Statistical Methods for Composite Analyses of Recurrent and Terminal Events

by

Yiyuan Huang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2026

Doctoral Committee:

Professor Peter X.K. Song, Co-Chair
Professor Min Zhang, Co-Chair
Professor Mousumi Banerjee
Associate Professor Salim Hayek

Yiyuan Huang
huangyy@umich.edu
ORCID iD: 0000-0001-5568-8517

© Yiyuan Huang 2026

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my Ph.D. advisors, Peter X.K. Song and Min Zhang, for their invaluable mentorship and unwavering support throughout my doctoral training. They played a central role in shaping the core ideas of this dissertation and provided extensive guidance on the design of simulation studies, methodological development, and scientific writing. Their detailed comments, careful edits, and high standards pushed me to think more rigorously and communicate more clearly, which has profoundly influenced my growth as a researcher.

I am sincerely grateful to my GSRA advisors, Mousumi Banerjee and Lili Zhao, for their mentorship and support. Through working with them, I learned how to translate statistical methodology into practical applications and how to effectively collaborate with clinicians and other non-technical partners. Their guidance helped me develop not only technical skills but also the communication and collaboration skills that are essential for interdisciplinary research.

I would like to thank my clinical collaborators, Salim Hayek and Rodica Pop-Busui, whose generous support funded my study for nearly five years. Working closely with them gave me invaluable exposure to real-world clinical problems, deepened my understanding of patient care and disease processes, and directly motivated the research questions explored in this dissertation.

I am also thankful to other clinicians and researchers, including Alexi Vasb inder, Anis Ismail, Lynn Ang, and Brian Schmidt, for the engaging discussions and collaborative work within the Hayek and Pop-Busui research groups. Our conversations, ranging from clinical interpretation to data analysis and study design, broadened my perspective and made the research process both intellectually stimulating and enjoyable.

My deepest gratitude goes to my parents, Shengyu Li and Wei Huang, for their unconditional love, support, and sacrifices. They have always placed great value on education and provided me

with every opportunity to pursue my goals. I am equally grateful to my grandparents, Dengxin Li, Xi Deng, Qiming Huang, and Mingxiu Xu, for raising me with care, instilling in me perseverance and curiosity, and laying the foundation for my academic journey. During the most challenging periods of my Ph.D., especially throughout the pandemic, their constant encouragement and reassurance gave me strength and confidence to continue.

Finally, I would like to thank my cohort mates and close friends, Qinmengge Li, Zhongzhe Ouyang, Tian Xie, Chang Wang, Bangyao Zhao, Yuqi Zhai, and Tong (Toni) Yang, for their friendship and companionship throughout this journey. The many evenings we spent cooking together, sharing homemade meals, celebrating holidays, and simply relaxing created a sense of warmth and belonging during my Ph.D. These moments of laughter and connection provided a meaningful balance to academic life and made these years especially memorable.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii

LIST OF FIGURES vi

LIST OF TABLES viii

LIST OF APPENDICES ix

LIST OF ACRONYMS x

ABSTRACT xi

CHAPTER

1 Introduction 1

2 Statistical Methods for Composite Analysis of Recurrent and Terminal Events in Clinical Trials 3

2.1 Introduction 3
2.2 Definitions and Notations 5
2.3 Statistical Methods 6
2.3.1 Composite-outcome Methods 6
2.3.2 Joint-analysis Methods 8
2.3.3 Win-ratio Type of Methods 11
2.4 Simulation Study 13
2.4.1 Single Type of Recurrent Events 14
2.4.2 Two Types of Recurrent Events 23
2.5 Discussion 25

3 Model-Assisted Causal Inference for the Treatment Effect on Recurrent Events in the Presence of Terminal Events 28

3.1 Introduction 28
3.2 While-Alive-based (WA) Test: A Revisit 31
3.3 Separable Effects Framework (SEF) 33
3.4 Proportional Rate Marginal Structural Model-assisted Test (PR-MSMaT) 39
3.5 Simulation Experiment 41

3.6 Application 43
3.7 Discussion 48

4 Causal Inference for Recurrent Events in the Presence of Terminal Events with Covariates and Informative Censoring: An Extension of PR-MSMaT 50

4.1 Introduction 50
4.2 Separable Effect Framework (SEF) 53
4.3 Extended PR-MSMaT (ext-PR-MSMaT) 58
4.4 Simulation Study 59
4.5 Application 65
4.6 Discussion 68

5 Future Directions 71

5.1 Multiple Types of Recurrent Events 71
5.2 Doubly Robust Estimation 72
5.3 Dynamic Treatment Regimes and Individualized Prediction 72

APPENDICES 73

BIBLIOGRAPHY 87

LIST OF FIGURES

FIGURE

2.1 Causal diagram of the relationship between the treatment, recurrent events, and terminal events in the data generation models with a single type of recurrent event. 15

2.2 Simulation results of scenario (i). H_0 indicates that the rejection rate is a type I error; otherwise, it is a power. Acronyms: TE: Cox-PH model with time to terminal events only; TTFE: Cox-PH model with the time-to-first-event outcome; CRE: Combined-recurrent-event analysis using the proportional rate model; JFM: joint frailty model; GL: Ghosh-Lin method; NA: Nelsen-Aalen method; WR: proportional win-fraction regression. 18

2.3 Simulation results of scenario (ii). H_0 indicates that the rejection rate is a type I error; otherwise, it is a power. Acronyms: TE: Cox-PH model with time to terminal events only; TTFE: Cox-PH model with the time-to-first-event outcome; CRE: Combined-recurrent-event analysis using the proportional rate model; JFM: joint frailty model; GL: Ghosh-Lin method; NA: Nelsen-Aalen method; WR: proportional win-fraction regression. 20

2.4 Additional simulations. H_0 indicates that the rejection rate is a type I error; otherwise, it is a power. Acronyms: TE: Cox-PH model with time to terminal events only; TTFE: Cox-PH model with the time-to-first-event outcome; CRE: Combined-recurrent-event analysis using the proportional rate model; JFM: joint frailty model; GL: Ghosh-Lin method; NA: Nelsen-Aalen method; WR: proportional win-fraction regression. 22

2.5 Causal diagram of the relationship between the treatment, recurrent events, and terminal events in the data generation model with two types of recurrent events. 24

2.6 Simulation results: Two types of recurrent events. Acronyms: TE: Cox-PH model with time to terminal events only; TTFE: Cox-PH model with the time-to-first-event outcome; CRE: Combined-recurrent-event analysis using the proportional rate model; JFM: joint frailty model; GL: Ghosh-Lin method; NA: Nelsen-Aalen method; WR: proportional win-fraction regression. 25

3.1 The simulation results illustrate the limitations of the While-alive-based test. WA.0: the survival-completed cumulative loss function for the control group; WA.1: the survival-completed cumulative loss function for the treatment group; RR: Recurrent event Ratio; HR: Hazard Ratio. 34

3.2 Intuition regarding the relationship between the treatment, recurrent events, and terminal events. 35

3.3 Directed acyclic diagram (DAG) illustrates: (a) the causal pathways for treatment A, recurrent event Y_k, and terminal event D_k; (b) the hypothetical decomposition of treatment A into A_Y (directly affecting recurrent events) and A_D (directly affecting terminal events). 37
3.4 Empirical type I error rate and power across time under three scenarios designed in the simulation study: constant, decreasing, and increasing baseline recurrent-event rates. The dashed horizontal line marks the nominal significance level 0.05. 43
3.5 Plots of times for recurrent event of gastrointestinal bleeding (GIB) marked by green dots, death by blue dots appearing at the edge, and censoring by red dots among the matched pairs across device type (LVAD vs. BiVAD). 46
3.6 Comparison of test statistics (Z) and their $-\log_{10}(p)$ values at five selected follow-up times (months) across two test methods: existing WA (While-alive-based test) and new PR-MSMaT (Proportional Rate Marginal Structural Model–assisted Test). The horizontal dashed line in the left panel indicates $Z = 0$, and the red dotted line in the right panel corresponds to the conventional significance threshold $p = 0.05$. 47

4.1 **Data-generating Models.** Panels: (a) recurrent events Y_k, (b) deaths D_k, (c) time-varying covariate $L_{D,k}$, and (d) censoring C_k. Other components: treatment A, hypothetical treatment components A_Y and A_D, baseline confounders L_0. 62
4.2 **Scenario (1)** No covariates, no censoring effects, and $Y \perp D$. Panels show Type I error and power at $n = 500, 1000$. 63
4.3 **Scenario (2)** L_D-related effects $(\alpha_{A_D}, \alpha_Y, \eta_{L_D}, \gamma_{L_D}) = 0.5$, censoring-related effects $(\delta_{A_Y}, \delta_{A_D}, \delta_Y, \delta_{L_0}, \delta_{L_D}) = 0.1$, and $Y \perp D$ ($\eta_Y = 0$). Panels show Type I error and power at $n = 500, 1000$. 64
4.4 **Scenario (3)** L_D-related and censoring-related effects are the same as scenario (2), while $\eta_Y = 0.5$ (recurrent events affect death). Panels show Type I error and power at $n = 500, 1000$. 65
4.5 **Scenario (4)** L_D-related effects are same as scenario (2), censoring-related effects increased to 0.3; $Y \perp D$. Panels show Type I error and power at $n = 500, 1000$. 66
4.6 Comparison of test statistics (Z) and their $-\log_{10}(p)$ values at five selected follow-up times (months) across two test methods: our previous method PR-MSMaT (Proportional Rate Marginal Structural Model–assisted Test) and our proposed new method ext-PR-MSMaT (extended PR-MSMaT). The horizontal dashed line in the left panel indicates $Z = 0$, and the red dotted line in the right panel corresponds to the conventional significance threshold $p = 0.05$. 68

LIST OF TABLES

TABLE

2.1 Type I Error Check ($\beta_1 = \beta_2 = 0$) 16
2.2 Simulation scenario (i) 17
2.3 Simulation scenario (ii) 19
2.4 Time Complexity Analysis (time in seconds) 22

C.1 Baseline characteristics by device type. 79

LIST OF APPENDICES

A Derivation of Estimating Equation in the Proportional Rate Marginal Structural Model 73

B The Asymptotic Normality for the Test Statistic of PR-MSMaT 76

C Summary Table of Baseline Covariates in Dataset after Matching 79

LIST OF ACRONYMS

CRE Combined-recurrent-event Analysis

ESHF End-stage Heart Failure

GIB Gastrointestinal Bleeding

GL Ghosh–Lin Method

H0 Null Hypothesis

HR Hazard Ratio

INTERMACS Interagency Registry for Mechanically Assisted Circulatory Support

JFM Joint Frailty Model

MCS Mechanical Circulatory Support

NA Nelsen–Aalen Method

PH Proportional Hazards

PR-MSMaT Proportional Rate Marginal Structural Model Assisted Test

RCT Randomized Controlled Trial

RMST Restricted Mean Survival Time

RR Recurrent Event Ratio

TE Terminal-event Analysis

TTFE Time-to-first-event Analysis

WA While-alive Method

WR Win-fractions Regression

ABSTRACT

In clinical studies, researchers are often interested in evaluating treatment effects on composite endpoints involving both recurrent and terminal events. For example, in the Interagency Registry for Mechanically Assisted Circulatory Support (INTERMACS) program, patients with end-stage heart failure (ESHF) treated by mechanical circulatory support (MCS) devices may experience recurrent complications such as gastrointestinal bleeding (GIB), while death is deemed as a terminal event that precludes further occurrences. Analyzing these events together can reduce estimation bias and increase statistical power by integrating more information of adverse outcomes during the treatment period. Chapter 2 is devoted to review and compare existing statistical methods for such composite analyses through comprehensive simulation studies, which demonstrates the need of evaluating treatment effects on recurrent events in the presence of terminal events. This motivates the new methodological developments of rigorous, robust hypothesis testing procedures on treatment effects in Chapters 3 and 4. Chapter 3 introduces a model-assisted testing procedure for recurrent events in the presence of terminal events, while Chapter 4 extends this framework to account for baseline and time-varying covariates, as well as informative censoring.

