male, demographic adjustment. Dashed lines indicate the null value for each win statistic.

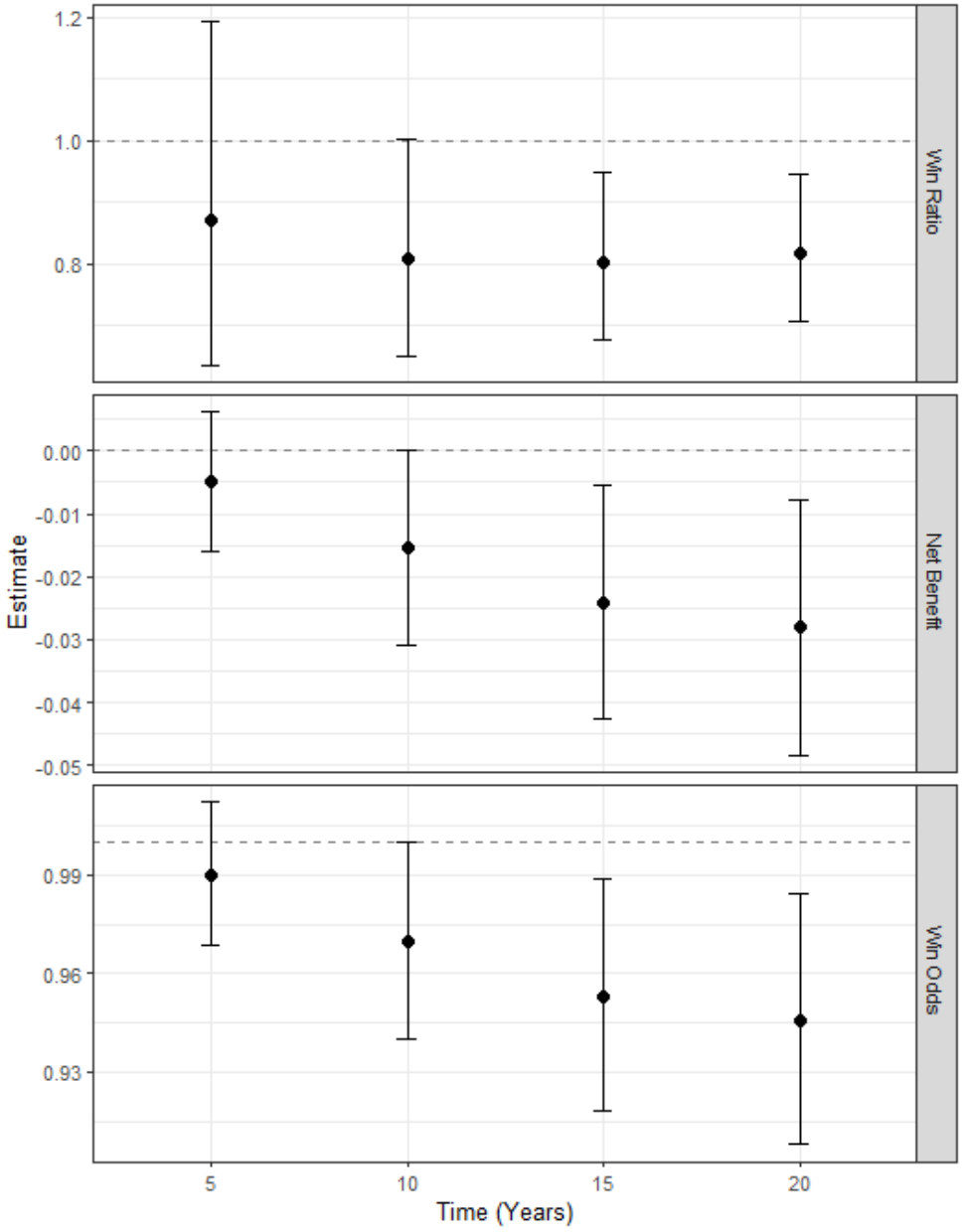

Figure 4.3: IPTW-adjusted win ratio, net benefit, and win odds over time for the simple composite outcome. Black vs. White, demographic adjustment. Dashed lines indicate the null value for each win statistic.

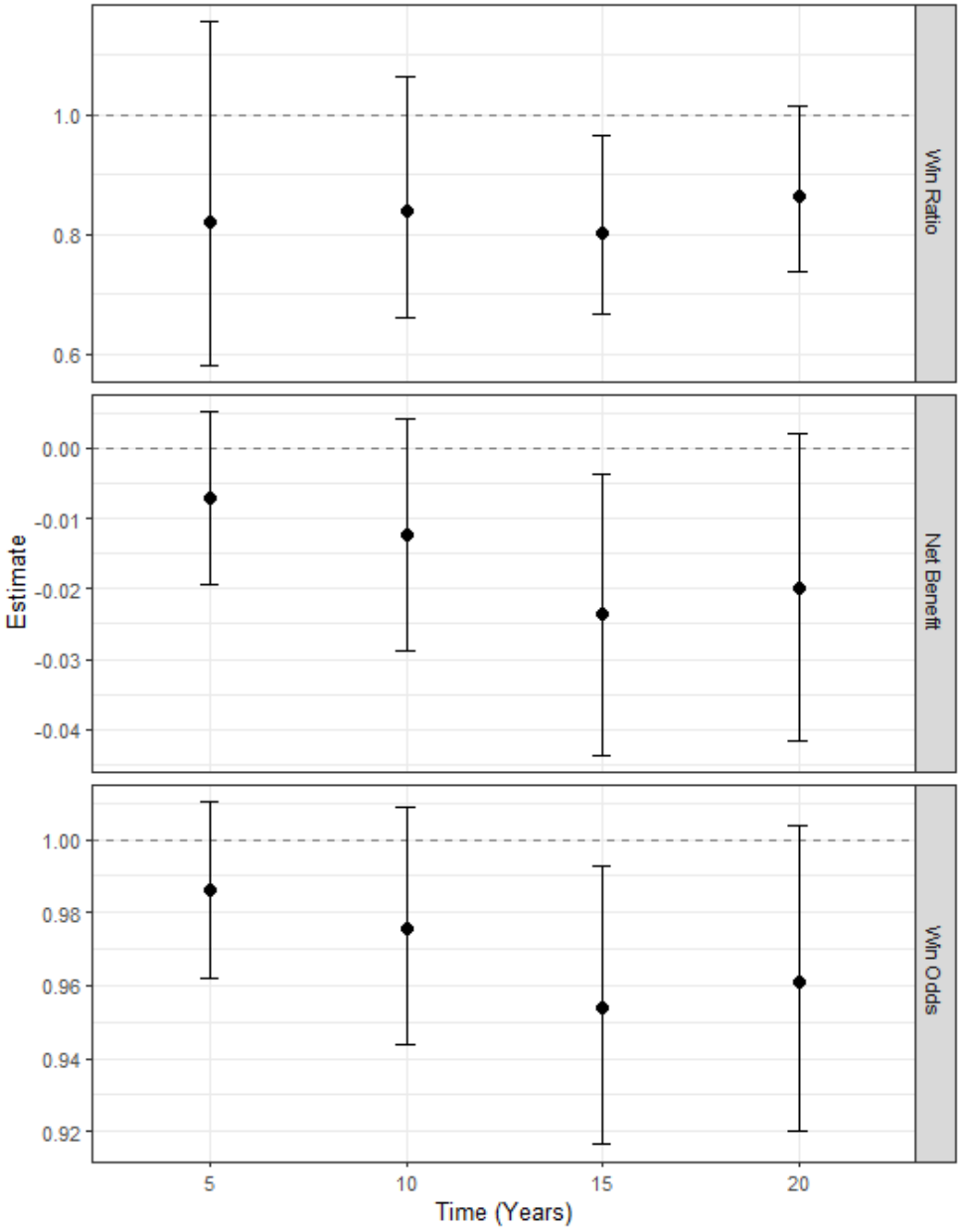

Figure 4.4: IPTW-adjusted win ratio, net benefit, and win odds over time for the simple composite outcome. Hispanic vs. White, demographic adjustment. Dashed lines indicate the null value for each win statistic.

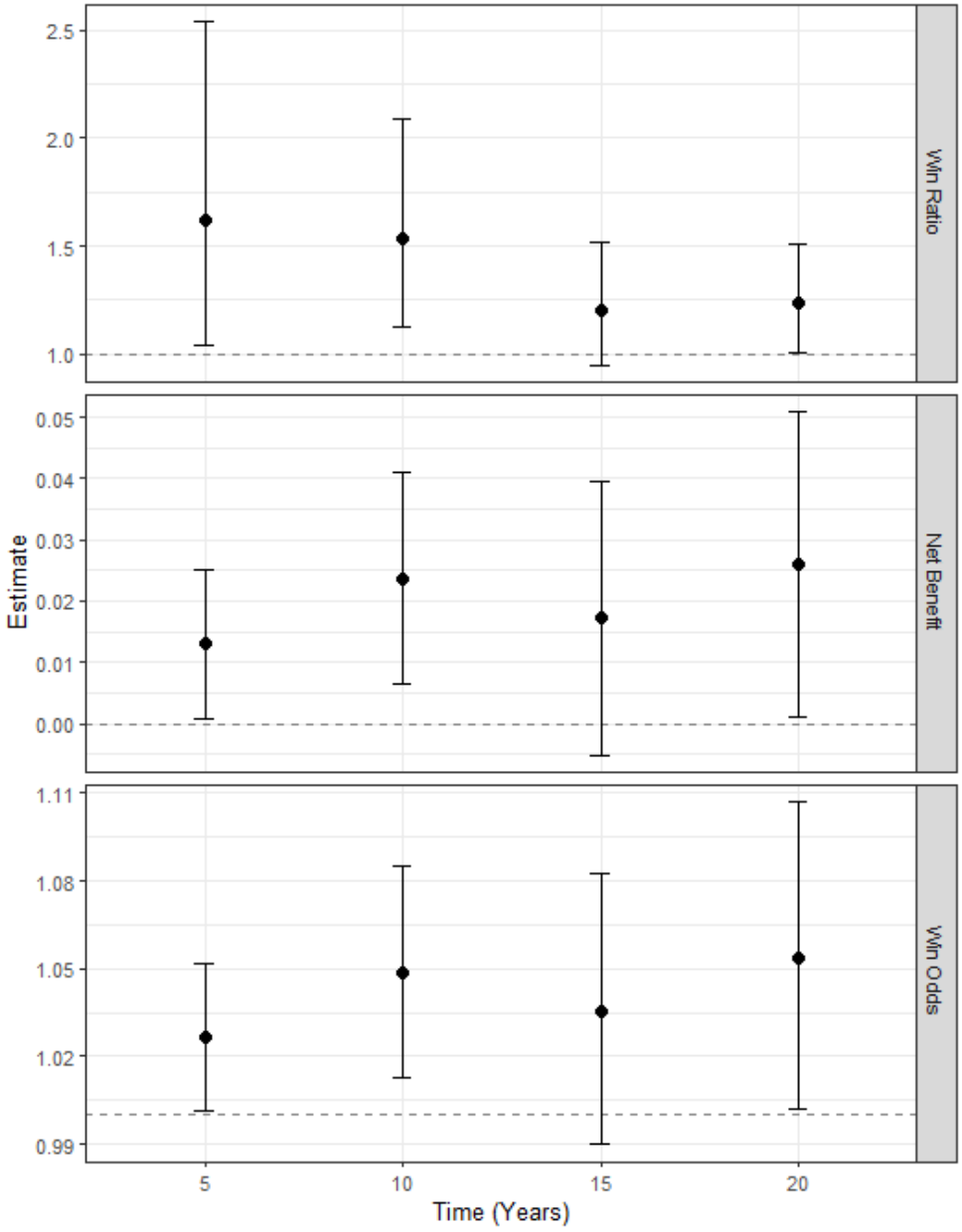

Figure 4.5: IPTW-adjusted win ratio, net benefit, and win odds over time for the simple composite outcome. Chinese vs. White, demographic adjustment. Dashed lines indicate the null value for each win statistic.

The Black versus White comparison exhibited a similar progressive attenuation of disparities as Post et al. (2022) showed using Cox regression [40]. In MESA, adjusting for age and sex, the IPTW-adjusted win ratio was 0.819 (95% CI: 0.708–0.946, $p = 0.007$), lower than the unadjusted estimate of 0.852. Adding years of education attenuated the win ratio further to 0.878 (95% CI: 0.757–1.017, $p = 0.084$), no longer significant. After full adjustment, the win ratio was 0.964 (95% CI: 0.825–1.126, $p = 0.642$), indicating that the unadjusted disparity was largely explained by measured risk factor differences. The Cox model exhibited a similar pattern, with $1/\widehat{HR} = 0.799$ ($p = 0.002$) at the demographic level, 0.853 ($p = 0.028$) after adding education, and 0.959 ($p = 0.580$) after full adjustment. The fact that both the marginal win ratio and the conditional hazard ratio converge toward the null under full adjustment provides complementary evidence that the unadjusted disparity is attributable to differences in measured risk factor distributions.

The Hispanic versus White comparison was not significant at any adjustment level for the IPTW win ratio (demographic: 0.864, $p = 0.069$; demographic + SES: 1.096, $p = 0.390$; full: 1.091, $p = 0.408$). The Cox model was significant only for demographic adjustment ($1/\widehat{HR} = 0.824$, $p = 0.013$), with further attenuation after adding education ($1/\widehat{HR} = 0.983$, $p = 0.854$). The discrepancy between the marginal IPTW-adjusted WR and the conditional Cox $1/\widehat{HR}$ at the demographic level is an expected consequence of non-collapsibility [32].

For the Chinese versus White comparison, the IPTW win ratio showed progressive attenuation from 1.232 ($p = 0.040$) at demographic adjustment to 1.060 ($p = 0.672$) after full adjustment. The Cox model followed a similar trajectory, from $1/\widehat{HR} = 1.225$ ($p = 0.053$) to 1.221 ($p = 0.093$).

Six-Component Outcome with CAC

Table 4.5 presents the IPTW-adjusted win ratio for the 6-component outcome. Because the Cox model is not applicable to the continuous CAC component and thus models a different outcome, Cox results are omitted from this table and reported separately in the appendix for the 5-component outcome.

Table 4.4: IPTW-adjusted win ratio and covariate-adjusted Cox inverse hazard ratio for the simple composite outcome (CV Death > Non-fatal CVD) across three progressive adjustment levels. Bold p-values indicate significance at $\alpha = 0.05$.

		IPTW Win Ratio			Adjusted Cox $1/\widehat{HR}$		
Adjustment	Subgroup	$\widehat{WR}$	95% CI	p	$1/\widehat{HR}$	95% CI	p
Unadjusted	Female	1.478	(1.314, 1.663)	**<0.001**	1.408	(1.255, 1.580)	**<0.001**
	Black	0.852	(0.738, 0.984)	**0.029**	0.845	(0.735, 0.973)	**0.019**
	Hispanic	0.945	(0.809, 1.104)	0.474	0.891	(0.765, 1.038)	0.138
	Chinese	1.252	(1.027, 1.527)	**0.026**	1.210	(0.985, 1.486)	0.070
Demographic	Female	1.472	(1.309, 1.656)	**<0.001**	1.459	(1.301, 1.637)	**<0.001**
	Black	0.819	(0.708, 0.946)	**0.007**	0.799	(0.694, 0.919)	**0.002**
	Hispanic	0.864	(0.738, 1.011)	0.069	0.824	(0.708, 0.960)	**0.013**
	Chinese	1.232	(1.010, 1.503)	**0.040**	1.225	(0.997, 1.506)	0.053
Demo. + SES	Female	1.513	(1.344, 1.703)	**<0.001**	1.514	(1.348, 1.701)	**<0.001**
	Black	0.878	(0.757, 1.017)	0.084	0.853	(0.739, 0.983)	**0.028**
	Hispanic	1.096	(0.890, 1.349)	0.390	0.983	(0.820, 1.179)	0.854
	Chinese	1.274	(1.029, 1.577)	**0.026**	1.347	(1.087, 1.671)	**0.007**
Full	Female	1.454	(1.266, 1.669)	**<0.001**	1.419	(1.246, 1.616)	**<0.001**
	Black	0.964	(0.825, 1.126)	0.642	0.959	(0.825, 1.114)	0.580
	Hispanic	1.091	(0.887, 1.343)	0.408	1.078	(0.899, 1.293)	0.416
	Chinese	1.060	(0.810, 1.388)	0.672	1.221	(0.968, 1.540)	0.093

For the Female versus Male comparison, the IPTW-adjusted win ratio ranged from 1.731 to 1.758 across adjustment levels, substantially larger than the corresponding simple-outcome estimates.

The Black versus White comparison for the CAC outcome showed a pattern that differs from the simple outcome. The CAC outcome yielded significant win ratios above 1.0 at every adjustment level: 1.153 ($p < 0.001$) at the demographic level, 1.202 ($p < 0.001$)

after adding education, and 1.322 ($p < 0.001$) after full adjustment. The win ratio increased with subsequent adjustment levels because accounting for clinical risk factors that are more prevalent among Black participants made their favorable CAC distribution even more apparent.

The Hispanic versus White comparison showed consistently significant IPTW-adjusted win ratios for the CAC outcome at all three levels: 1.134 ($p = 0.004$), 1.173 ($p = 0.007$), and 1.207 ($p = 0.001$).

For the Chinese versus White comparison, the IPTW WR was significant at all adjustment levels, ranging from 1.206 ($p = 0.006$) to 1.378 ($p < 0.001$).

4.3.6 Proportional Win-Fractions Regression

Table 4.6 presents the results of the PW regression models for the simple composite outcome under demographic adjustment. Each model estimates the covariate-adjusted win ratio for the primary subgroup indicator while simultaneously adjusting for age and the other demographic variable (sex or race/ethnicity).

The Female coefficient was the largest. After adjusting for age and race/ethnicity, the covariate-adjusted win ratio for Female versus Male was $e^{\hat{\beta}} = 1.853$ (95% CI: 1.434–2.393, $p < 0.001$). This is larger than the demographically adjusted IPTW estimate of 1.472 (Table 4.4), consistent with the expected difference between a conditional and marginal estimand. Age was a strong independent predictor in all four models, with each additional year of age associated with an approximately 8–10% reduction in the win ratio ($e^{\hat{\beta}} \approx 0.905$–$0.920$, $p < 0.001$ in all models). The sex covariate, when included as an adjustment variable in the three race/ethnicity models, was also significant in every case, confirming the strong female cardiovascular advantage observed throughout this analysis.

For the race/ethnicity comparisons, none of the primary subgroup indicators reached statistical significance at $\alpha = 0.05$ after demographic adjustment. In each case, the direction was consistent with the IPTW estimates at the demographic level: below 1.0 for Black and Hispanic, above 1.0 for Chinese. The wider confidence intervals relative to the IPTW approach reflect both the smaller effective sample sizes in the race-specific comparisons and

Table 4.5: IPTW-adjusted win ratio for the 6-component outcome with CAC across three progressive adjustment levels. Bold p-values indicate significance at $\alpha = 0.05$.

		IPTW Win Ratio		
Adjustment	Subgroup	$\widehat{WR}$	95% CI	p
Unadjusted	Female	1.740	(1.629, 1.860)	**<0.001**
	Black	1.201	(1.109, 1.301)	**<0.001**
	Hispanic	1.228	(1.127, 1.337)	**<0.001**
	Chinese	1.348	(1.214, 1.496)	**<0.001**
Demographic	Female	1.731	(1.620, 1.850)	**<0.001**
	Black	1.153	(1.065, 1.249)	**<0.001**
	Hispanic	1.134	(1.041, 1.235)	**0.004**
	Chinese	1.334	(1.202, 1.480)	**<0.001**
Demo + SES	Female	1.755	(1.641, 1.876)	**<0.001**
	Black	1.202	(1.108, 1.304)	**<0.001**
	Hispanic	1.173	(1.045, 1.317)	**0.007**
	Chinese	1.378	(1.231, 1.544)	**<0.001**
Full	Female	1.758	(1.627, 1.898)	**<0.001**
	Black	1.322	(1.212, 1.441)	**<0.001**
	Hispanic	1.207	(1.075, 1.355)	**0.001**
	Chinese	1.206	(1.055, 1.378)	**0.006**

the different variance properties of the regression estimator. The PW regression results are broadly consistent with the IPTW findings. Both approaches identified the female advantage as the dominant demographic signal and found no statistically significant race/ethnicity effect after demographic adjustment.

The standardized score process plots for the primary subgroup indicator in each model are presented in Appendix B (Figures B.4–B.7). In all four models, the process remained

stable throughout the follow-up period, providing no evidence against the proportional win-fractions assumption.

Table 4.6: Proportional win-fractions regression for the simple composite outcome (CV Death > Non-fatal CVD) with demographic adjustment. Each panel reports a separate model. The win ratio ($e^{\hat{\beta}}$) represents the multiplicative change in the win ratio per unit change in the covariate, holding other covariates fixed. Bold p-values indicate significance at $\alpha = 0.05$.

Model	Variable	$\hat{\beta}$	SE	$e^{\hat{\beta}}$	95% CI	p
Female vs. Male	Female	0.617	0.131	1.853	(1.434, 2.393)	< **0.001**
	Age	−0.088	0.007	0.916	(0.903, 0.929)	< **0.001**
	Race/Ethnicity[†]	−0.093	0.052	0.911	(0.823, 1.010)	0.077
Black vs. White	Black	−0.105	0.167	0.900	(0.650, 1.248)	0.528
	Age	−0.084	0.009	0.920	(0.904, 0.936)	< **0.001**
	Female	−0.528	0.161	0.590	(0.430, 0.808)	< **0.001**
Hispanic vs. White	Hispanic	−0.281	0.162	0.755	(0.549, 1.038)	0.084
	Age	−0.087	0.009	0.916	(0.901, 0.933)	< **0.001**
	Female	−0.637	0.160	0.529	(0.387, 0.723)	< **0.001**
Chinese vs. White	Chinese	0.280	0.214	1.323	(0.869, 2.014)	0.192
	Age	−0.099	0.010	0.905	(0.888, 0.923)	< **0.001**
	Female	−0.504	0.181	0.604	(0.424, 0.862)	**0.005**

[†]Categorical race/ethnicity indicator included as an adjustment variable; this model was fit on the full cohort.

4.4 Discussion

This applied analysis demonstrated the practical utility of win ratio methods for characterizing demographic disparities in cardiovascular outcomes using data from the MESA cohort. The analysis was organized around three aims: evaluating the win ratio dynamically over

follow-up, exploring the influence of outcome complexity and the incorporation of a continuous component, and applying IPTW-adjusted win ratios alongside covariate-adjusted Cox models under progressive adjustment.

The WR-over-time analysis of the simple composite outcome revealed that the win ratio magnitude generally declined over the follow-up period, consistent with the theoretical results of Dong et al. (2020) [18]. For the Female versus Male comparison, the win ratio decreased from approximately 1.7 at 5 years to 1.45 at 20 years. This attenuation occurs because, as the follow-up time horizon increases, more events are observed and more pairwise comparisons are resolved, increasing both win and loss rates in absolute terms. Even when absolute differences are maintained, the ratio of wins to losses tends to decrease as both quantities grow.