Chapter 2 is devoted to review the mainstream composite analysis methods in three categories: (A) composite-outcome methods, which combine multiple events into a composite outcome before analysis, e.g., combining events into a time-to-event outcome in time-to-first-event analysis (TTFE) and into a single recurrent event process in the combined-recurrent-event analysis (CRE); (B) joint-analysis methods, which test for the recurrent event process and the terminal event jointly, e.g., Joint Frailty Model (JFM), Ghosh-Lin Method (GL), and Nelson-Aalen Method (NA); and (C) win-ratio type methods that account for the ordering of two types of events, e.g., Win-fraction Regression (WR). Through comprehensive simulation studies to evaluate type I error control and

power under a wide range of scenarios, the non-parametric joint testing approach (GL/NA) and CRE demonstrate overall the best performance, whereas TTFE and WR exhibit relatively low power. Also, incorporating events in the composite analyses that have no or weak association with treatments usually decreases power.

Chapter 3 is motivated by comparing the implant MCS devices [left ventricular assist devices (LVADs) versus biventricular assist devices (BiVADs)] as an essential treatment option for patients with ESHF. A task of central interest is to conduct a hypothesis test for the influence of such treatment option on the risk of recurrent events, which is greatly challenged by death as a terminal event. Existing methods based on cumulative event frequency may disadvantage individuals with longer survival, as they tend to experience more recurrent events simply because they remain at risk for a longer period. The While-Alive–based (WA) test addresses this issue of survival-length bias and appears to perform well under a temporally constant recurrent event rate. Such constant-rate assumption is often violated in practice, leading to inflated type I error and inaccurate estimates of treatment effects for the WA test. To address this methodological gap, a Proportional Rate Marginal Structural Model-assisted Test (PR-MSMaT) is proposed within the causal inference framework of separable treatment effects for recurrent and terminal events. Through simulation studies, the proposed PR-MSMaT exhibits a proper control of type I error while achieving power comparable to the WA test under time-varying recurrent event rates. Applying the propensity score matching method to create match-pairs, I perform PR-MSMaT to compare the postoperative risk of GIB across different MCS devices among pair-matched ESHF patients enrolled in the INTERMACS program and conclude that the choice of BiVAD has a lower risk of GIB than the choice of LVAD, particularly after 5 months of the implant surgery.

Chapter 4 focuses on an important extension of PR-MSMaT to accommodate baseline and time-varying covariates, as well as informative censoring, so that the resulting new test can be applied to test for treatment effects in observational study settings such as INTERMACS. The resulting test provides a desirable methodological flexibility by incorporating appropriate weighting to adjust for sampling heterogeneities within the separable effects framework. Simulation results show

that the extended PR-MSMaT improves power over the original version of PR-MSMaT procedure introduced in Chapter 3 while preserving a valid type I error control under scenarios with time-varying covariates and informative censoring. I re-analyze the postoperative risk of GIB in the INTERMACS program using the extended PR-MSMaT, adjusting for the time-varying all-cause infection process and informative censoring. I confirm the finding in Chapter 3 that the choice of BiVAD has a lower risk of GIB than the choice of LVAD, and such benefit begins right after the implant surgery and lasts over the postoperative follow-up.

Chapter 5 outlines potential future directions, including accommodating multiple types of recurrent events, developing doubly robust estimating equations, and extending the proposed testing framework to dynamic treatment regimes and individualized risk prediction for recurrent events.

CHAPTER 1

Introduction

In clinical studies, evaluating treatment effects often involves outcomes that are more complex than a single time-to-event endpoint. In many settings, patients may experience recurrent events over time, such as repeated hospitalizations or complications, while also being subject to a terminal event, such as death, that precludes further occurrences. For example, in the Interagency Registry for Mechanically Assisted Circulatory Support (INTERMACS), patients with end-stage heart failure treated with mechanical circulatory support devices frequently experience recurrent complications such as gastrointestinal bleeding, while death terminates the observation of these events. This dual-outcome structure presents unique challenges for statistical analysis.

Analyzing recurrent and terminal events jointly is essential for making efficient and unbiased use of available data. Traditional approaches that consider only a single outcome, such as time to first event or survival alone, may fail to fully capture the burden of disease and treatment effects. By incorporating both recurrent and terminal events into a composite analysis framework, researchers can improve statistical efficiency and reduce bias by leveraging more comprehensive information about patient outcomes over time.

A wide range of statistical methods have been developed to address this problem. These include composite-outcome methods that combine multiple events into a single endpoint, joint-analysis methods that model recurrent and terminal events simultaneously, and win-ratio–type approaches that prioritize event ordering. While these methods provide useful tools for composite analysis, they also raise important methodological questions, particularly regarding how to formally evaluate

treatment effects on recurrent events in the presence of terminal events. In particular, the dependence between recurrent and terminal events and the truncation of recurrent processes by death complicate both estimation and hypothesis testing.

Motivated by these challenges, this dissertation focuses on developing rigorous statistical methods for inference on recurrent events in the presence of terminal events. The primary objective is to construct testing procedures that are both statistically valid and practically interpretable, while appropriately accounting for the complex interplay between event processes.

This dissertation is organized as follows. Chapter 2 provides a comprehensive review and comparison of existing statistical methods for composite analysis of recurrent and terminal events through extensive simulation studies. The results highlight the strengths and limitations of current approaches and demonstrate the need for improved methods for hypothesis testing in this setting.

Building on these findings, Chapter 3 introduces a novel model-assisted testing procedure for evaluating treatment effects on recurrent events in the presence of terminal events. This approach is designed to address key limitations of existing methods and to provide robust inference under a broad range of scenarios.

Chapter 4 further extends the proposed framework by incorporating baseline and time-varying covariates, as well as informative censoring. This extension enables the method to be applied in more realistic settings, including observational studies where confounding and censoring mechanisms must be properly addressed.

Finally, Chapter 5 discusses potential directions for future research, including extensions to multiple types of recurrent events, development of doubly robust estimation procedures, and applications to dynamic treatment regimes and individualized risk prediction.

CHAPTER 2

Statistical Methods for Composite Analysis of Recurrent and Terminal Events in Clinical Trials

2.1 Introduction

In clinical studies, often various types of outcomes (events/endpoints) are measured, including recurrent and terminal events. Examples include the major adverse cardiac events and deaths due to cardiovascular diseases (Kip et al. 2008, Nicholls et al. 2020), nonfatal events and in-hospital deaths in COVID-19-related researches (Vasbinder et al. 2022, Pan et al. 2021), deaths, recurrences, morbidity, and side-effects after surgery in cancer researches (Rullier et al. 2017). When testing treatment effects, which is often the primary analysis in clinical studies, it may not be immediately apparent what outcome(s) should be used. Therefore, exploring statistical methods that can effectively handle multiple outcomes is of great interest.

When there are multiple outcomes, using a composite outcome as a primary endpoint is a frequently adopted methodology. This strategy combines multiple outcomes or different types of events into a singular measure (Freemantle et al. 2003), for example, the most widely-used time-to-first-event-outcome (TTFE). Composite outcome approaches are considered to have several inherent advantages for testing treatment effects. Among these, the most obvious ones are its conceptual and implementation simplicity. Primarily, compared with testing each endpoint separately, combining multiple events into a single one in analysis can avoid the need of adjustment for multiple testing (Heddle & Cook 2010), hence enhancing the statistical power. Such enhancement facilitates the

effective detection of treatment effects even with constrained sample sizes. Also, by combining diverse event data into a single metric, composite outcomes streamline the reporting process of the statistical results, precluding the need to delineate the treatment effects on each event in isolation. Nonetheless, on the other hand, it may also lead to a major drawback of the approach, that is, the limited specificity in capturing the treatment effect on each component of the composite outcome. The issue is especially of concern when the treatment has contrary effects on different components, making it difficult to determine the specific mechanisms of how the treatment impacts outcomes (Heddle & Cook 2010).

Time-to-first-event analysis (TTFE) is the most popular composite outcome approach, and it has been widely adopted in both practice and research (Zannad et al. 2013, Anker et al. 2016, Rauch, Kieser, Binder, Bayes-Genis & Jahn-Eimermacher 2018, Pan et al. 2021, Vasbinder et al. 2022). TTFE has emerged as a popular choice because it provides a single, easily interpretable summary outcome that can succinctly capture the essential information on the timing of all events, i.e., the minimum order statistic of times to all events of interest. However, as TTFE ignores subsequent events other than the first one, it leads to loss of information and possibly inefficient statistical inference, especially when the rate of recurrent events is high. Naturally, one would like to exploit further information on subsequent events and the relationship between the recurrent and terminal events; thus, other methods for the composite analysis of both recurrent and terminal events have been developed (Lin et al. 2000, Ghosh & Lin 2000, Huang & Wang 2004, Mao & Wang 2021). Compared to TTFE, those methods may reduce information loss and potentially achieve higher statistical power when testing the treatment effect in clinical studies. However, they have been less adopted in practice than TTFE. A few simulation studies have compared TTFE and other composite-analysis methods (Mao & Kim 2021, Fritsch et al. 2023). These studies have provided valuable information for understanding the practical performance of different methods. However, while most of these simulation studies included the TTTE method and some other methods, to the best of our knowledge, there is a lack of a comprehensive review and comparison of different types of methods under a wide range of scenarios.

The chapter aims to comprehensively review different types of composite analysis methods when both recurrent and terminal events exist. Performances of various composite outcome approaches for hypothesis testing are compared through simulation studies under a wide range of scenarios regarding both type I error control and power. The remainder of this chapter is organized as follows: section 2.2 introduces the notations and defines the problem. Section 2.3 reviews the three classes of statistical methods and representative methods in each class. Section 2.4 details the simulation setups and results, followed by a discussion in section 2.5.

2.2 Definitions and Notations

Suppose there are n subjects in a clinical trial, indexed by $i = 1, 2, \cdots, n$. For the i-th subject, let A_i denote the assigned treatment group, which can either be 0 (the control group) or 1 (the treatment group), D_i denote the time to the terminal event, and the sequence $\{T_{ik}\}_{k=1}^{r_i}$ denote the times of the recurrent events, satisfying $0 < T_{i1} < T_{i2} < \cdots < T_{i,r_i} < D_i$, where r_i is the number of recurrent events the subject experiences before D_i.

We denote the counting processes of the recurrent and terminal events, respectively, for subject i as

$$N_i^{T,*}(t) = \sum_{k=1}^{r_i} I(T_{ik} \leq t), \tag{2.1}$$

$$N_i^D(t) = I(D_i \leq t), \tag{2.2}$$

where $I(\cdot)$ is the indicator function. Note that $N_i^{T,*}(t)$ counts the number of recurrent events by time t. For subject i, let C_i denote the right-censoring time that is assumed to be independent of all the recurrent and terminal events of the subject. Then, the observation time corresponding to the terminal event would be $X_i = \min(D_i, C_i) := D_i \wedge C_i$, and the censoring indicator $\delta_i = I(D_i \leq C_i)$, with $\delta_i = 1$ denoting not being censored. Accordingly, the observed counting process of recurrent events can be denoted as $N_i^T(t) = N_i^{T,*}(t \wedge X_i)$. The observed data can be summarized by

$\left\{A_i, X_i, \delta_i, N_i^T(t)\right\}$ (independently and identically distributed across $i = 1, 2, \cdots, n$).

Since different methods for composite analysis may target different estimands and hence may not be directly comparable in terms of estimation. In this article, we focus on hypothesis testing of treatment effects. Specifically, we consider the following hypotheses:

H_0 :There are no differences in recurrent or terminal events between the control and treatment group,

and

H_1 : Other than H_0,

where H_0 and H_1 are the null and alternative hypotheses, respectively.

2.3 Statistical Methods

In this section, we review the statistical methods that handle multiple types of events in clinical studies, including both recurrent and terminal events. We refer to these methods generally as composite analysis approaches, while the more common term "composite outcome method" is used to refer to a specific class within the general composite analysis methods. We categorize these methods into three classes and select six of the most representative methods for a detailed introduction in this section and for comparison in the simulation study.

2.3.1 Composite-outcome Methods

The composite-outcome methods combine all information (i.e., type of events and timing) into a single process/a composite outcome (i.e., time-to-event outcome or recurrent events outcome), and then standard analysis for the corresponding type of outcome is applied. As noted in Section 2.1, using composite outcomes is the most widely employed strategy when dealing with multiple endpoints. In this class, the most popular one is by treating all events the same without distinguishing

the type. For instance, when the component outcomes are binary, such as whether COVID patients have hospitalized outcomes (e.g., need for mechanical ventilation, need for dialysis, or in-hospital death), the commonly used composite outcome is "whether a subject has experienced any of them" (Vasbinder et al. 2022). When there is censoring, and the interest also lies in the timing of the composite event, then one can combine all information on events into time to the first of all events, leading to the time-to-first-event analysis. Alternatively, one may want to keep information on the timing of each component event, while still not distinguishing the type of events. Then it combines all information into a recurrent event outcome. Methods originally developed for recurrent event outcomes can be applied, which we refer to as combined-recurrent-event analysis (CRE).

2.3.1.1 Time-to-first-event Analysis (TTFE)

When time to multiple types of events or endpoints are simultaneously considered, the time-to-first-event (TTFE) analysis is the most commonly used method in practice (Rauch & Beyersmann 2013, Mao & Kim 2021). This method focuses on the time to the first event, whichever type, but may neglect subsequent events that a subject may experience. Standard survival analysis techniques, such as the log-rank test or Cox proportional hazard (Cox-PH) model, are typically used for analysis. For the subject i, the counting process considered in TTFE analysis is $N_{TTFE,i}(t) = I(D_i \wedge T_{i1} < t)$. The TTFE analysis treats each type of event equally and only considers the timing of the first event, potentially resulting in information loss and reduced statistical power, particularly when the number of recurrent events is high (Rauch & Beyersmann 2013, Rauch, Kieser, Binder, Bayes-Genis & Jahn-Eimermacher 2018).

2.3.1.2 Combined-recurrent-event Analysis (CRE)

In contrast to TTFE, which may lose information on subsequent events, the combined-recurrent-event analysis incorporates all events, including both recurrent and terminal events, as a single counting process, and then the comparison between treatment groups is based on the combined counting process. The Andersen-Gill model (Andersen & Gill 1982) and the proportional rate

model (Lin et al. 2000) are the most common models for modeling recurrent events. When only the treatment indicator is included as an independent variable, and the robust sandwich estimator is used, the aforementioned two models are equivalent in terms of estimation and inference (Furberg et al. 2022). Specifically, for subject i, the composite counting process of the two types of events (recurrent and terminal events) is denoted as $N_{CRE,i}(t) = N_i^T(t) + N_i^D(t)$. Then, the proportional rate model can be written as,

$$\lambda_{CRE,i}(t|A_i) = \lambda_{CRE,0}(t) \exp(\beta_{CRE} A_i),$$

where $\lambda_{CRE,0}(t) = E\{\mathrm{d}N_{CRE}(t)\}$ is the non-negative baseline event rate of the composite process and β_{CRE} is the treatment effect on the rate of the composite process. Then, the Wald test on the hypothesis $\beta_{CRE} = 0$ is conducted. To account for within-subject correlation, the robust standard error obtained by the sandwich estimator is used for the test.

To summarize, the proportional rate model incorporates information on both recurrent and terminal events but treats them as equivalent, resulting in the loss of information on the relationship between them. One may argue that the proportional rate or intensity model assumes proportionality of either rate or intensity, which may not hold in practice. However, we note that proportionality always holds under the null hypothesis, and therefore, the test based on the model will have a type I error rate at the nominal level, leading to a valid test. When the alternative hypothesis is true, and the proportionality assumption is violated, the test will still remain valid despite not having the optimal power.

2.3.2 Joint-analysis Methods

With the development of statistical methodologies, it has become clear that terminal events play a distinct role in the composite analysis of multiple endpoints. First, the terminal event (such as death) differs from the other types of recurrent events regarding clinical importance and interpretation. Also, as it can only occur once and preclude the occurrence of other events, the terminal event is a

different type of event from other recurrent events. It would be desirable to separate the terminal event from other recurrent events and consider the two processes of events jointly. There has been a growing interest in developing statistical methods that jointly model both the recurrent and terminal events (Ghosh & Lin 2000, 2002, Huang & Wang 2004, Zeng & Cai 2010, Mao 2023). We generally refer to these methods as joint-analysis methods.

2.3.2.1 Joint Frailty Model (JFM)

One popular method for jointly modeling recurrent and terminal events is the joint frailty model (JFM) studied by Huang & Wang (2004). The joint frailty model models recurrent and terminal events using the semiparametric proportional rate model and the proportional hazard model, respectively. The two-component models share a subject-specific frailty term, which captures the correlation between recurrent and terminal events. Specifically, the joint frailty model is

$$\lambda_D(t|A) = z\lambda_{D,0}(t)\exp(\beta_D A), \tag{2.3}$$

$$\lambda_T(t|A) = z\lambda_{T,0}(t)\exp(\beta_T A), \tag{2.4}$$

where $\lambda_{D,0}(t)$, $\lambda_{T,0}(t)$ are the baseline hazard and rate function of terminal and recurrent events, respectively; β_D and β_T quantify treatment effects on the terminal and recurrent events, and z is a shared frailty term. Under the joint frailty model, testing of the treatment effect can be based on testing the following hypotheses:

$$H_0 : \beta_D = \beta_T = 0 \quad \text{vs.} \quad H_1 : \beta_D \neq 0 \text{ or } \beta_T \neq 0.$$

The estimator of coefficients of interests $\mathbf{T} = (\hat{\beta}_D, \hat{\beta}_T)'$ as well as the corresponding covariance matrix $\hat{\Sigma}$ can be obtained by maximizing the penalized likelihood (Rondeau et al. 2012), then the Wald test statistic is $\mathbf{Q} = \mathbf{T}'\hat{\Sigma}^{-1}\mathbf{T}$, which follows a Chi-square distribution with 2 degrees of

freedom asymptotically under the null hypothesis (H_0). Although the JFM can utilize all available data on recurrent and terminal events, it can be challenging to verify its parametric assumptions, such as proportional rate in (2.4) and proportional hazard in (2.3). This has led some to question the applicability of the model. Additionally, the treatment effect cannot be interpreted as the rate (hazard) ratio between groups due to the subject-specific frailty and the non-linear link function in the model (Furberg et al. 2022), which restricts the interpretation of the estimated treatment effect at the subject level. However, as we commented previously in the discussion of the proportional rate model in section 2.3.1.2, we focus on hypothesis testing but not on estimating the treatment effect in this paper.

We comment that the joint frailty model is chosen to represent a class of modeling approaches where part of the joint processes is modeled parametrically. For example, copula-based models are alternatives to shared frailty for modeling the dependence structure. Also, the frailty term could be modeled using more flexible distributions. We refer readers to Emura et al. (2019) for more details.

2.3.2.2 Non-parametric Joint Testing

Ghosh & Lin (2000) proposed a nonparametric joint testing approach for recurrent and terminal events. Unlike the joint frailty model, which quantifies the treatment effect by parameters in a semiparametric model, the joint testing approach of Ghosh & Lin (2000) (GL) directly conducts hypothesis testing on two processes: the counting process of terminal events and the cumulative count of recurrent events, acknowledging that there would be no further recurrent events after the terminal event. The estimator for the mean cumulative count of recurrent events is developed to respect this restriction and estimates the mean cumulative count of events in the real world when the terminal event exists. The cumulative hazard of the terminal event is estimated by the Nelson-Aalen estimator. The two estimators for the two processes are then integrated into a joint test statistic. Denote the mean frequency function of recurrent events in the control and treatment group as $\mu_0(t)$ and $\mu_1(t)$, respectively; the cumulative hazard function of the terminal event as $\Lambda_0^D(t)$ and $\Lambda_1^D(t)$

for control and treatment groups, respectively. Then the hypothesis to be tested is

$$H_0 : \mu_0(t) = \mu_1(t) \text{ and } \Lambda_0^D(t) = \Lambda_1^D(t) \text{ for all } t;$$
$$H_1 : \mu_0(t) \neq \mu_1(t) \text{ or } \Lambda_0^D(t) \neq \Lambda_1^D(t) \text{ for some } t.$$

The corresponding test statistic Q_T is asymptotically chi-squared with 2 degrees of freedom under H_0 (Ghosh & Lin 2000).