For the Black versus White comparison, the trajectory of the WR over time was informative in a different way. The win ratio was approximately 0.85 at 5 years and declined further to approximately 0.80 by 20 years. The confidence intervals excluded the null at later time points but not at 5 years, indicating that the excess cardiovascular burden among Black participants relative to White participants accumulated gradually and became more statistically apparent with longer follow-up. This pattern shows the importance of long follow-up periods for detecting disparities in composite cardiovascular outcomes, especially when fatal events are relatively rare in the early years. It also highlights a practical advantage of the win ratio, where because of its prioritization of mortality in the hierarchy, the growing signal was driven by the accumulating mortality differential rather than being masked by early non-fatal events, as might occur in a conventional time-to-first-event analysis.

Constructing composite hierarchical outcomes of increasing complexity produced several important insights. The monotone increase in the win ratio for the Female comparison from 1.478 (simple) to 1.740 (CAC outcome) demonstrates that adding clinically relevant components to the hierarchy can improve the power of the win ratio by reducing the proportion of tied pairs. After all five time-to-event components, 61% of pairs remained tied. The addition of baseline CAC resolved the majority of these, reducing the tie proportion to 18%. This finding supports the recommendation of Pocock et al. (2024) to incorporate

continuous patient-centered measures at the bottom of the hierarchy to enhance the discriminatory power of the composite endpoint [4]. In the MESA context, CAC functions as a measure of cumulative subclinical atherosclerosis that discriminates among participants who remained clinically event-free during the follow-up period.

The most striking finding of this analysis was the reversal of the Black versus White win ratio from below 1.0 for the simple outcome ($\widehat{WR} = 0.852$) to above 1.0 for the CAC outcome ($\widehat{WR} = 1.201$). This occurred because Black participants had substantially lower baseline CAC burden than White participants (56% vs. 43% with CAC score of 0), and the CAC tie-breaker resolved a large pool of previously tied event-free pairs in favor of Black participants, illustrating the sensitivity of the WR estimate to composition and priority ordering of the clinical hierarchy.

The progressive adjustment analysis using the IPTW-adjusted win ratio closely mirrored the Cox regression findings reported by Post et al. (2022) for all-cause and cardiovascular mortality in the MESA cohort [40]. For the Black versus White comparison using the simple composite, the win ratio was progressively attenuated from 0.819 at the demographic level to 0.964 after full adjustment, with the largest increment of attenuation occurring when education was added. The Cox $1/\widehat{HR}$ exhibited a nearly identical trajectory, from 0.799 to 0.959. The convergence of two methodologically distinct approaches toward the null under full adjustment provides complementary evidence that the unadjusted Black–White disparity in composite cardiovascular outcomes is largely attributable to differences in measured risk factor distributions.

At several points in the analysis, the IPTW-adjusted win ratio and the covariate-adjusted Cox $1/\widehat{HR}$ yielded modestly different point estimates and, in some instances, different significance conclusions. The most notable example was the Hispanic comparison at the demographic adjustment level, where the Cox model was significant ($1/\widehat{HR} = 0.824$, $p = 0.013$) while the IPTW win ratio was not ($\widehat{WR} = 0.864$, $p = 0.069$). The conditional estimate is generally further from the null than the corresponding marginal quantity even when both are estimating a genuine group-level difference [32]. These discrepancies are therefore expected consequences of the distinct estimands rather than contradictions, and they reinforce the value of reporting both approaches side by side. A recent paper by Dong

et al. (2026) provides a thorough investigation of the non-collapsibility of win statistics and recommends standardized (stratified) win statistics as one approach to addressing this issue [36].

The PW regression model provided a complementary regression-based perspective on the demographic comparisons. The covariate-adjusted Female win ratio from the PW model (1.853) was notably larger than the IPTW estimate at the same adjustment level (1.472). The conditional estimate is generally further from the null than the marginal estimate, analogous to the well-known non-collapsibility of the odds ratio and hazard ratio [32, 36]. The PW model also directly quantified the association of age with the composite hierarchical outcome, a covariate association that is implicitly absorbed into the IPTW weights but not separately estimated. For the race/ethnicity comparisons, the PW model yielded non-significant results consistent with the IPTW findings under demographic adjustment. The standardized score processes provided no evidence against the proportional win-fractions assumption in any of the four models, supporting the validity of the model over the observed follow-up period. These models used only the demographic adjustment level and the simple composite outcome due computational limitations.

There are several limitations to this applied analysis. First, unmeasured confounding cannot be ruled out in this observational setting. The simulation study demonstrated that even moderate unmeasured confounding can induce substantial bias in the IPTW-adjusted win ratio. Second, only the baseline CAC score was used for this analysis, and its inclusion in a composite outcome with time-to-event endpoints that were accumulated over the follow-up period changes the interpretation of the win ratio in a qualitative way. The practice of incorporating continuous baseline measures follows recent cardiovascular trial designs, but the interpretation in an observational context may warrant some caution [4, 13]. Third, the win ratio is known to depend on the length of follow-up [18, 24]. MESA has a median follow-up of 19.4 years, which is long by the standards of most observational cardiovascular studies. The temporal analysis showed that the win ratio estimates were generally stable after 10 years, but the values at earlier time points may not be directly comparable to estimates from studies with shorter follow-up. Fourth, the MESA cohort was free of clinical cardiovascular disease at baseline by design. The results may not generalize to populations

with prevalent cardiovascular disease, who would have different baseline risk profiles and potentially different patterns of hierarchical outcome resolution.

This analysis is among the first applications of the IPTW-adjusted win ratio to demographic subgroup comparisons in a large observational cohort, demonstrating the practical feasibility of the approach outside of the randomized trial setting. The long follow-up (median 19.4 years), diverse cohort composition, and rich event data of MESA permitted the construction of hierarchical composite outcomes of up to six components, providing a detailed view of how outcome complexity affects the win ratio. The side-by-side presentation of the win ratio and Cox inverse hazard ratio throughout the analysis illustrates the complementary information provided by the two approaches and makes the circumstances under which they agree and disagree more transparent. The progressive adjustment strategy also provides a structured framework for interpreting the sources of observed disparities.

Chapter 5

CONCLUSIONS

This thesis investigated the win ratio and related methods for analyzing composite hierarchical outcomes in observational cardiovascular studies. We addressed a gap in a literature that has developed almost exclusively in the context of randomized clinical trials. The simulation study demonstrated that the IPTW-adjusted win ratio is approximately unbiased with near-ideal coverage and type I error when the propensity score model is correctly specified, but that not accounting for unmeasured confounding in the treatment model produces monotonically increasing bias, inflated type I error, and reduced coverage. These findings apply equally to the covariate-adjusted Cox model, showing that win ratio methods are not uniquely vulnerable to unmeasured confounding. The simulation also revealed that the win ratio and Cox model can reach different conclusions when treatment effects are concentrated in different components of the hierarchy. The win ratio shows a larger apparent treatment benefit when the treatment primarily reduces mortality, while the Cox model does so when the treatment primarily reduces non-fatal events. With these complementary strengths, there may be a case for reporting both methods in certain circumstances.

The applied analysis in the MESA cohort demonstrated the practical use and value of the win ratio for demographic subgroup comparisons in a large observational setting. The progressive attenuation of the Black–White cardiovascular disparity under IPTW adjustment replicated the pattern reported by Post et al. (2022) using Cox regression [40], providing complementary evidence of the observed disparity and its attribution to the measured risk factors. The incorporation of baseline CAC score as a tie-breaking component at the bottom of the hierarchy greatly reduced the proportion of tied pairs from 61% to 18% and reversed the direction of the Black–White win ratio, illustrating the influence the composition of the clinical hierarchy has on the win ratio. The PW regression model supported the IPTW findings and directly quantified the strong independent effects of age and sex on

the composite outcome.

As hierarchical composite outcomes continue to play an important role in both clinical trials and observational studies, a thorough understanding of the properties and limitations of the win ratio and related methods under realistic conditions will be essential for their responsible use. The win ratio literature is still expanding at a rapid pace, as researchers continue to address these challenges. Extensions of the work presented in this thesis to dependent censoring, non-exponential marginals, high-dimensional data, and other realistic scenarios would further broaden this understanding and support the viability of the win ratio in observational settings.

BIBLIOGRAPHY

[1] Nancy M. Heddle and Richard J. Cook. Composite outcomes in clinical trials: What are they and when should they be used? 51(1):11–13.

[2] Victor Dayan, Stuart W Grant, James M Brophy, Fabio Barili, and Nick Freemantle. Composite end points and competing risks analysis. 39(1):ivae126.

[3] Robert Goldberg, Joel M. Gore, Bruce Barton, and Jerry Gurwitz. Individual and Composite Study Endpoints: Separating the Wheat from the Chaff. 127(5):379–384.

[4] Stuart J Pocock, John Gregson, Timothy J Collier, Joao Pedro Ferreira, and Gregg W Stone. The win ratio in cardiology trials: Lessons learnt, new developments, and wise future use. 45(44):4684–4699.

[5] Ignacio Ferreira-González, Gaiet Permanyer-Miralda, Antònia Domingo-Salvany, Jason W. Busse, Diane Heels-Ansdell, Victor M. Montori, Elie A. Akl, Dianne M. Bryant, Pablo Alonso-Coello, Jordi Alonso, Andrew Worster, Suneel Upadhye, Roman Jaeschke, Holger J. Schünemann, Valeria Pacheco-Huergo, Ping Wu, Edward J. Mills, and Gordon H. Guyatt. Problems with use of composite end points in cardiovascular trials: Systematic review of randomised controlled trials. 334(7597):786.

[6] Stuart J. Pocock, Cono A. Ariti, Timothy J. Collier, and Duolao Wang. The win ratio: A new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. 33(2):176–182.

[7] João Pedro Ferreira, Pardeep S. Jhund, Kévin Duarte, Brian L. Claggett, Scott D. Solomon, Stuart Pocock, Mark C. Petrie, Faiez Zannad, and John J.V. McMurray. Use of the Win Ratio in Cardiovascular Trials. 8(6):441–450.

[8] Dianne M. Finkelstein and David A. Schoenfeld. Combining mortality and longitudinal measures in clinical trials. 18(11):1341–1354.

[9] Marc Buyse. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. 29(30):3245–3257.

[10] Gaohong Dong, Bo Huang, Johan Verbeeck, Ying Cui, James Song, Margaret Gamalo-Siebers, Duolao Wang, David C. Hoaglin, Yodit Seifu, Tobias Mütze, and John Kolassa. Win statistics (win ratio, win odds, and net benefit) can complement one another to show the strength of the treatment effect on time-to-event outcomes. 22(1):20–33.

[11] John Gregson, Dylan Taylor, Ruth Owen, Tim Collier, David J. Cohen, and Stuart Pocock. Hierarchical Composite Outcomes and Win Ratio Methods in Cardiovascular Trials: A Review and Consequent Guidance. 151(22):1606–1619.

[12] Mathew S. Maurer, Jeffrey H. Schwartz, Balarama Gundapaneni, Perry M. Elliott, Giampaolo Merlini, Marcia Waddington-Cruz, Arnt V. Kristen, Martha Grogan, Ronald Witteles, Thibaud Damy, Brian M. Drachman, Sanjiv J. Shah, Mazen Hanna, Daniel P. Judge, Alexandra I. Barsdorf, Peter Huber, Terrell A. Patterson, Steven Riley, Jennifer Schumacher, Michelle Stewart, Marla B. Sultan, and Claudio Rapezzi. Tafamidis Treatment for Patients with Transthyretin Amyloid Cardiomyopathy. 379(11):1007–1016.