Alternatively, a closely related approach is to estimate the mean cumulative count of recurrent events using the Nelson-Aalen estimator, treating the terminal event as censoring instead of a competing risk in the GL approach. The approach will also lead to a joint 2-degree-of-freedom test, referred to as the Nelsen-Aalen method (NA).

Although both methods jointly test the recurrent and terminal events, the mean cumulative count of events targeted by the two methods differ and have different interpretations. For example, the GL method estimates the mean cumulative count of events in a world where terminal events may occur. Therefore, if a treatment has no effect on the recurrent events but increases the survival time, then it will also lead to an increase in the mean cumulative count of the recurrent events. As a result, for the GL method, the interpretation of the 1-degree-of-freedom test corresponding to the recurrent event process may be problematic. Nonetheless, the null hypothesis targeted by the 2-degree-of-freedom test of GL method always holds, and hence, we expect it to be a valid test with correct type I error control.

2.3.3 Win-ratio Type of Methods

Methods reviewed in the previous sections, regardless of whether they distinguish recurrent and terminal events or not, do not explicitly consider the clinical importance of different types of events. Neither do they prioritize or weight different types of events based on clinical relevance or other considerations. For example, the CRE method treats all event the same, regardless of whether it is a recurrent or terminal event. Terminal events such as deaths may have greater clinical significance

than other recurrent events. The it's reasonable to incorporate the importance or clinical relevance of different types of events in analysis. The third class of methods we reviewed here has taken this into account. The weighting methods are commonly used when information on the importance or clinical relevance of different types of events is available. The weighting methods can work on different levels (Ozga & Rauch 2022). For instance, Lachin & Bebu (2015) proposed a method based on the weighted event-type-specific hazard ratio, while Rauch, Kunzmann, Kieser, Wegscheider, König & Eulenburg (2018) proposed a method based on the "all-cause hazard ratio", and Bakal et al. (2015) proposed a method based on the weighted Kaplan-Meier curve. Furthermore, Mao & Lin (2016) proposed a weighting framework to combine multiple types of recurrent events with terminal events. However, all of these methods require pre-specified weights based on previous research or expertise from clinical workers, which is subjective and may be difficult to justify. Alternatively, the win-ratio type method addresses the importance of events by comparing the events in a pre-specified order that is easier to justify (Mao & Wang 2021, Mao et al. 2022).

The win ratio, a composite outcome measure initially proposed by Pocock et al. (2012), is based on a pairwise comparison between treatment groups for multiple types of events, considering the clinical relevance ordering of the events (denoted by $E_1, E_2, ...$) with decreasing significance. For each pair of subjects, the time to E_1 is compared first, and the subject with the longer time is declared the winner. If the comparison cannot be made due to missing or censored data, the time to E_2 is compared. This process continues until a winner is determined or the time to the last type of event is compared. In an unmatched study (as used in this paper), every subject in the control group is compared with every subject in the treatment group. The win ratio is defined as the fraction of wins divided by the fraction of losses when comparing the treatment to the control. In this study, if we prioritize terminal events over recurrent events and only focus on the first recurrent event, the indicator of subject i winning over subject j can be written as

$$\mathcal{W}_{ij}(t) = I(D_j < D_i \wedge t) + I(t < D_i \wedge D_j, T_{1j} < T_{1i} \wedge t), \tag{2.5}$$

where i and j are indexes of subjects from the treatment group and control group, respectively. The first term in (2.5) pertains to the comparison of the terminal event, while the second term in (2.5) pertains to the comparison of the first recurrent event when the winner cannot be determined by the terminal events. Then the win-loss function can be defined as

$$\omega_1(t|A_i = 1, A_j = 0) = E\left\{\mathcal{W}_{ij}(t)|A_i = 1, A_j = 0\right\},$$
$$\omega_0(t|A_i = 1, A_j = 0) = E\left\{\mathcal{W}_{ji}(t)|A_i = 1, A_j = 0\right\},$$

which is the expected number of wins and losses of all pairs of the subjects with $A_i = 1$ (treatment) compared to $A_j = 0$ (control). The ratio of ω_1 to ω_0, i.e., ω_1/ω_0, is the win-ratio proposed by Pocock et al. (2012). Mao & Wang (2021) generalized the idea from the two-sample setting to regression and proposed the proportional win-fraction regression model,

$$\frac{\omega_1(t|Z_i, Z_j)}{\omega_0(t|Z_i, Z_j)} = \exp\left\{\beta_{WR}(Z_i - Z_j)\right\},$$

where Z_i and Z_j are the vectors of covariates from the subjects i and j that come from two treatment groups respectively. When the only covariate under consideration is the treatment indicator, then β_{WR} is the log of the win-ratio studied by Pocock et al. (2012). Other extensions of the win-ratio type of regressions are proposed by Mao et al. (2022), which uses more information on recurrent events.

2.4 Simulation Study

Simulation studies are carried out to evaluate and compare the various methods reviewed above. We focus on scenarios that include a single type of recurrent event in addition to deaths, while scenarios with two types of recurrent events are also investigated.

2.4.1 Single Type of Recurrent Events

2.4.1.1 Simulation Setup

In a simulation dataset, a total of 200 subjects were generated, and the treatment variable $A \sim Bernoulli(0.5)$. The data for each subject consisted of recurrent and terminal events. The recurrent events were generated from a model with an intensity function that was influenced by the subject's previous history of recurrent events, namely

$$\lambda(t) = \lambda_0 \exp\left\{\beta_1 A + \beta_4 N^T(t)\right\},$$

where $N^T(t)$ is the cumulative number of recurrent events by the time t as defined in (2.1), λ_0 is the constant baseline intensity.

The terminal events were generated according to a model with a hazard function that also depended on the subject's previous history of recurrent events. The terminal events were generated from the process with the following hazard function

$$\gamma(t) = \gamma_0 \exp\left\{\beta_2 A + \beta_3 N^T(t)\right\},$$

where γ_0 is the constant baseline hazard function. Figure 2.1 illustrates the causal relationship between treatment, recurrent events, and terminal events, as the data generating models specify.

Each subject was followed up to 5 years, and any events beyond that time were considered administratively censored. Additionally, an independent censoring time C was generated for each subject through an exponential random variable with a pre-specified rate parameter, resulting in a sample-wide censorship rate of 20% (±0.1%).

We consider two scenarios based on the relative intensity of recurrent and terminal events. Scenario (i) is characterized by a high baseline hazard of terminal events ($\gamma_0 = 1$) and low baseline intensity of recurrent events ($\lambda_0 = 0.3$), resulting in earlier termination and a low frequency of recurrent events. (Table 2.2) In contrast, scenario (ii) involves terminal events with low baseline

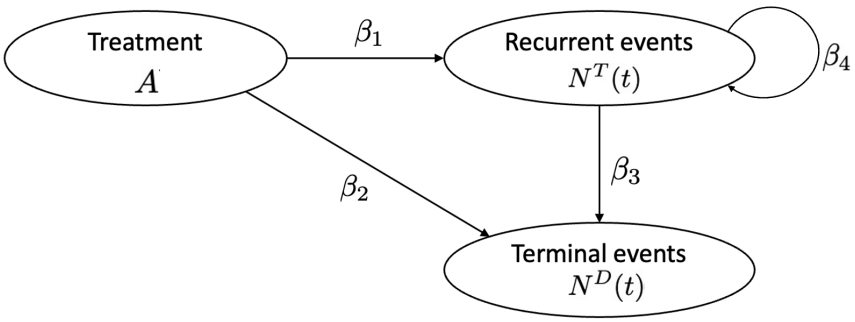


Figure 2.1: Causal diagram of the relationship between the treatment, recurrent events, and terminal events in the data generation models with a single type of recurrent event.

hazard ($\gamma_0 = 0.3$) and recurrent events with high baseline intensity ($\lambda_0 = 1$). (Table 2.3)

Furthermore, multiple sub-scenarios were generated depending on whether the treatment directly affects recurrent events (β_1) or not, the treatment has a direct effect on terminal events (β_2) or not, as well as past recurrent events affect the future recurrence rate (β_4) or not. In all scenarios, recurrent events always have a direct effect on terminal events with $\beta_3 = 0.5$. There are a total of four sub-scenarios generated under scenario (i) (Table 2.2) and six in scenario (ii) (Table 2.3).

Hypothesis testing at the level of 0.05 was carried out by various methods, including time-to-first-event analysis (TTFE), combined-recurrent-event analysis (CRE), Joint Frailty Model (JFM), Ghosh-Lin Method (GL), Nelsen-Aalen Method (NA), and Proportional win-fraction regression (WR). The testing based on the Cox-PH model with time to terminal event only (TE) was also conducted as a reference. We evaluate the performance of methods by the empirical type I error rate and the power. For that, we report the rejection rate out of 1000 Monte Carlo data sets for each method. The rejection rate is the empirical type I error rate when both β_1 and β_2 are 0 and the power otherwise.

2.4.1.2 Results

The type I error control is essential when comparing the power of different methods. The type I error rates under different scenarios are shown in Table 2.1, where the treatment has no effect on recurrent or terminal events ($\beta_1 = \beta_2 = 0$). When the null hypothesis (H_0) is true, the empirical

Table 2.1: Type I Error Check ($\beta_1 = \beta_2 = 0$)

Scenario settings	λ_0	0.3	1.0
	γ_0	1.0	0.3
	β_1	0	0
	β_2	0	0
	β_3	0.5	0.5
Type I error rates	TE	0.050	0.051
	TTFE	0.050	0.051
	CRE	0.048	0.050
	JFM	0.043	0.066
	GL	0.055	0.058
	NA	0.051	0.053
	WR	0.047	0.048

TE: Cox-PH model with time-to-terminal-event outcome; TTFE: Cox-PH model with the time-to-first-event outcome; CRE: Combined-recurrent-event analysis using the proportional rate model; JFM: the Joint Frailty Model; GL: Ghosh-Lin method; NA: Nelsen-Aalen method; WR: the Proportional Win-fraction regression.

type I error rate is close to the nominal level for all methods under all settings, except for the method based on the joint frailty model in some settings. We note that the proportional rate model leads to a test with a well-controlled type I error rate, even though the test is based on some specified model. This is expected because, under the null hypothesis, the model is correctly specified, with the coefficient of the treatment indicator being zero. When the alternative hypothesis is true, and the specified model is misspecified, then the power of the test would be affected. However, the test based on the joint frailty model sometimes has an inflated type I error rate (Table 2.1). An explanation is that even under the null hypothesis, the modeling part regarding the frailty may be misspecified, leading to incorrect type I error control.

We focus on comparing the power of different methods below. Under scenario (i) (Table 2.2 and Figure 2.2), where the hazard of mortality is high, and the intensity of recurrent events is low, most subjects experience only one or two recurrent events during the study period. Under scenario (i.a), the treatment only directly affects the terminal events but not the recurrent events. As Figure 2.2a shows, TE, which is based on the time to terminal event only, has overall the largest power.

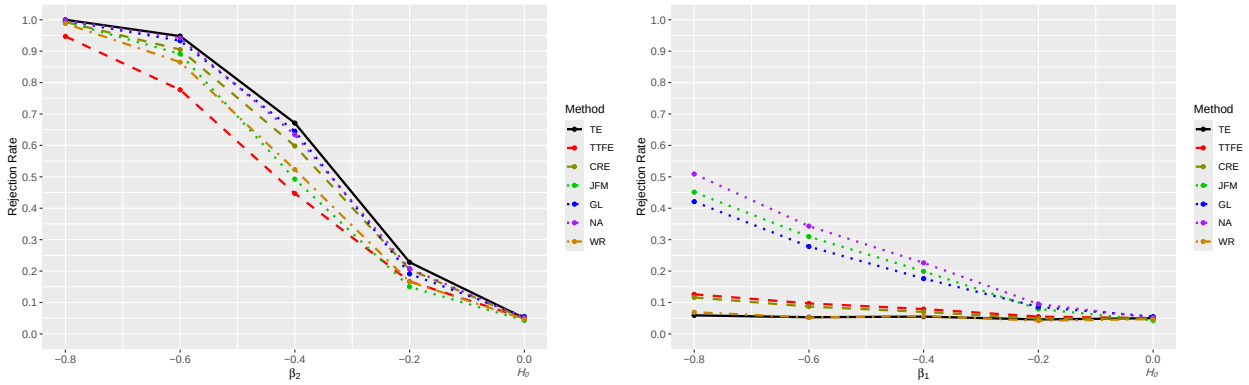
Table 2.2: Simulation scenario (i)

λ_0^a	γ_0^b	β_1	β_2	β_3	β_4	Sub-scenarios
0.3	1	0	varying^c	0.5	0	(i.a)
		varying^c	0	0.5	0	(i.b)
		-0.3	varying^c	0.5	0	(i.c)
		varying^c	-0.3	0.5	0	(i.d)

^{a} Baseline recurrent event rate;
^{b} Baseline terminal event hazard;
^{c} 0, -0.2, -0.4, -0.6, -0.8.

This is reasonable since there is little information from the recurrent events, and hence, statistical methods incorporating information on recurrent events will not help but add complexity. GL, NA, and CRE perform similarly and are slightly worse than TE. JFM and WR are similar and slightly worse than GL, NA, and CRE. We note that TTFE is the poorest-performing method in almost all instances. Under scenario (i.b), the treatment only directly affects the recurrent events but not the terminal events, and the treatment indirectly affects the terminal events through its effect on the recurrent events. Methods that jointly consider both the recurrent and terminal event processes, namely, JFM, NA and GL, have much larger power than other methods. In contrast, the TTFE and CRE have much lower power, and the win-ratio method (WR) and the method based on the terminal event only (TE) exhibit almost no power. Under scenarios (i.c) and (i.d), the treatment directly affects both the recurrent and terminal events, and we vary the strength of effect on terminal events (β_2) in (i.c) and on recurrent events (β_1) in (i.d), respectively. The joint analysis methods, NA and GL, have good power overall. The CRE accounts for both types of events by combining them together and also has large power when the treatment effect (β_2) is large, but the power is smaller than the joint analyses when the treatment effect is small under some scenarios. TTFE has less power than NA, GL and CRE. WR has the worst power overall.

Under scenario (ii) (Table 2.3 and Figure 2.3), where the hazard of mortality is low and the intensity of recurrent events is high, most subjects experience more recurrent events and longer survival compared to scenario (i). In scenario (ii.a), the treatment only affects terminal events directly but not recurrent events. As shown in Figure 3a, the non-parametric joint analysis methods



(a) Scenario (i.a)

(b) Scenario (i.b)

(c) Scenario (i.c)

(d) Scenario (i.d)

Figure 2.2: Simulation results of scenario (i). H_0 indicates that the rejection rate is a type I error; otherwise, it is a power. Acronyms: TE: Cox-PH model with time to terminal events only; TTFE: Cox-PH model with the time-to-first-event outcome; CRE: Combined-recurrent-event analysis using the proportional rate model; JFM: joint frailty model; GL: Ghosh-Lin method; NA: Nelsen-Aalen method; WR: proportional win-fraction regression.

Table 2.3: Simulation scenario (ii)

λ_0^a	γ_0^b	β_1	β_2	β_3	β_4	Sub-scenarios
1	0.3	0	varying^c	0.5	0	(ii.a)
		varying^c	0	0.5	0	(ii.b)
		-0.3	varying^c	0.5	0	(ii.c)
		varying^c	-0.3	0.5	0	(ii.d)
		-0.3	varying^c	0.5	0.1	(ii.e)
		varying^c	-0.3	0.5	0.1	(ii.f)

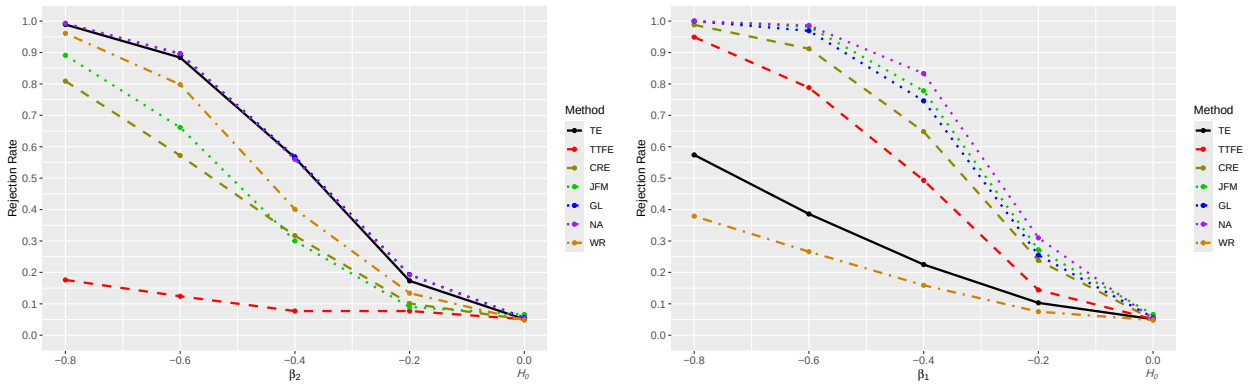
^{a} Baseline recurrent event rate;
^{b} Baseline terminal event hazard;
^{c} 0, -0.2, -0.4, -0.6, -0.8.

(GL and NA) have the largest power among all methods, followed by WR, JFM and CRE. And TTFE has significantly much lower power than other methods. In scenario (ii.b), the treatment only directly affects the recurrent events but not the terminal events, and treatment can indirectly affect the terminal event through its effect on recurrent events. As shown in Figure 3b, the joint analysis methods (NA, JFM, and GL) have larger power, followed by the CRE method. Of note, the WR method has considerably much lower power than all other methods. Under scenarios (ii.c) and (ii.e), the treatment directly affects both the recurrent (β_1) and terminal events (β_2), with β_1 fixed and β_2 varying. Overall, methods that jointly account for both the recurrent and terminal events (NA, JFM, and GL) have the best performance, especially when the direct effect of treatment on recurrent events is small (β_2 close to 0). CRE that considers both recurrent and terminal events also performs well when the treatment effect (β_2) is large. The TTFE has much lower power overall. WR suffers from very low power when the treatment effect on terminal events (β_2) is small.

Under scenarios (ii.d) and (ii.f), the treatment directly affects both the terminal (β_1) and recurrent events (β_2), with β_1 varying and β_2 fixed. Overall, CRE and joint analysis methods (JFM, GL, and NA) show larger power. However, the TTFE and WR have undesirable power compared to other methods.

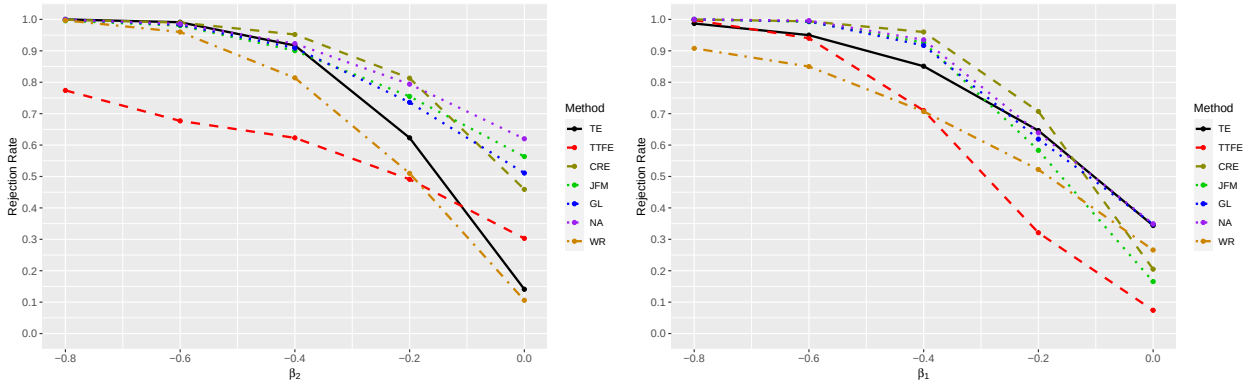
The key findings of the results presented are summarized below.

- The non-parametric joint testing methods (GL/NA) and Combined-recurrent-event analysis (CRE) demonstrate superior performance in power, and their relative performance depends



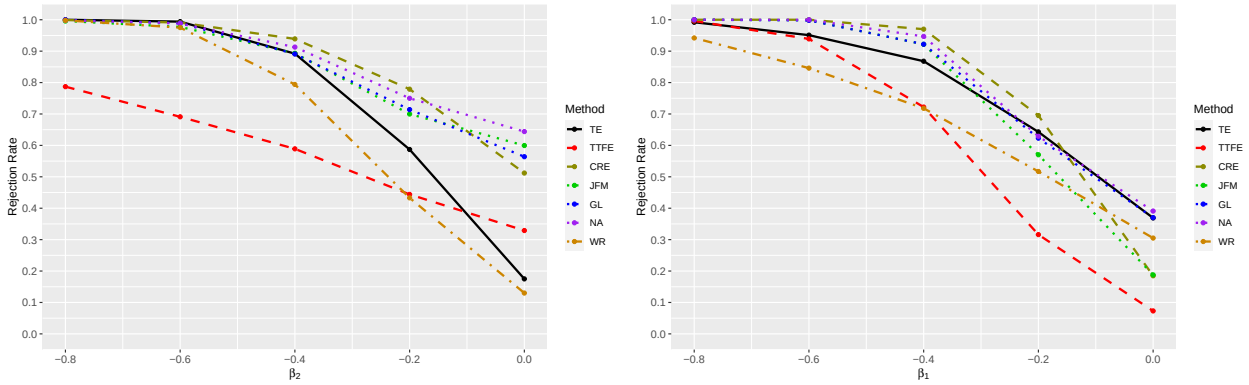
(a) Scenario (ii.a)

(b) Scenario (ii.b)



(c) Scenario (ii.c)

(d) Scenario (ii.d)



(e) Scenario (ii.e)

(f) Scenario (ii.f)

Figure 2.3: Simulation results of scenario (ii). H_0 indicates that the rejection rate is a type I error; otherwise, it is a power. Acronyms: TE: Cox-PH model with time to terminal events only; TTFE: Cox-PH model with the time-to-first-event outcome; CRE: Combined-recurrent-event analysis using the proportional rate model; JFM: joint frailty model; GL: Ghosh-Lin method; NA: Nelsen-Aalen method; WR: proportional win-fraction regression.

on specific scenarios.