[13] Stuart J. Pocock, João Pedro Ferreira, Timothy J. Collier, Christiane E. Angermann, Jan Biegus, Sean P. Collins, Mikhail Kosiborod, Michael E. Nassif, Piotr Ponikowski, Mitchell A. Psotka, John R. Teerlink, Jasper Tromp, John Gregson, Jonathan P. Blatchford, Cordula Zeller, and Adriaan A. Voors. The Win Ratio Method in Heart Failure Trials: Lessons Learnt From EMPULSE. 25(5):632–641.

[14] Xiaodong Luo, Hong Tian, Surya Mohanty, and Wei Yann Tsai. An Alternative Approach to Confidence Interval Estimation for the Win Ratio Statistic. 71(1):139–145.

[15] Ionut Bebu and John M. Lachin. Large sample inference for a win ratio analysis of a composite outcome based on prioritized components. 17(1):178–187.

[16] Gaohong Dong, Junshan Qiu, Duolao Wang, and Marc Vandemeulebroecke. The stratified win ratio. 28(4):778–796.

[17] Lu Mao. On causal estimation using U-statistics. 105(1):215–220.

[18] Gaohong Dong, Bo Huang, Yu-Wei Chang, Yodit Seifu, James Song, and David C. Hoaglin. The win ratio: Impact of censoring and follow-up time and use with nonproportional hazards. 19(3):168–177.

[19] Gaohong Dong, Lu Mao, Bo Huang, Margaret Gamalo-Siebers, Jiuzhou Wang, GuangLei Yu, and David C. Hoaglin. The inverse-probability-of-censoring weighting (IPCW) adjusted win ratio statistic: An unbiased estimator in the presence of independent censoring. 30(5):882–899.

[20] Gaohong Dong, Bo Huang, Duolao Wang, Johan Verbeeck, Jiuzhou Wang, and David C. Hoaglin. Adjusting win statistics for dependent censoring. 20(3):440–450.

[21] Lu Mao and Tuo Wang. A Class of Proportional Win-Fractions Regression Models for Composite Outcomes. 77(4):1265–1275.

[22] Di Zhang and Jong-Hyeon Jeong. Inference on win ratio for cluster-randomized semi-competing risk data. 4(2):1263–1292.

[23] Di Zhang, Stephen R. Wisniewski, and Jong-Hyeon Jeong. Causal Inference on Win Ratio for Observational Data with Dependent Subjects.

[24] Lu Mao. Defining estimand for the win ratio: Separate the true effect from censoring. 21(5):584–594.

[25] Duolao Wang, Sirui Zheng, Ying Cui, Nengjie He, Tao Chen, and Bo Huang. Adjusted win ratio using the inverse probability of treatment weighting. 35(1):21–36.

[26] Tuo Wang, Ying Li, and Yongming Qu. Restricted Time Win Ratio: From Estimands to Estimation. 17(1):136–148.

[27] Maria Vittoria Chiaruttini, Giulia Lorenzoni, Gaya Spolverato, and Dario Gregori. Win Statistics in Observational Cancer Research: Integrating Clinical and Quality-of-Life Outcomes. 13(11):3272.

[28] Alexander Schmidt, Ikram Ullah, Sergio Luis Teruya, Denise M. Fine, Natalia Sabogal, Rachel Foster, Stephen Helmke, Carlos Rodriguez, Brian Spencer, Morgan L. Winburn, Dimitrios Bampatsias, Abdirahman Wardhere, Cinthia De Freitas, Tatiana Prokaeva, Cesia Maria Gallegos Kattan, Edward James Miller, Michael Paul LaValley, Mathew S. Maurer, and Frederick L. Ruberg. Transthyretin amyloid cardiomyopathy is associated with worse outcomes when compared to patients with non-amyloid related heart failure of similar race and ethnicity: The scan-mp study. 83:520–520.

[29] Diane E. Bild, David A. Bluemke, Gregory L. Burke, Robert Detrano, Ana V. Diez Roux, Aaron R. Folsom, Philip Greenland, David R. JacobsJr., Richard Kronmal, Kiang Liu, Jennifer Clark Nelson, Daniel O'Leary, Mohammed F. Saad, Steven Shea, Moyses Szklo, and Russell P. Tracy. Multi-Ethnic Study of Atherosclerosis: Objectives and Design. 156(9):871–881.

[30] Lu Mao. On the Alternative Hypotheses for the Win Ratio. 75(1):347–351.

[31] Dianne M. Finkelstein and David A. Schoenfeld. Graphing the Win Ratio and its components over time. 38(1):53–61.

[32] D. Oakes. On the win-ratio statistic in clinical trials with multiple types of event. 103(3):742–745.

[33] David Oakes. Bivariate Survival Models Induced by Frailties. 84(406):487–493.

[34] Gaohong Dong, David C. Hoaglin, Junshan Qiu, Roland A. Matsouaka, Yu-Wei Chang, Jiuzhou Wang, and Marc Vandemeulebroecke. The Win Ratio: On Interpretation and Handling of Ties. 12(1):99–106.

[35] Tyler J. VanderWeele and Whitney R. Robinson. On causal interpretation of race in regressions adjusting for confounding and mediating variables. 25(4):473–484.

[36] Gaohong Dong, Margaret Gamalo-Siebers, Ying Cui, Bo Huang, Xiaolong Luo, and Lu Tian. Win statistics (win ratio, win odds, and net benefit): Noncollapsibility and standardization for randomized clinical trials. 0(0):1–17.

[37] G. Enderlein. Cox, D. R.; Oakes, D.: Analysis of Survival Data. Chapman and Hall, London – New York 1984, 201 S., £ 12,–. 29(1):114–114.

[38] Ying Cui and Bo Huang. WINS: The R WINS Package.

[39] Sirui Zheng, Duolao Wang, Junshan Qiu, Tao Chen, and Margaret Gamalo. A win ratio approach for comparing crossing survival curves in clinical trials. 33(4):488–501.

[40] Wendy S. Post, Karol E. Watson, Spencer Hansen, Aaron R. Folsom, Moyses Szklo, Steven Shea, R. Graham Barr, Gregory Burke, Alain G. Bertoni, Norrina Allen, James S. Pankow, Joao A.C. Lima, Jerome I. Rotter, Joel D. Kaufman, W. Craig Johnson, Richard A. Kronmal, Ana V. Diez-Roux, and Robyn L. McClelland. Racial and Ethnic Differences in All-Cause and Cardiovascular Disease Mortality: The MESA Study. 146(3):229–239.

[41] Lu Mao and KyungMann Kim. Statistical models for composite endpoints of death and non-fatal events: A review. 13(3):260–269.

[42] Björn Redfors, John Gregson, Aaron Crowley, Thomas McAndrew, Ori Ben-Yehuda, Gregg W Stone, and Stuart J Pocock. The win ratio approach for composite endpoints: Practical guidance based on previous experience. 41(46):4391–4399.

[43] Philip Hougaard. A Class of Multivariate Failure Time Distributions. 73(3):671–678.

[44] S Von Koch, S Koul, D Erlinge, and M Mohammad. Percutaneous coronary intervention and medical therapy versus medical therapy alone in chronic coronary syndrome: A hierarchical win-ratio analysis from SCAAR. 44:ehad655.1327.

[45] Miguel A Hernan and James M Robins. Causal Inference: What If.

[46] Tyler J. VanderWeele and Peng Ding. Sensitivity Analysis in Observational Research: Introducing the E-Value. 167(4):268–274.

Appendix A

ADDITIONAL SIMULATION STUDY FIGURES

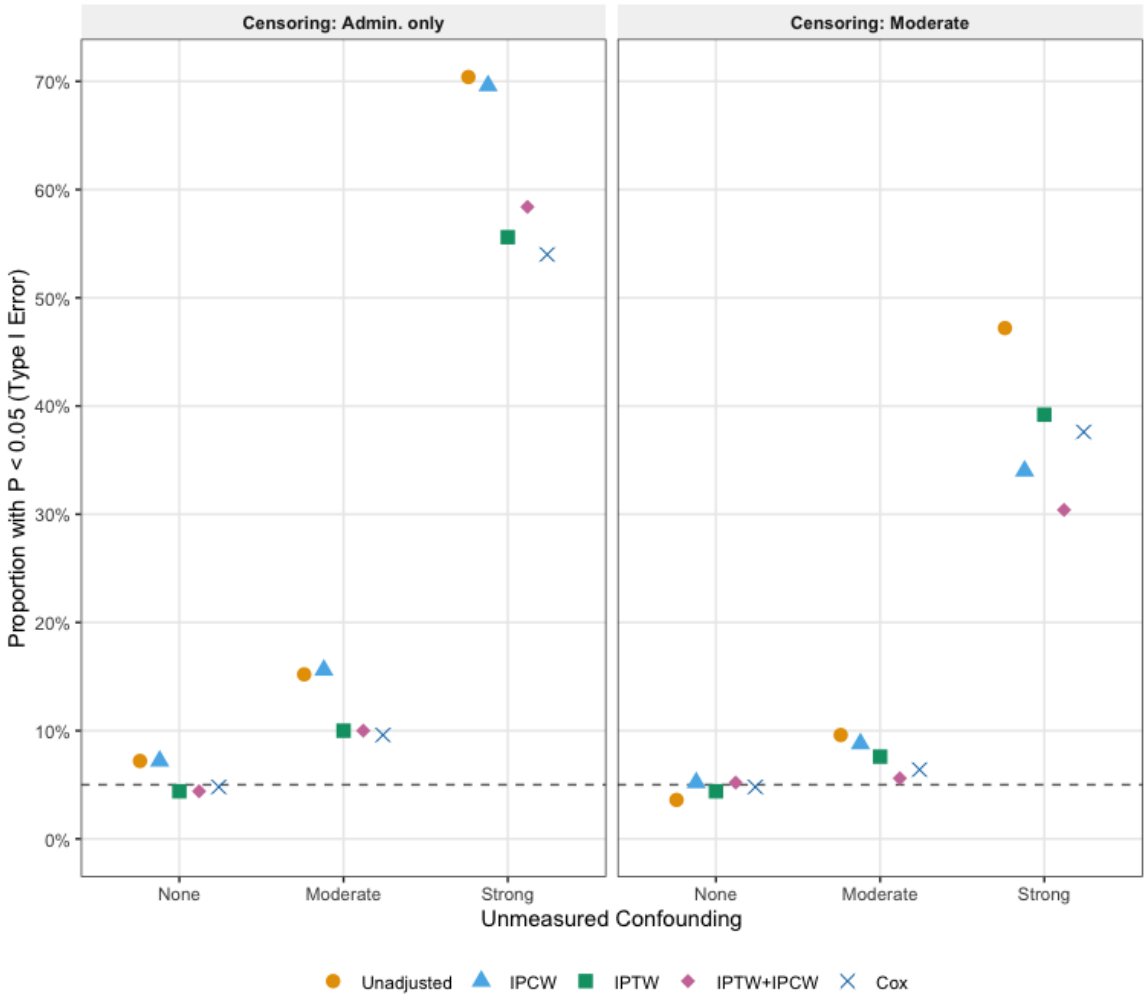

Figure A.1: (Study 1) Type I error under the null by confounding level.

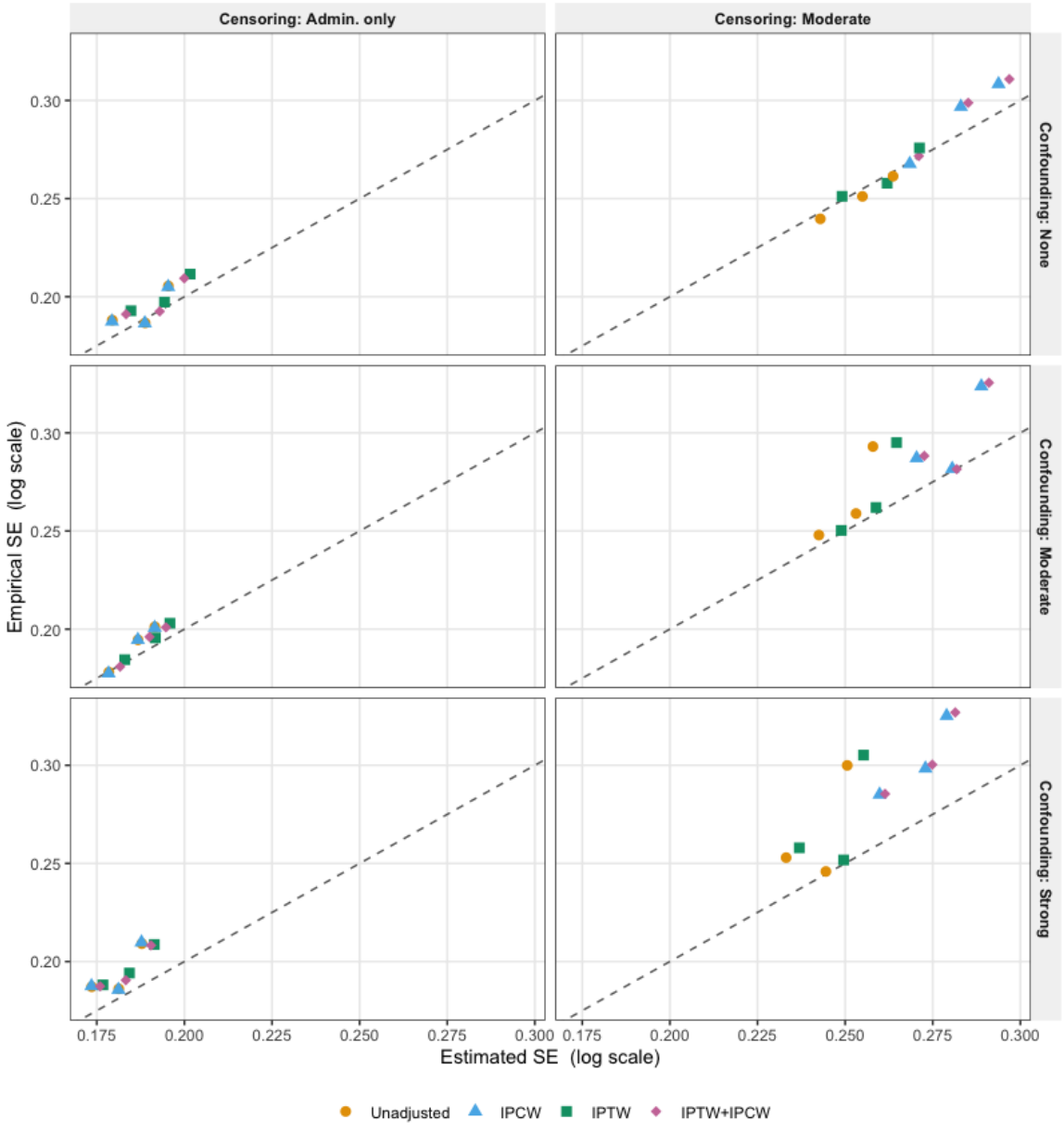

Figure A.2: (Study 1) SE calibration by confounding and censoring.

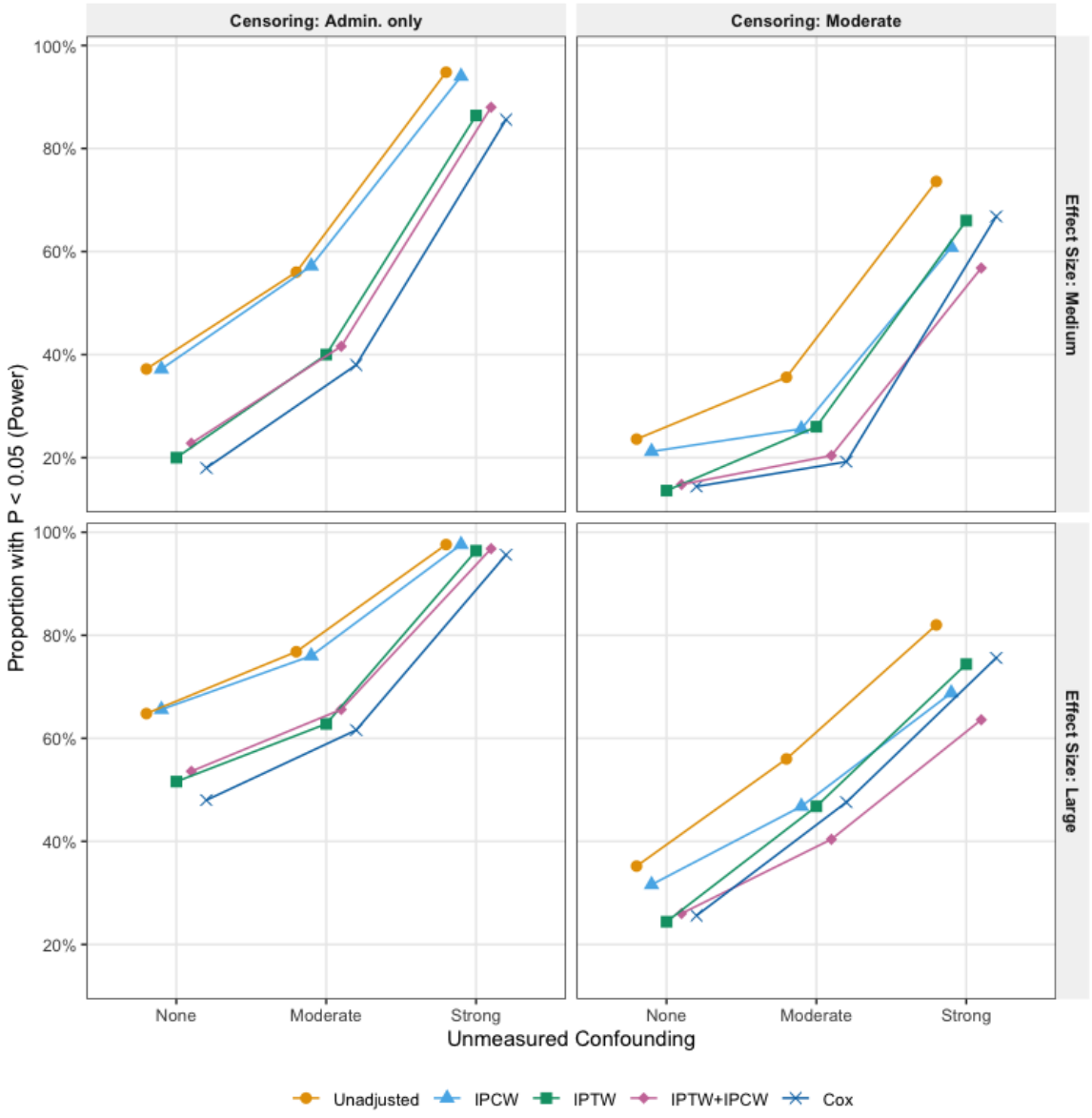

Figure A.3: (Study 1) Power under confounding (non-null scenarios).

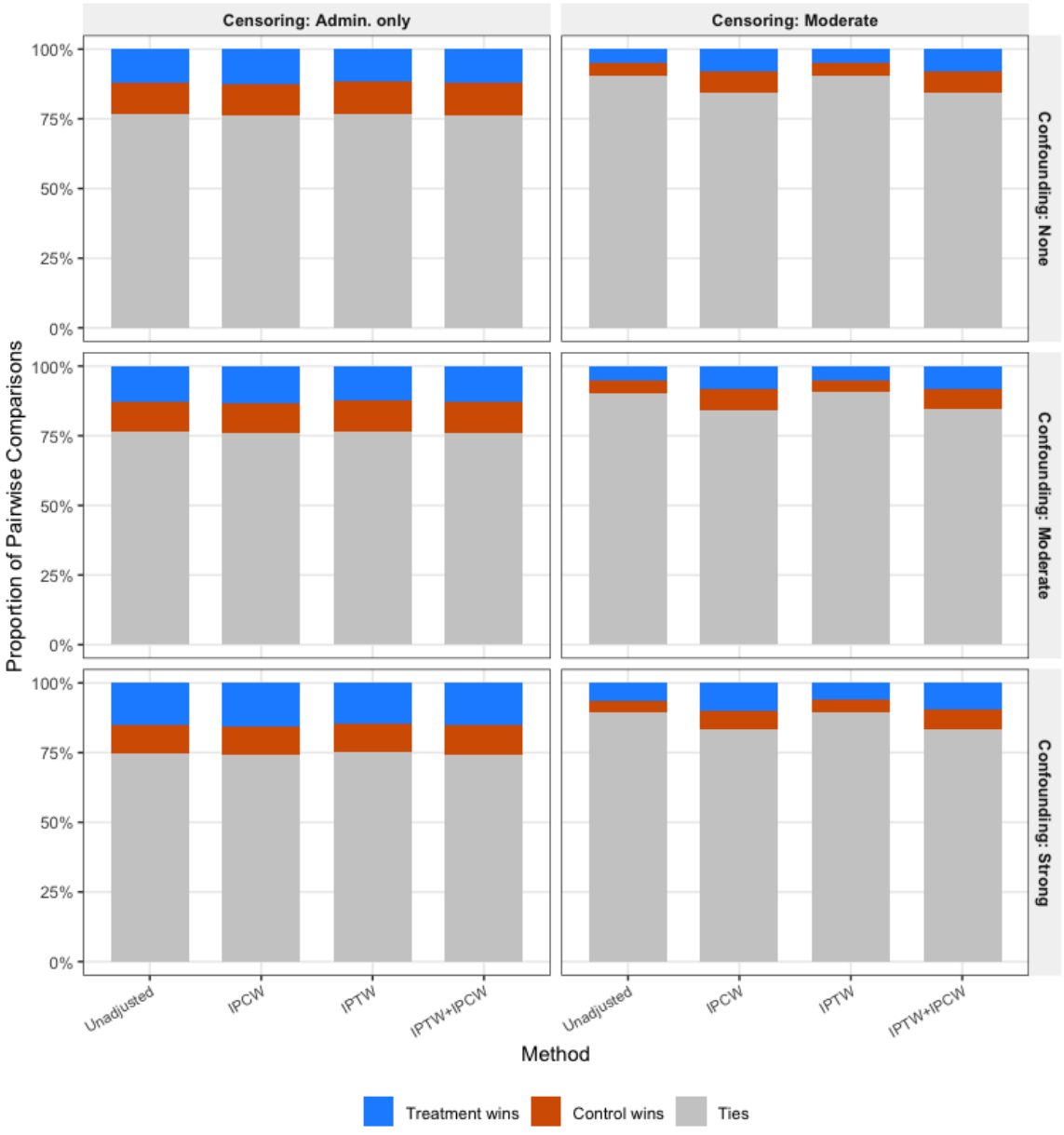

Figure A.4: (Study 1) Win proportions and ties under censoring (null scenarios).

Appendix B

ADDITIONAL APPLIED ANALYSIS TABLES AND FIGURES

Table B.1: (Table B.1) Four-variant IPW win ratio and adjusted Cox model for the simple composite outcome (CV Death > Non-fatal CVD), demographic adjustment. Bold p-values indicate significance at $\alpha = 0.05$.