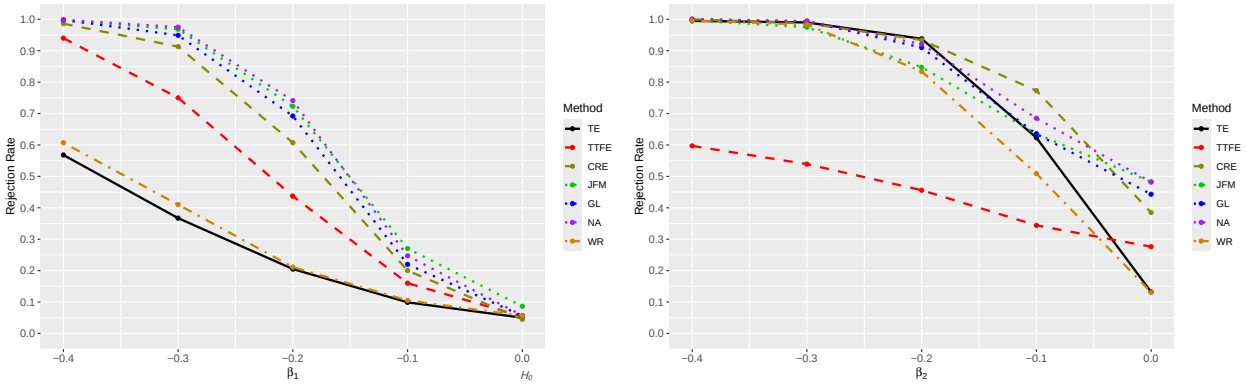
- Time-to-first-event Analysis (TTFE) and Win-fractions Regression (WR) exhibit low power across most scenarios.

2.4.1.3 Additional Simulations and Computation

We conducted additional simulations under modified scenarios to evaluate the robustness of our findings. Specifically, we altered the following parameters in the data generation process: (a) the censoring rate was increased from 20% (±0.1%) to 30% (±0.1%); (b) terminal event times were generated from a Weibull distribution with a shape parameter of 2, instead of an exponential distribution. All other parameters remained unchanged. The additional scenarios, labeled (iii.a) and (iii.b), correspond to the original scenarios (ii.b) and (ii.c), respectively.

The results for scenarios (iii.a) and (iii.b) are presented in Figures 2.4a and 2.4b, respectively. Overall, the relative performance patterns of the methods are similar to those observed in scenarios (ii.b) and (ii.c), shown in Figure 2.3.

In addition to evaluating statistical properties, we assessed the computational performance of the methods under scenario (ii.b) with $\beta_1 = -0.2$. Specifically, we examined computation time and convergence rate. Table2.4 reports the average computation time (in seconds) for seven methods across 1,000 Monte Carlo replications, for sample sizes of n = 200, 500, and 1,000. Overall, the computation times for all methods remain within a reasonable range and do not present practical concerns. JFM is relatively more computationally intensive than the other methods, except for GL and NA where bootstrapping is used for variance estimation. The computation time of WR increases faster than that of other methods as the sample size increases, likely due to its pairwise comparison procedure. In terms of convergence, JFM failed to converge in 50, 23, and 17 times out of 1000 replications for sample sizes of 200, 500, and 1000, respectively. These results suggest that JFM is less computationally appealing than other methods.



(a) Scenario (iii.a)

(b) Scenario (iii.b)

Figure 2.4: Additional simulations. H_0 indicates that the rejection rate is a type I error; otherwise, it is a power. Acronyms: TE: Cox-PH model with time to terminal events only; TTFE: Cox-PH model with the time-to-first-event outcome; CRE: Combined-recurrent-event analysis using the proportional rate model; JFM: joint frailty model; GL: Ghosh-Lin method; NA: Nelsen-Aalen method; WR: proportional win-fraction regression.

Table 2.4: Time Complexity Analysis (time in seconds)

Sample size (n)	TE	TTFE	CRE	JFM	GL	NA	WR
200	0.019	0.028	0.127	8.973	34.188	33.651	1.405
500	0.013	0.014	0.123	21.132	57.221	57.370	9.995
1000	0.007	0.019	0.219	40.454	100.812	101.357	41.152

2.4.2 Two Types of Recurrent Events

Simulations in this section consider the setting where there are two types of recurrent events in addition to a terminal event. Specifically, simulation studies are conducted to provide some insight into the following question. Suppose there is a primary type of recurrent event (type A) of interest. Researchers may be uncertain whether it is beneficial to include another type of recurrent event (type B), which has lower clinical relevance but a higher event rate, in terms of statistical power.

2.4.2.1 Simulation Setup

Similar to section 2.2, we denote the counting processes of type A and B recurrent events as $N^{T1}(t)$ and $N^{T2}(t)$. The general settings of $n = 200$, the randomized design, and the maximum follow-up time $\tau = 5$ remain the same as in section 2.4.1.1. Formally, for each subject, the recurrent events were generated by the Poisson processes with the following intensities

$$\lambda_1(t) = \lambda_{01} \exp\left\{\beta_{11} A + \beta_{41} N^{T1}(t)\right\},$$
$$\lambda_2(t) = \lambda_{02} \exp\left\{\beta_{12} A + \beta_{42} N^{T2}(t)\right\},$$

where $N^{T1}(t)$ and $N^{T2}(t)$ are the cumulative numbers of type A and B recurrent events by time t, λ_{01} and λ_{02} are the baseline intensity functions of type A and B recurrent events, respectively.

The terminal events were generated by the model with the following hazard function

$$\gamma(t) = \gamma_0 \exp\left\{\beta_2 A + \beta_{31} N^{T1}(t) + \beta_{32} N^{T2}(t)\right\},$$

where γ_0 is the constant baseline hazard. A censoring indicator was generated for each subject to ensure the sample-wide censoring rate is 20% (±0.1%). Figure 2.5 demonstrates the causal relationship between the treatment, type A and B recurrent events, and terminal events specified by the data generating models.

In the simulation study, we consider type A recurrent events as the primary type, with a baseline

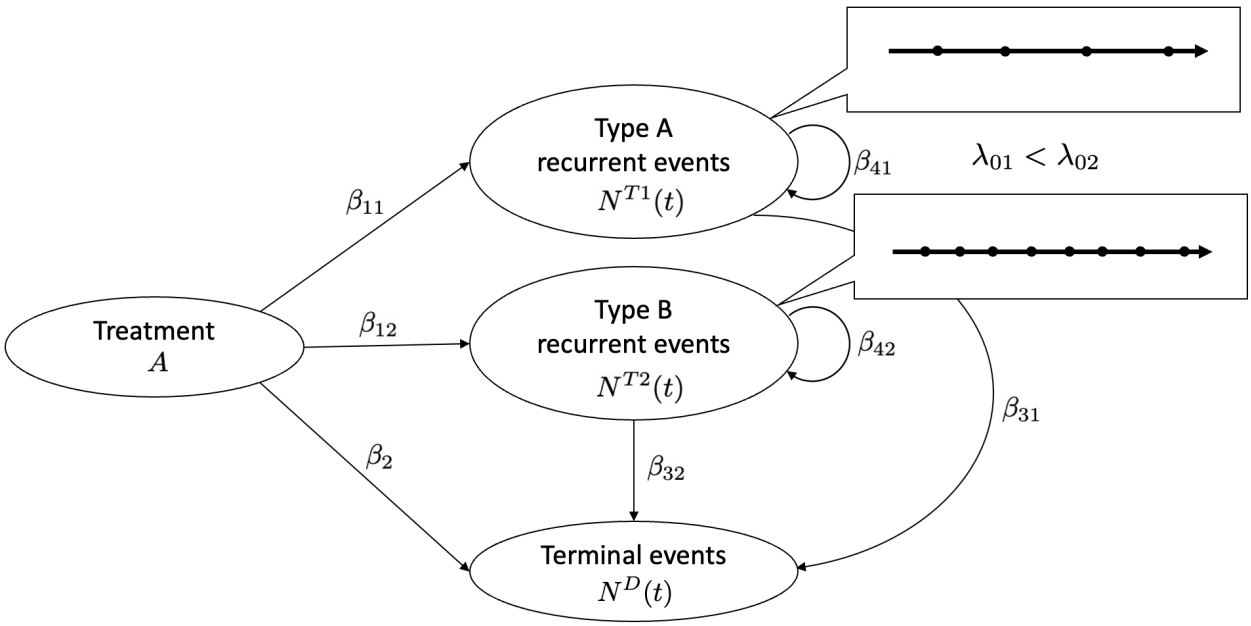
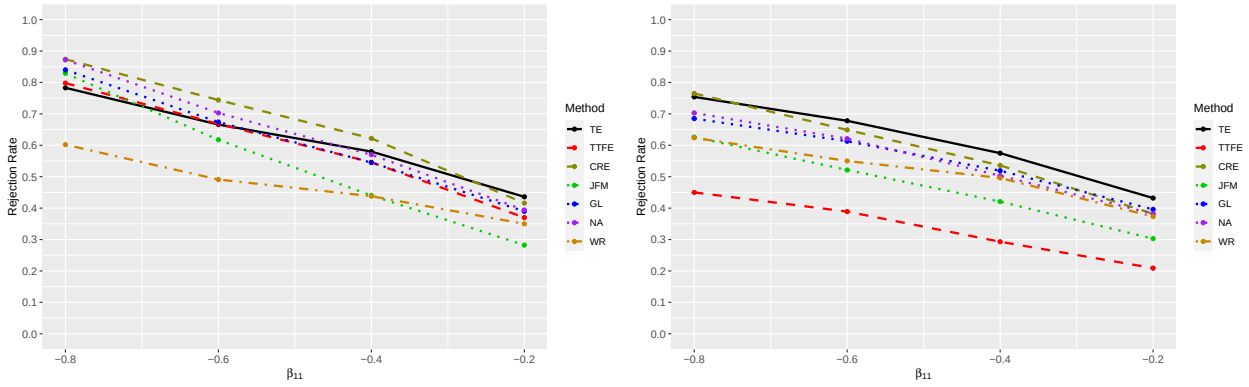


Figure 2.5: Causal diagram of the relationship between the treatment, recurrent events, and terminal events in the data generation model with two types of recurrent events.

rate $\lambda_{01} = 0.5$, and a direct effect on terminal events with $\beta_{31} = 1$. Compared to type A, type B recurrent events have a higher frequency with a baseline rate of $\lambda_{02} = 1$ and less clinical relevance in the sense that its effect on death is smaller with $\beta_{32} = 0.5$. The treatment effect on type A recurrent events varies as $\beta_{11} = -0.2, -0.4, -0.6, -0.8$, while type B recurrent events are less affected by the treatment with a fixed treatment effect of $\beta_{12} = -0.1$. Other parameters are set as $\gamma_0 = 0.5$, $\beta_2 = -0.3$, and $\beta_{41} = \beta_{42} = 0.1$. We implemented each method in two ways. In (iv.a), only type A recurrent events but not type B are considered as outcomes of interest, and in (iv.b), both type A and type B events are considered recurrent events of interest without differentiating the type. In both ways, the terminal event is handled the same as before.

2.4.2.2 Results

The results of (iv.a) (excluding type B recurrent events) and (iv.b) (including type B recurrent events) are presented in Figure 2.6a and 2.6b, respectively. Both (iv.a) and (iv.b) are under the same scenario where the treatment can directly affect both recurrent and terminal events. CRE outperforms all other methods in most scenarios.



(a) (iv.a) excluding type B recurrent events

(b) (iv.b) including type B recurrent events

Figure 2.6: Simulation results: Two types of recurrent events. Acronyms: TE: Cox-PH model with time to terminal events only; TTFE: Cox-PH model with the time-to-first-event outcome; CRE: Combined-recurrent-event analysis using the proportional rate model; JFM: joint frailty model; GL: Ghosh-Lin method; NA: Nelsen-Aalen method; WR: proportional win-fraction regression.

Compared to (iv.a), both the joint analysis methods (JFM, GL, and NA) and composite-outcome methods (TTFE and CRE) perform worse in (iv.b) with type B recurrent events included. WR performs slightly better in (iv.b).

2.5 Discussion

In this paper, we first reviewed the mainstream statistical methods for analyzing the composite of recurrent and terminal events. We categorized the methods into three classes: composite-outcome methods, joint-analysis methods, and win-ratio type methods. For the composite-outcome methods, the Time-to-First-Event (TTFE) analysis leverages solely the information on the first event experienced by a subject, irrespective of the type of the event. That is, the outcome information is summarized as the minimal order statistic of all event times. The Combined-Recurrent-Event (CRE) analysis exploits the data of all events without distinguishing between event types. That is, the outcome information is summarized as the order statistics of all event times without distinguishing the types. In contrast to the composite-outcome methods, where information is condensed into a single process (time-to-event process in TTFE and recurrent-event process in CRE), joint-analysis

methods introduce further granularity by differentiating between recurrent and terminal events. The outcomes are represented by two random processes. From joint-analysis methods, CRE, to TTFE, more and more information is lost during the data processing. The lost information cannot be recovered by any statistical techniques, which potentially causes reduced efficiency. Win-ratio type methods are a somewhat special type of method in that they incorporate clinical relevance and importance into the testing procedure. It has been compared with TTFE through simulations in the literature, but its performance relative to other methods is less studied.

Comprehensive simulation studies are conducted to evaluate the performance of various methods reviewed above under various scenarios regarding statistical power and type I error. We emphasize that, in theory, no single method outperforms other methods uniformly across all scenarios; performance always depends on the specific context. However, some general patterns emerge across simulations. These patterns are not only meaningful and interpretable but can also provide practical guidance for method selection. Some important results and messages are highlighted here. First, all methods except the joint frailty model (JFM) have a well-controlled type I error rate, including the testing methods based on the proportional rate model. This is because those methods, except for JFM, regardless of parametric or non-parametric, are based on correct models when the null hypothesis is true, that is, there is no treatment effect on recurrent or terminal events. The JFM, on the other hand, could be misspecified due to the frailty term even under the null hypothesis, leading to potentially incorrect type I error control. Second, non-parametric joint testing (GL/NA) and CRE have the best power overall across different scenarios, especially compared to the TTFE method. These results are intuitively reasonable because, as explained above, the joint test approach and the CRE inherently exploited more information than the TTFE method. Additionally, the relative power of the non-parametric joint testing methods to CRE depends on the scenarios. Although joint testing methods are based on two processes, which contain more information than the single process of the composite outcome used in the CRE method, the latter may outperform in certain scenarios. This is because CRE employs a one-degree-of-freedom test, in contrast to the non-parametric joint testing methods (GL/NA), which depend on two-degree-of-freedom tests. Overall, given their (CRE, GL,

and NA) robustness regarding type I error control, convenience in implementation, and superior power, they are the recommended methods in practice. In other words, they should be preferred to the popular method of TTFE. Third, the Win-Ratio type method has more undesirable power than other methods, especially compared to the non-parametric joint testing methods (GL/NA) and CRE. The Win-Ratio type method was mainly proposed to improve power over the TTFE method, with the main considerations in the order of priority being clinical relevance and interpretation. It is not expected, both theoretically and intuitively, to have larger power than the nonparametric joint testing approach and CRE, which also align with our simulation results. Lastly, we point out the joint frailty model approach not only may have incorrect type-I error control, but it also may suffer from issues with the convergence of the algorithm (Toenges & Jahn-Eimermacher 2020). The incorrect control of type I error is often due to model misspecifications. Although this issue can be mitigated through model diagnostics and by selecting alternative frailty distributions or alternative dependence structures such as copulas, these approaches are generally not robust and may not be well suited for hypothesis testing in clinical trial settings, for which robustness is a key consideration.

The simulation results in section 2.4.2.2 align with the observation in the CAPRICORN trial (Freemantle et al. 2003), where adding cardiovascular hospital admissions to the primary outcome of all-cause mortality alone diluted the estimated treatment effect. These results provide insight into how one decides whether to incorporate a type of event into the outcome, warning of the danger of the false belief that more events lead to more statistical power. Instead, one needs to carefully balance considerations of both the number of events and the treatment effect.

CHAPTER 3

Model-Assisted Causal Inference for the Treatment Effect on Recurrent Events in the Presence of Terminal Events

3.1 Introduction

Heart transplantation remains the gold standard for patients with end-stage heart failure. However, its use is severely limited by the shortage of donor organs, with only about 3,000 adult transplants performed annually in the United States. Durable mechanical circulatory support (MCS) has become an important treatment option in clinical practice, with more than 2,500 ventricular assist device (VAD) implantations reported annually in the Interagency Registry for Mechanically Assisted Circulatory Support (INTERMACS) since 2006. These data motivate the methodological development of this paper and are analyzed in our empirical study. In current clinical practice, continuous-flow left ventricular assist devices (LVADs) account for the majority of implants, whereas biventricular assist devices (BiVADs) are used less frequently and are typically reserved for patients with severe right ventricular dysfunction. Although these devices have markedly improved survival and quality of life, their use is accompanied by a substantial burden of adverse events, including bleeding, infection, stroke, and device malfunction. Among these, gastrointestinal bleeding (GIB) is one of the most frequent and clinically consequential complications, often requiring repeated hospitalizations, transfusions, and endoscopic procedures. The recurrent GIB

has been linked to increased mortality and reduced quality of life (Hammer et al. 2024). A clinically important question is whether the choice between BiVAD and LVAD implantation leads to different risks of recurrent GIB events among patients receiving mechanical circulatory support, while accounting for the fact that death terminates the observation of future bleeding events.

Despite its clinical importance, this question remains unresolved for two main reasons. First, single-center studies often lack sufficient sample sizes of patients receiving MCS devices to support adequately powered statistical comparisons. Second, randomized clinical trials are logistically difficult and ethically complex; patients and clinicians may be reluctant to randomize device type due to uncertainty about comparative risks. Consequently, the existing literature provides inconclusive evidence regarding the relative benefits of the two devices. Some studies suggest that early planned BiVAD support improves certain clinical outcomes relative to delayed LVAD-to-BiVAD conversion (Fitzpatrick et al. 2009), whereas others report higher bleeding and infection rates among BiVAD recipients (Cleveland et al. 2011). Large-scale registry data such as INTERMACS may provide an opportunity to address this question using a substantially larger patient population. Of note, INTERMACS aggregates patient-level data from hospitals nationwide, and preliminary findings have uncovered lower average survival for BiVAD patients (Kirklin et al. 2015). It is worth pointing out that prior analyses cannot disentangle whether differences in bleeding reflect true device effects or underlying patient survival differences, which are the major confounding factors of the effect of the devices to the risk of GIB. The differential survival distributions between the two device groups complicates the analysis and interpretation. These challenges motivate the development of rigorous and interpretable statistical methods that can leverage large clinical registry databases such as INTERMACS to obtain valid evidence on device-related risks.

Our investigation focuses primarily on the risk of recurrent GIB, a leading post-surgical complication that substantially contributes to morbidity, healthcare cost, and resource utilization. Because GIB may occur multiple times during postoperative follow-up, recurrent event analysis provides a natural framework for studying this outcome (Andersen & Gill 1982, Pepe & Cai 1993, Lawless & Nadeau 1995). A major challenge in applying these methods arises from the presence of a terminal

event such as death, which precludes the occurrence of future bleeding episodes. Treating death simply as a censoring event yields inference conditional on survival, which may bias comparisons between devices when their survival profiles differ and may lack causal interpretability.

Several statistical approaches have been proposed to address survival bias in recurrent event analyses. The Ghosh–Lin estimator (Ghosh & Lin 2000) incorporates death into the estimation of the marginal mean number of recurrent events, but it may disadvantage groups with longer survival because more events accumulate over a longer period at risk. To mitigate this issue, while-alive (WA) estimands have been proposed to normalize the mean number of events by the restricted mean survival time, thereby providing a measure of the event rate per unit time alive (Mao 2023, Schmidli et al. 2023, Wei et al. 2023). Although the WA framework represents an important advance, it implicitly assumes that the recurrent event rate is constant over time. In practice, event intensities often evolve over the course of follow, for example, increasing shortly after surgery and declining thereafter. When the constant-rate assumption is violated, the WA test may exhibit inflated type I error when the baseline event rate varies over time and when survival distributions differ substantially between treatment groups. These limitations motivate the development of new methods that can accommodate both time-varying recurrent-event rates and differential survival distributions across treatment groups.

The concerns described above are particularly relevant in our motivating example. The type of ventricular assist device may influence GIB recurrence through two distinct pathways: a direct effect on the recurrent bleeding process (e.g., via shear stress and anticoagulation requirements) and an indirect effect mediated through survival, since patients who live longer remain at risk for bleeding for a longer period. Disentangling these two pathways is crucial for valid inference on the device-specific bleeding risk. In alignment with this clinical insight, we adopt the separable effects framework proposed by Stensrud et al. (2022) and extended to recurrent–terminal event settings by Janvin et al. (2024). This framework allows the treatment (i.e., device type) to be decomposed into two hypothetical components acting separately on the recurrent-event and terminal-event pathways; see Figure 3.2 and Section 3.3 for details. In brief, this decomposition yields an

interpretable estimand for the direct effect of treatment on recurrent events without conditioning on survival and accommodates time-varying baseline rates. Our contribution in this paper is to develop a new statistical inference method that builds on this framework. While existing work has primarily focused on estimation under separable effects, formal hypothesis testing procedures remain underdeveloped.