	Unadjusted WR			IPCW WR			IPTW WR			IPCW+IPTW WR			Adjusted Cox		
Subgroup	WR	95% CI	p	WR	95% CI	p	WR	95% CI	p	WR	95% CI	p	1/HR	95% CI	p
Female	1.478	(1.314, 1.663)	**<.001**	1.402	(1.240, 1.585)	**<.001**	1.472	(1.309, 1.656)	**<.001**	1.399	(1.237, 1.581)	**<.001**	1.459	(1.301, 1.637)	**<.001**
Black	0.852	(0.738, 0.984)	**.029**	0.842	(0.725, 0.978)	**.024**	0.819	(0.708, 0.946)	**.007**	0.824	(0.710, 0.957)	**.011**	0.799	(0.694, 0.919)	**.002**
Hispanic	0.945	(0.809, 1.104)	.474	0.891	(0.754, 1.053)	.177	0.864	(0.738, 1.011)	.069	0.847	(0.716, 1.002)	.053	0.824	(0.708, 0.960)	**.013**
Chinese	1.252	(1.027, 1.527)	**.026**	1.129	(0.895, 1.426)	.306	1.232	(1.010, 1.503)	**.040**	1.118	(0.885, 1.413)	.349	1.225	(0.997, 1.506)	.053

Table B.2: IPW win ratio and adjusted Cox model for the simple composite outcome (CV Death > Non-fatal CVD), demographic + SES adjustment.

	Unadjusted WR			IPCW WR			IPTW WR			IPCW+IPTW WR			Adjusted Cox		
Subgroup	WR	95% CI	p	WR	95% CI	p	WR	95% CI	p	WR	95% CI	p	1/HR	95% CI	p
Female	1.478	(1.314, 1.663)	**<.001**	1.402	(1.240, 1.585)	**<.001**	1.513	(1.344, 1.703)	**<.001**	1.409	(1.246, 1.593)	**<.001**	1.514	(1.348, 1.701)	**<.001**
Black	0.852	(0.738, 0.984)	**.029**	0.842	(0.725, 0.978)	**.024**	0.878	(0.757, 1.017)	.084	0.845	(0.727, 0.982)	**.028**	0.853	(0.739, 0.983)	**.028**
Hispanic	0.945	(0.809, 1.104)	.474	0.891	(0.754, 1.053)	.177	1.096	(0.890, 1.349)	.390	0.916	(0.761, 1.103)	.354	0.983	(0.820, 1.179)	.854
Chinese	1.252	(1.027, 1.527)	**.026**	1.129	(0.895, 1.426)	.306	1.274	(1.029, 1.577)	**.026**	1.088	(0.853, 1.387)	.496	1.347	(1.087, 1.671)	**.007**

Table B.3: IPW win ratio and adjusted Cox model for the simple composite outcome (CV Death > Non-fatal CVD), full adjustment.

	Unadjusted WR			IPCW WR			IPTW WR			IPCW+IPTW WR			Adjusted Cox		
Subgroup	WR	95% CI	p	WR	95% CI	p	WR	95% CI	p	WR	95% CI	p	1/HR	95% CI	p
Female	1.478	(1.314, 1.663)	**<.001**	1.402	(1.240, 1.585)	**<.001**	1.454	(1.266, 1.669)	**<.001**	1.380	(1.212, 1.571)	**<.001**	1.419	(1.246, 1.616)	**<.001**
Black	0.852	(0.738, 0.984)	**.029**	0.842	(0.725, 0.978)	**.024**	0.964	(0.825, 1.126)	.642	0.895	(0.768, 1.044)	.158	0.959	(0.825, 1.114)	.580
Hispanic	0.945	(0.809, 1.104)	.474	0.891	(0.754, 1.053)	.177	1.091	(0.887, 1.343)	.408	0.909	(0.755, 1.096)	.319	1.078	(0.899, 1.293)	.416
Chinese	1.252	(1.027, 1.527)	**.026**	1.129	(0.895, 1.426)	.306	1.060	(0.810, 1.388)	.672	0.990	(0.755, 1.297)	.940	1.221	(0.968, 1.540)	.093

Table B.4: IPW win ratio and adjusted Cox model for the 5-component outcome (CV Death > MI/RCA/Stroke > HF > Angina/TIA > Revasc.), demographic adjustment.

	Unadjusted WR			IPCW WR			IPTW WR			IPCW+IPTW WR			Adjusted Cox		
Subgroup	WR	95% CI	p	WR	95% CI	p	WR	95% CI	p	WR	95% CI	p	1/HR	95% CI	p
Female	1.551	(1.407, 1.710)	**<.001**	1.568	(1.422, 1.730)	**<.001**	1.545	(1.401, 1.704)	**<.001**	1.563	(1.417, 1.723)	**<.001**	1.571	(1.427, 1.729)	**<.001**
Black	0.945	(0.839, 1.065)	.353	0.959	(0.851, 1.080)	.491	0.910	(0.807, 1.026)	.122	0.924	(0.820, 1.042)	.199	0.929	(0.826, 1.044)	.216
Hispanic	0.985	(0.868, 1.118)	.819	0.996	(0.878, 1.131)	.954	0.911	(0.801, 1.035)	.153	0.922	(0.811, 1.049)	.217	0.898	(0.792, 1.017)	.090
Chinese	1.421	(1.205, 1.676)	**<.001**	1.431	(1.214, 1.687)	**<.001**	1.402	(1.189, 1.654)	**<.001**	1.412	(1.197, 1.665)	**<.001**	1.460	(1.226, 1.739)	**<.001**

Table B.5: IPW win ratio and adjusted Cox model for the 5-component outcome, demographic + SES adjustment.

	Unadjusted WR			IPCW WR			IPTW WR			IPCW+IPTW WR			Adjusted Cox		
Subgroup	WR	95% CI	p	WR	95% CI	p	WR	95% CI	p	WR	95% CI	p	1/HR	95% CI	p
Female	1.551	(1.407, 1.710)	**<.001**	1.568	(1.422, 1.730)	**<.001**	1.582	(1.434, 1.746)	**<.001**	1.602	(1.451, 1.768)	**<.001**	1.617	(1.468, 1.781)	**<.001**
Black	0.945	(0.839, 1.065)	.353	0.959	(0.851, 1.080)	.491	0.970	(0.858, 1.097)	.629	0.986	(0.873, 1.115)	.823	0.982	(0.872, 1.106)	.767
Hispanic	0.985	(0.868, 1.118)	.819	0.996	(0.878, 1.131)	.954	1.028	(0.865, 1.224)	.751	1.050	(0.885, 1.247)	.575	1.011	(0.872, 1.174)	.881
Chinese	1.421	(1.205, 1.676)	**<.001**	1.431	(1.214, 1.687)	**<.001**	1.466	(1.227, 1.750)	**<.001**	1.474	(1.235, 1.759)	**<.001**	1.579	(1.317, 1.893)	**<.001**

Table B.6: IPW win ratio and adjusted Cox model for the 5-component outcome, full adjustment.

	Unadjusted WR			IPCW WR			IPTW WR			IPCW+IPTW WR			Adjusted Cox		
Subgroup	WR	95% CI	p	WR	95% CI	p	WR	95% CI	p	WR	95% CI	p	1/HR	95% CI	p
Female	1.551	(1.407, 1.710)	**<.001**	1.568	(1.422, 1.730)	**<.001**	1.562	(1.392, 1.753)	**<.001**	1.586	(1.414, 1.779)	**<.001**	1.576	(1.414, 1.757)	**<.001**
Black	0.945	(0.839, 1.065)	.353	0.959	(0.851, 1.080)	.491	1.102	(0.968, 1.254)	.141	1.113	(0.978, 1.266)	.104	1.150	(1.015, 1.302)	**.029**
Hispanic	0.985	(0.868, 1.118)	.819	0.996	(0.878, 1.131)	.954	1.064	(0.894, 1.266)	.483	1.087	(0.916, 1.291)	.338	1.105	(0.952, 1.282)	.188
Chinese	1.421	(1.205, 1.676)	**<.001**	1.431	(1.214, 1.687)	**<.001**	1.254	(1.007, 1.562)	**.043**	1.271	(1.023, 1.580)	**.030**	1.412	(1.162, 1.715)	**<.001**

Table B.7: IPW win ratio for the 6-component outcome with CAC, demographic adjustment. The Cox model is not applicable to the continuous CAC component and is omitted.

	Unadjusted WR			IPCW WR			IPTW WR			IPCW+IPTW WR		
Subgroup	WR	95% CI	p	WR	95% CI	p	WR	95% CI	p	WR	95% CI	p
Female	1.740	(1.629, 1.860)	**<.001**	1.740	(1.629, 1.860)	**<.001**	1.731	(1.620, 1.850)	**<.001**	1.731	(1.620, 1.850)	**<.001**
Black	1.201	(1.109, 1.301)	**<.001**	1.201	(1.109, 1.301)	**<.001**	1.153	(1.065, 1.249)	**<.001**	1.153	(1.065, 1.249)	**<.001**
Hispanic	1.228	(1.127, 1.337)	**<.001**	1.228	(1.127, 1.337)	**<.001**	1.134	(1.041, 1.235)	**.004**	1.134	(1.041, 1.235)	**.004**
Chinese	1.348	(1.214, 1.496)	**<.001**	1.348	(1.214, 1.496)	**<.001**	1.334	(1.202, 1.480)	**<.001**	1.334	(1.202, 1.480)	**<.001**

Table B.8: IPW win ratio for the 6-component outcome with CAC, demographic + SES adjustment. The Cox model is not applicable to the continuous CAC component and is omitted.

	Unadjusted WR			IPCW WR			IPTW WR			IPCW+IPTW WR		
Subgroup	WR	95% CI	p	WR	95% CI	p	WR	95% CI	p	WR	95% CI	p
Female	1.740	(1.629, 1.860)	**<.001**	1.740	(1.629, 1.860)	**<.001**	1.755	(1.641, 1.876)	**<.001**	1.755	(1.641, 1.876)	**<.001**
Black	1.201	(1.109, 1.301)	**<.001**	1.201	(1.109, 1.301)	**<.001**	1.202	(1.108, 1.304)	**<.001**	1.202	(1.108, 1.304)	**<.001**
Hispanic	1.228	(1.127, 1.337)	**<.001**	1.228	(1.127, 1.337)	**<.001**	1.173	(1.045, 1.317)	**.007**	1.173	(1.045, 1.317)	**.007**
Chinese	1.348	(1.214, 1.496)	**<.001**	1.348	(1.214, 1.496)	**<.001**	1.378	(1.231, 1.544)	**<.001**	1.378	(1.231, 1.544)	**<.001**

Table B.9: IPW win ratio for the 6-component outcome with CAC, full adjustment. The Cox model is not applicable to the continuous CAC component and is omitted.

	Unadjusted WR			IPCW WR			IPTW WR			IPCW+IPTW WR		
Subgroup	WR	95% CI	p	WR	95% CI	p	WR	95% CI	p	WR	95% CI	p
Female	1.740	(1.629, 1.860)	**<.001**	1.740	(1.629, 1.860)	**<.001**	1.758	(1.627, 1.898)	**<.001**	1.758	(1.627, 1.898)	**<.001**
Black	1.201	(1.109, 1.301)	**<.001**	1.201	(1.109, 1.301)	**<.001**	1.322	(1.212, 1.441)	**<.001**	1.322	(1.212, 1.441)	**<.001**
Hispanic	1.228	(1.127, 1.337)	**<.001**	1.228	(1.127, 1.337)	**<.001**	1.207	(1.075, 1.355)	**.001**	1.207	(1.075, 1.355)	**.001**
Chinese	1.348	(1.214, 1.496)	**<.001**	1.348	(1.214, 1.496)	**<.001**	1.206	(1.055, 1.378)	**.006**	1.206	(1.055, 1.378)	**.006**

Table B.10: Component-wise win and loss proportions for the 6-component outcome (Black vs. White, 1,841 × 2,590 = 4,768,190 pairs).

Component	Wins (%)	Ties (%)	Losses (%)	Win Diff. (%)
CV Death	6.36	84.97	8.68	−2.32
MI/Stroke/RCA	6.50	72.05	6.42	+0.08
Heart Failure	2.35	66.99	2.72	−0.37
Angina/TIA	3.32	61.29	2.37	+0.95
Revascularization	0.69	60.46	0.15	+0.54
CAC Score	25.50	18.07	16.89	+8.61

Table B.11: Component-wise win and loss proportions for the 6-component outcome (Hispanic vs. White, 1,486 × 2,590 = 3,848,740 pairs).

Component	Wins (%)	Ties (%)	Losses (%)	Win Diff. (%)
CV Death	6.81	86.75	6.45	+0.36
MI/Stroke/RCA	6.52	72.57	7.65	−1.13
Heart Failure	2.35	68.25	1.97	+0.39
Angina/TIA	3.32	61.29	3.64	−0.31
Revascularization	0.68	60.35	0.27	+0.41
CAC Score	25.51	18.01	16.83	+8.68

Table B.12: Component-wise win and loss proportions for the 6-component outcome (Chinese vs. White, $798 \times 2{,}590 = 2{,}066{,}820$ pairs).

Component	Wins (%)	Ties (%)	Losses (%)	Win Diff. (%)
CV Death	6.95	86.49	6.55	+0.40
MI/Stroke/RCA	6.94	75.00	4.54	+2.40
Heart Failure	2.59	71.41	1.00	+1.59
Angina/TIA	3.66	65.40	2.35	+1.31
Revascularization	0.76	64.39	0.26	+0.50
CAC Score	26.41	17.58	20.39	+6.02

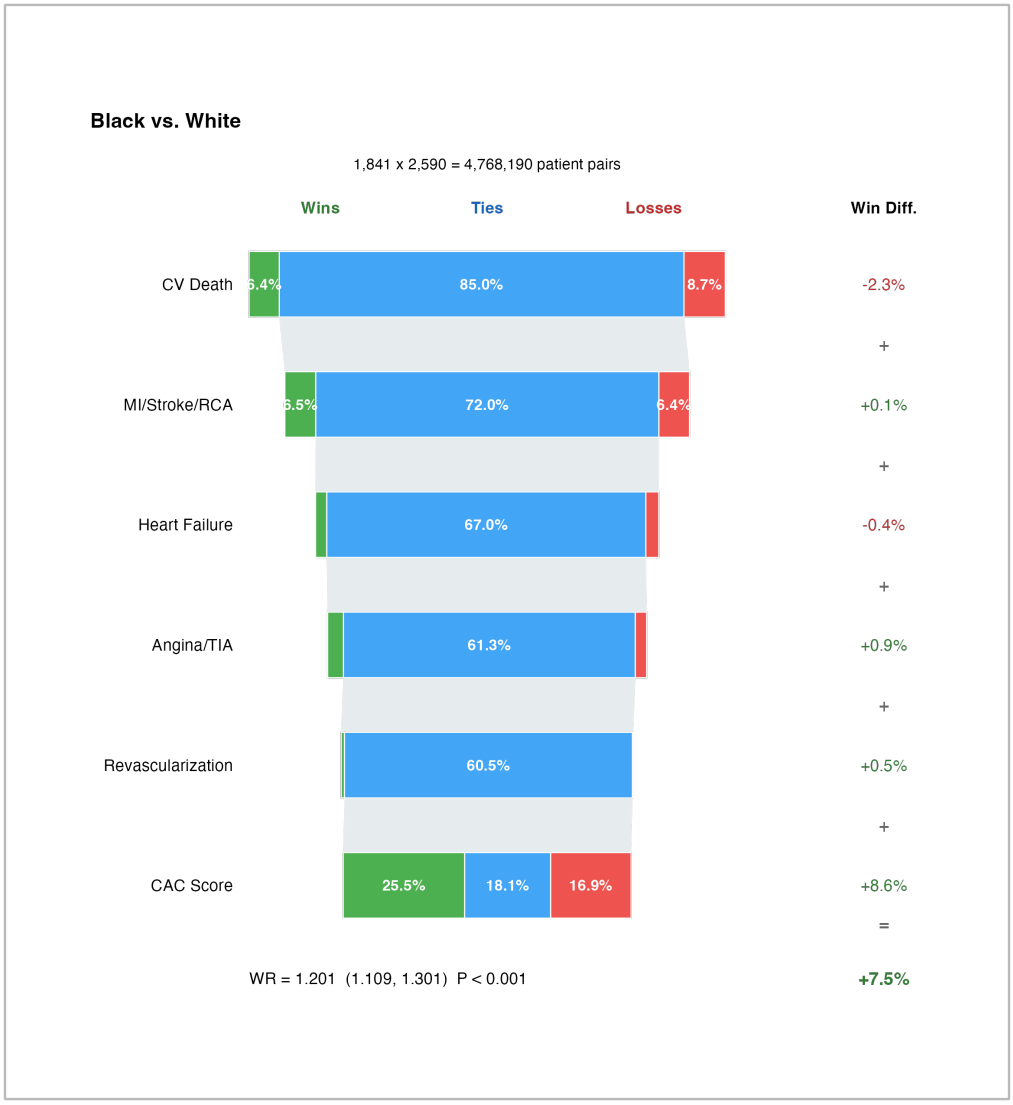

Figure B.1: Hierarchical cascade plot of pairwise comparisons for the 6-component outcome (Black vs. White). See Figure 4.1 for interpretation.

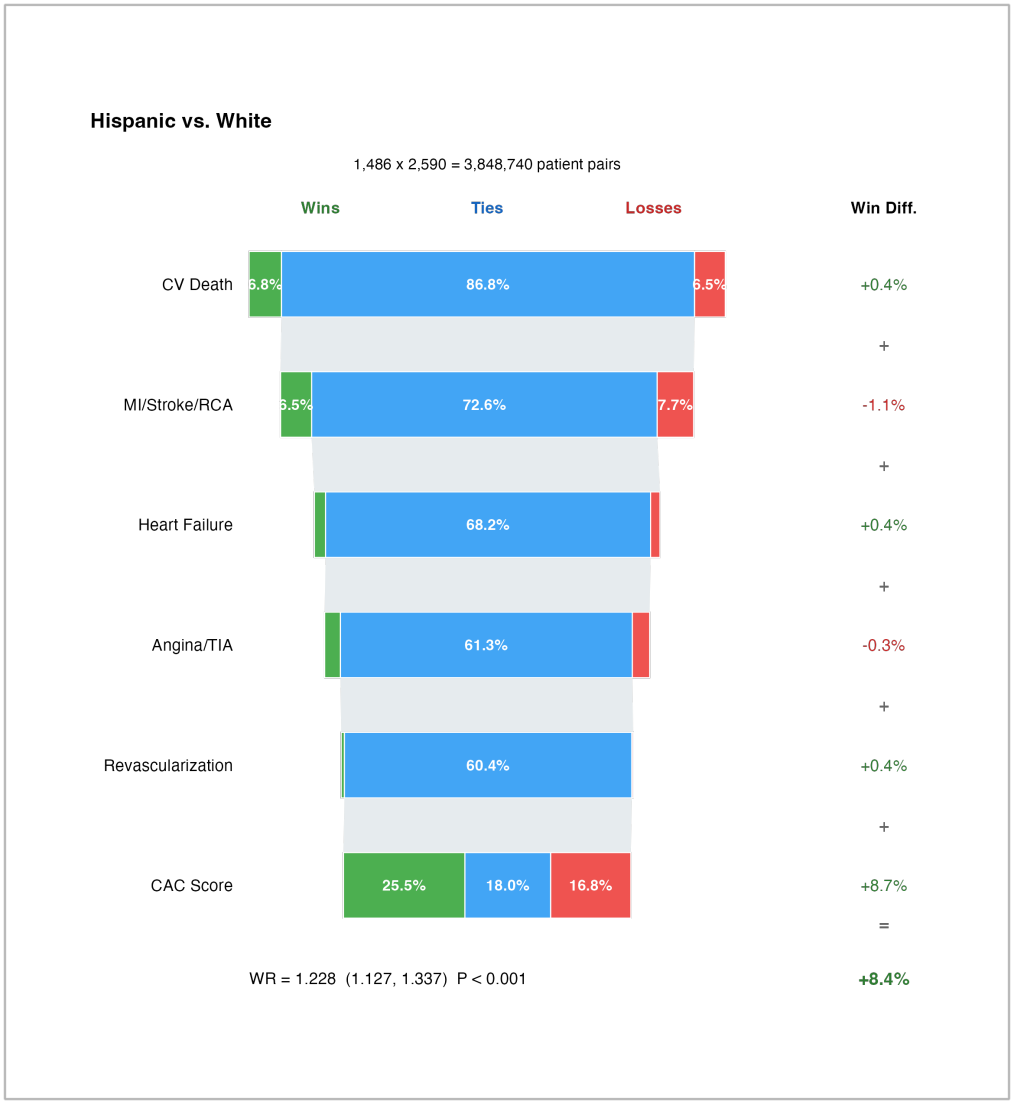

Figure B.2: Hierarchical cascade plot of pairwise comparisons for the 6-component outcome (Hispanic vs. White).

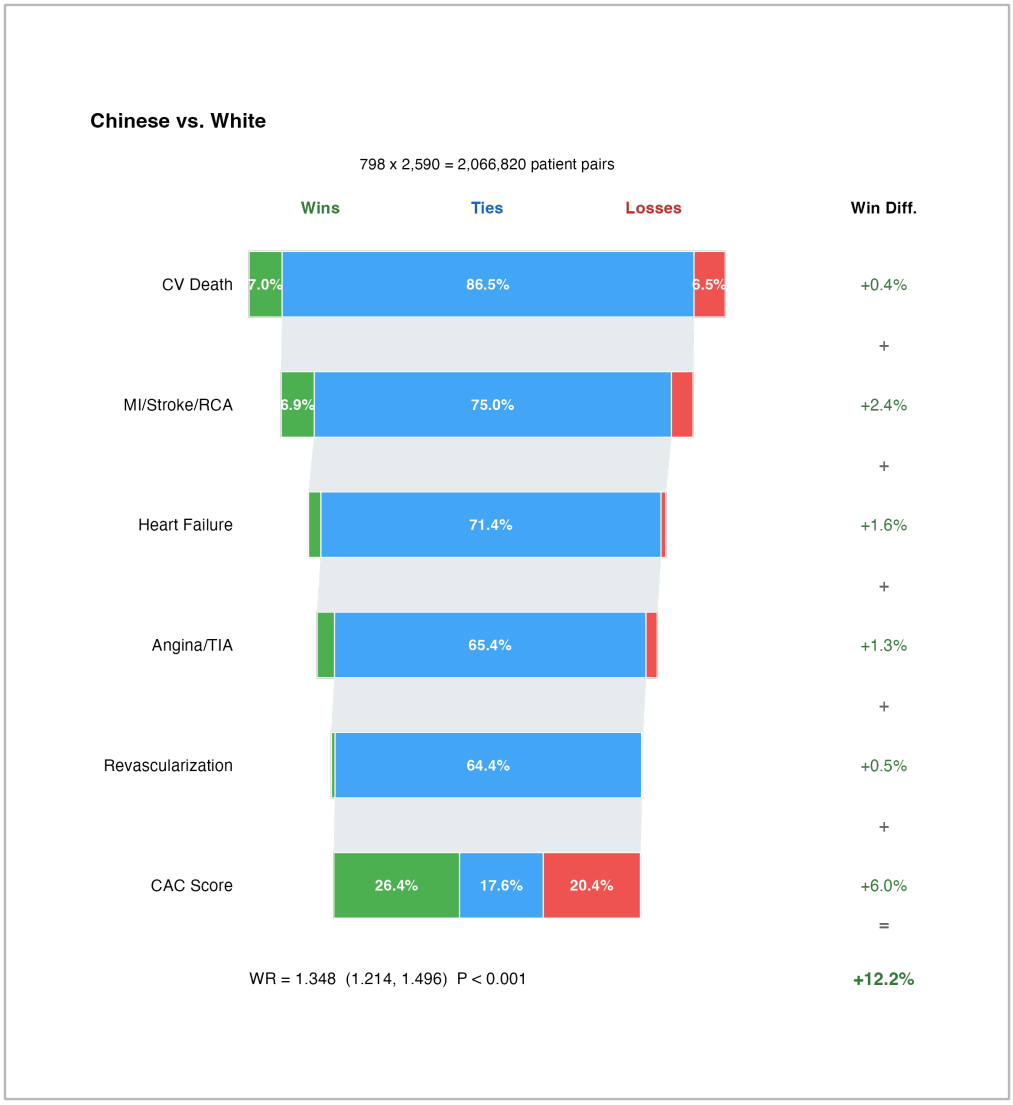

Figure B.3: Hierarchical cascade plot of pairwise comparisons for the 6-component outcome (Chinese vs. White).

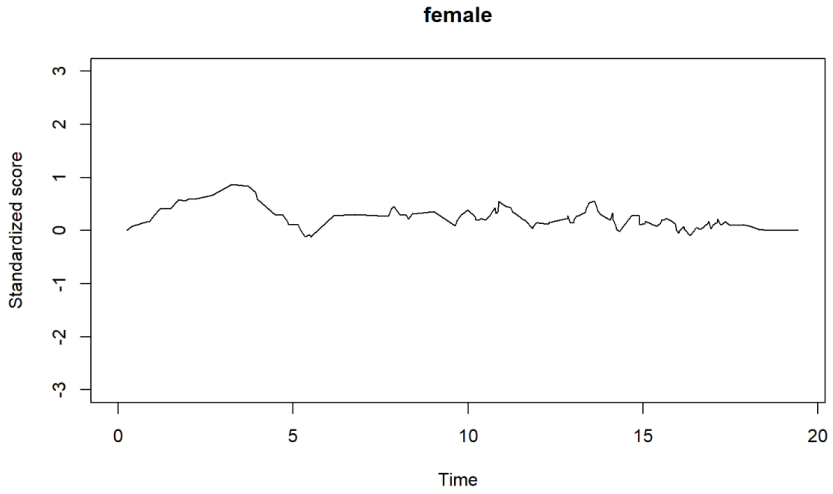

Figure B.4: Standardized score process for the Female indicator in the PW regression model (Female vs. Male).

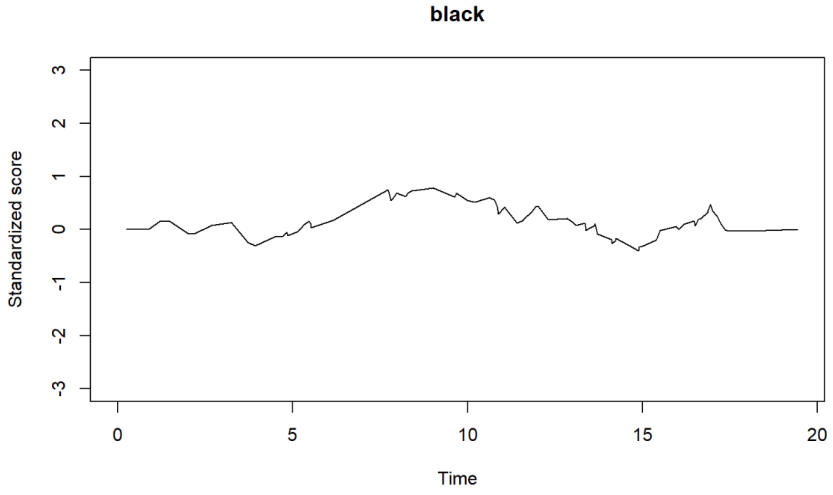

Figure B.5: Standardized score process for the Black indicator in the PW regression model (Black vs. White).

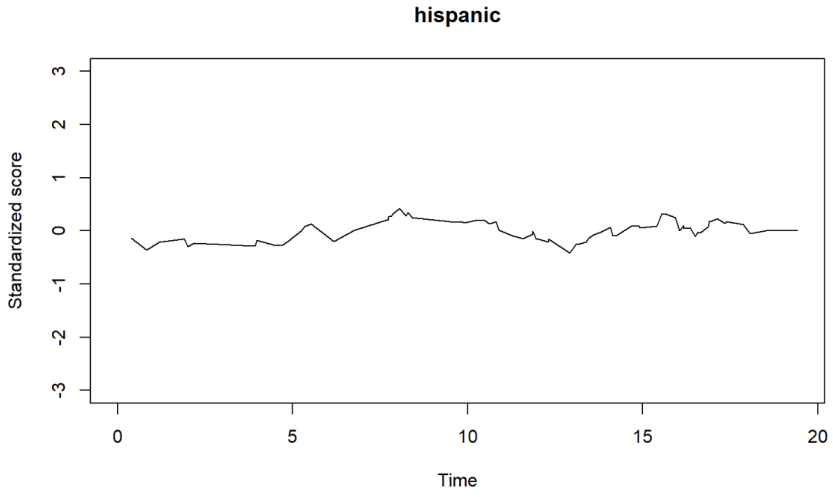

Figure B.6: Standardized score process for the Hispanic indicator in the PW regression model (Hispanic vs. White).

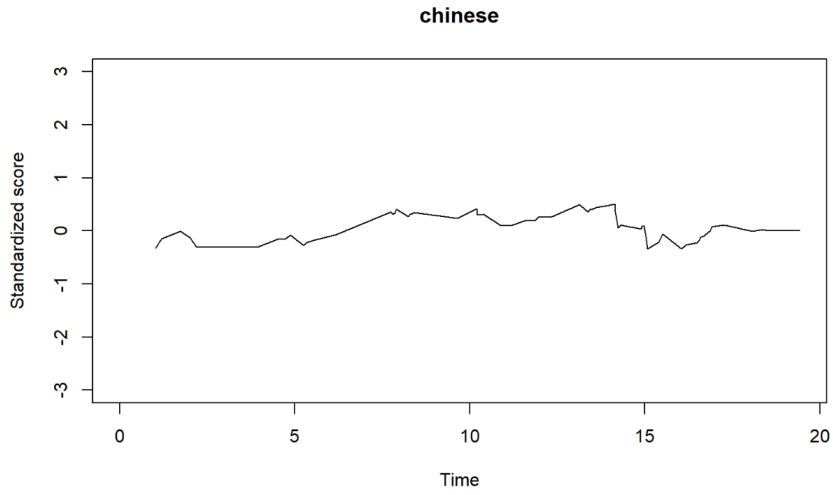

Figure B.7: Standardized score process for the Chinese indicator in the PW regression model (Chinese vs. White).

Appendix C

DERIVING THE TRUE WIN RATIO

Let X and U denote measured and unmeasured covariates with joint density $f_{X,U}(x,u)$. For treatment $a \in \{0,1\}$, define the latent event times $(T_{D,a}, T_{H,a})$ with respective survival functions:

$$S_{D,a}(t \mid x,u) = \exp\{-\Lambda_{D,a}(x,u)\cdot t)\}$$

$$S_{H,a}(t \mid x,u) = \exp\{-\Lambda_{H,a}(x,u)\cdot t\}$$

with respective proportional hazards models:

$$\Lambda_{D,a}(x,u) = \lambda_D \exp[-(\beta_{A,D}a + \beta_{X,D}^\top x + \beta_{U,D}^\top u)]$$

$$\Lambda_{H,a}(x,u) = \lambda_H \exp[-(\beta_{A,H}a + \beta_{X,H}^\top x + \beta_{U,H}^\top u)]$$

The joint survival function for $(T_{D,a}, T_{H,a})$ under treatment a is given by the Gumbel–Hougaard copula:

$$S_{D,H,a}(y_2, y_1 \mid x,u) = \exp -[(\Lambda_{D,a}(x,u)y_2)^\theta + (\Lambda_{H,a}(x,u)y_1)^\theta]^{1/\theta}$$

Define the conditional survival of T_H beyond y_1 given survival of T_D beyond y_2 as

$$G_a(y_1, y_2 \mid x,u) = \frac{S_{D,H,a}(y_2, y_1 \mid x,u)}{S_{D,a}(y_2 \mid x,u)}$$

Define the density of T_D at y_2 under treatment a:

$$\lambda_{D,a}(y_2 \mid x,u) = \frac{\partial}{\partial y_2}[1 - S_{D,a}(y_2 \mid x,u)]$$

Define the hazard for T_H at y_1 conditional on $T_D = y_2$ under treatment a:

$$\lambda_{H|D,a}(y_1 \mid y_2, x,u) = \frac{\frac{\partial}{\partial y_1} G_a(y_1, y_2 \mid x,u)}{G_a(y_1, y_2 \mid x,u)}$$

Define the censoring survival and hazards under treatment a:

$$Q_a(t) = \exp(-\Lambda_{C,a}t), \quad \Lambda_{C,a} = \lambda_C \exp(-\beta_{A,C}a),$$

with:

$$\lambda_{C,a}(t) = \Lambda_{C,a}.$$

Following Zhang and Jeong (2021), we define the four terms:

$$\begin{aligned}
A(x,u) &= \int_0^\infty S_{D,1}(y_2 \mid x,u) Q_1(y_2) S_{D,0}(y_2 \mid x,u) Q_0(y_2) \lambda_{D,0}(y_2 \mid x,u) dy_2 \\
B(x,u) &= \int_0^\infty \int_0^{y_2} G_1(y_1, y_2 \mid x,u) Q_1(y_2) G_0(y_1, y_2 \mid x,u) Q_0(y_2) \lambda_{H|D,0}(y_1 \mid y_2, x, u) \\
&\quad \times \{\lambda_{C,0}(y_2) + \lambda_{C,1}(y_2)\} dy_1 dy_2 \\
C(x,u) &= \int_0^\infty S_{D,1}(y_2 \mid x,u) Q_1(y_2) S_{D,0}(y_2 \mid x,u) Q_0(y_2) \lambda_{D,1}(y_2 \mid x,u) dy_2 \\
D(x,u) &= \int_0^\infty \int_0^{y_2} G_1(y_1, y_2 \mid x,u) Q_1(y_2) G_0(y_1, y_2 \mid x,u) Q_0(y_2) \lambda_{H|D,1}(y_1 \mid y_2, x, u) \\
&\quad \times \{\lambda_{C,0}(y_2) + \lambda_{C,1}(y_2)\} dy_1 dy_2
\end{aligned}$$

Then the true causal win ratio is defined as:

$$\begin{aligned}
\mathrm{WR}_{\mathrm{true}} &= \frac{\mathbb{E}_{X,U}[A(X,U) + B(X,U)]}{\mathbb{E}_{X,U}[C(X,U) + D(X,U)]} \\
&= \frac{\int (A(x,u) + B(x,u)) f_{X,U}(x,u) dx du}{\int (C(x,u) + D(x,u)) f_{X,U}(x,u) dx du}
\end{aligned}$$

When there are a small number of covariates this expression can feasibly be calculated analytically using numerical integration. When the number of covariates is large, the integrals become numerically intractable.

Appendix D

SOFTWARE

All analyses were conducted in R (version 4.4.1). The `WINS` package was used for unadjusted two-sample win statistics, inverse probability weighted win statistics, and computation of win statistics over time [38]. The `WR` package was used for PW regression and related sensitivity analyses. The `survival` package was used for Cox proportional hazards models. The `ipw` package was used to estimate stabilized inverse-probability-of-treatment weights.

The R package `WinStatObservationalSim` was developed for the simulation study, and contains functions for data generation, calculation of the true win ratio, and a simulation engine. The package can be found on GitHub at https://github.com/trurob/WinStatObservationalSim.