We develop a model-assisted, score-type hypothesis test, termed PR-MSMaT, which aims to test the direct effect of treatments on recurrent events in the presence of terminal events. Our approach builds on the separable effects framework to yield a global, time-integrated test with a clear causal mechanistic interpretation. We show that PR-MSMaT maintains correct type I error control under time-varying baseline event rates and differential survival distributions, thereby overcoming key limitations of existing WA-based tests.

The remainder of the chapter is organized as follows. Section 3.2 reviews hypothesis testing under the while-alive (WA) paradigm and demonstrates the limitations of the WA-based test through simulation. Section 3.3 presents the Separable Effects Framework (SEF) (Stensrud et al. 2022, Janvin et al. 2024), which is adopted in our development of a hypothesis testing procedure PR-MSMaT. Section 3.4 develops the Proportional Rate Marginal Structural Model-assisted Test (PR-MSMaT) and studies its theoretical properties. Section 3.5 presents simulation studies comparing PR-MSMaT with the WA-based test in terms of type I error control and statistical power under various scenarios. Section 3.6 applies the method to the INTERMACS dataset to compare recurrent gastrointestinal bleeding between ventricular assist devices, followed by some concluding remarks in Section 3.7.

3.2 While-Alive-based (WA) Test: A Revisit

Methods for analyzing recurrent events in the presence of a terminal event have evolved considerably over the past decades. The Nelson–Aalen (NA) estimator was among the earliest approaches, where death was treated as a censoring event for the recurrence process. This assumption implicitly

allows individuals to continue accruing events after death, leading to systematic overestimation of the recurrent event burden in settings with non-negligible mortality. To address this issue, acknowledging no further event after the survival time D, Ghosh & Lin (2000) (GL) proposed a nonparametric framework based on the mean frequency function, $\mu(t) = \mathbb{E}\{N^*(t)\}$, where $N^*(t)$ is the cumulative number of recurrent events up to time t. Ghosh & Lin (2000) also developed the corresponding hypothesis test procedure. However, the GL test does not properly account for the reality among patients receiving device implants; that is, the device type that prolongs survival may increase recurrent event counts simply because patients live longer, not necessarily because the device elevates morbidity.

To address this limitation, Mao (2023) considered an approach based on the while-alive (WA) estimand, which averages the number of cumulative recurrent events over survival time. Specifically, at a prespecified time τ, the While-Alive Loss Rate (WALR) is given by

$$\ell(\tau) = \frac{\mathbb{E}\{N^*(\tau)\}}{\mathbb{E}(D \wedge \tau)}.$$

This quantity represents the expected number of recurrent events per unit time in life over the interval $[0, \tau]$, so it corrects for the survival-induced bias inherent in both NA and GL methods. Mao (2023) further derived an inference for treatment comparisons with WALR, in which a modified estimand $L(t) = \ell(t)t$ is further considered to avoid the unstable behavior of $\ell(t)$ for small t, termed *the survival-completed cumulative loss function*.

It is worth pointing out that the test based on the WA estimand only incurs a fair comparison of the recurrent event if the true baseline recurrent event rate is constant over time. Arguably, this condition may not hold in many practical studies. For instance, in our motivating example, it is irrational to believe that patients' GIB risk trajectory remains constant over time. In such cases, the existing WA-based test may become invalid with inflated type I errors.

To illustrate this potential pitfall, we conducted a simulation experiment, emulating a randomized trial under time-varying recurrent event rates. For simplicity, we assumed independent recurrent

events and independence between recurrent and terminal events. The baseline recurrent event rate was set as a time-varying stepwise function, starting at 2 events per year and then rising by a factor of $\exp(0.5)$ every 1.2 years within a maximum follow-up of 5 years. Treatment effects were specified through two recurrent event rate ratios (RR = $1.00, 0.75$) together with three terminal event hazard ratios (HR = $1.0, 0.5, 0.2$), yielding six scenarios. Each scenario was simulated with $n = 500$ subjects and 100 rounds of Monte Carlo replications, and the WA function (survival-completed cumulative loss function) was estimated for treatment and control arms using an existing R package *WA* (Mao 2023). Figure 3.1 displays the WA functions averaged over 100 replications. From which, we learned:

(i) In those scenarios where the treatment has no impact on recurrent events (RR = 1) but does affect survival (HR $\neq$ 1), such as the scenario of RR = 1.0 and HR = 0.5, the WA estimates appeared substantially different between the two treatment groups. Consequently, a hypothesis test based on such WA estimates could have a poor type I error control.

(ii) In the settings where the treatment can truly affect both recurrent and terminal events, for example, the scenario of RR = 0.75 and HR = 0.2, the estimated WA curves crossed at year 4 or so (i.e., not really separable), which implied that the WA-based test could lose statistical power to test the treatment effect around year 4.

Such numerical evidence illustrates that with time-varying baseline event rates, the WA method may have inflated type I error and lose statistical power for hypothesis testing when the treatment acts through both the recurrent process and survival. This calls for the need for a new inference approach that can explicitly consider the pathways linking treatment, recurrent events, and terminal events.

3.3 Separable Effects Framework (SEF)

To better evaluate treatment effects on recurrent events, it is useful to consider two pathways: as shown in Figure 3.2, a ***direct effect***, where a treatment directly affects on the recurrent process not

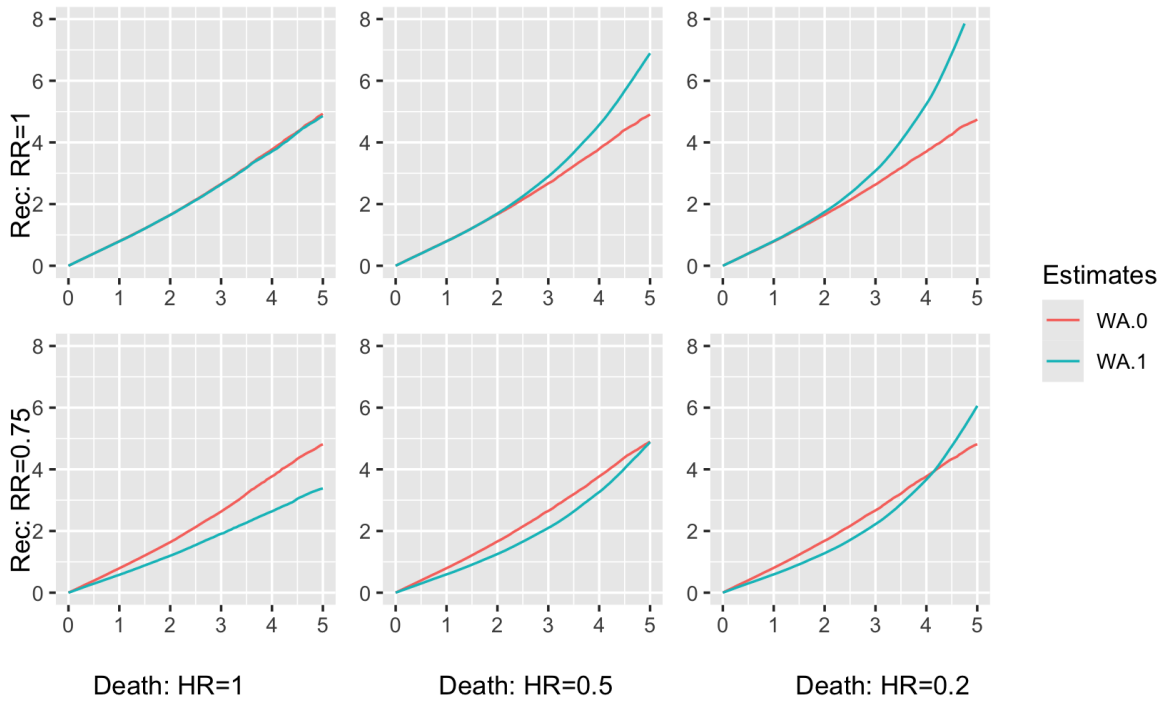


Figure 3.1: The simulation results illustrate the limitations of the While-alive-based test. WA.0: the survival-completed cumulative loss function for the control group; WA.1: the survival-completed cumulative loss function for the treatment group; RR: Recurrent event Ratio; HR: Hazard Ratio.

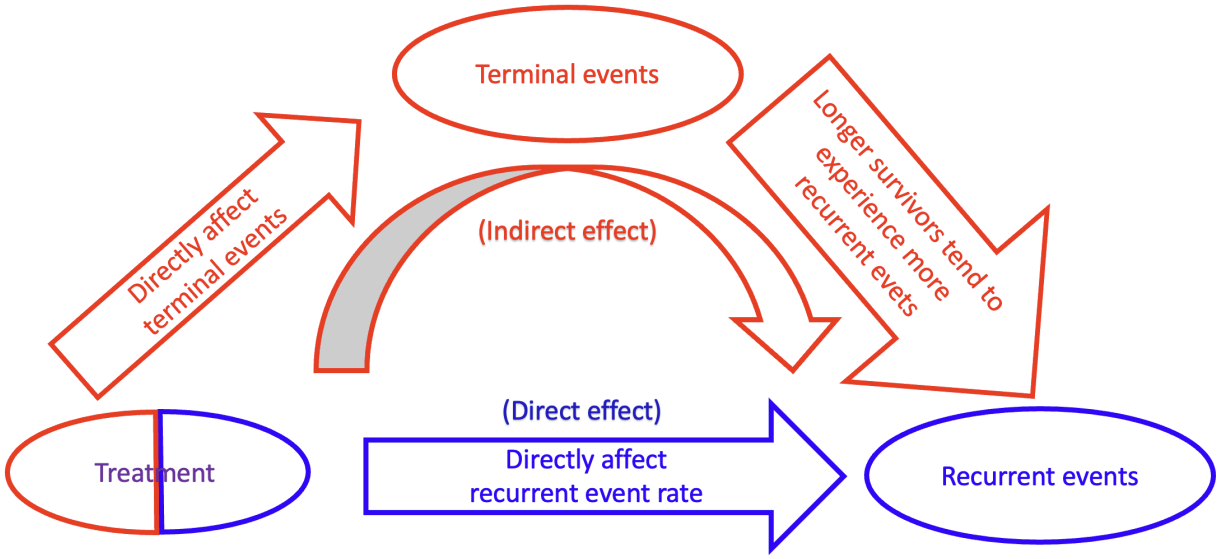


Figure 3.2: Intuition regarding the relationship between the treatment, recurrent events, and terminal events.

through survival, such as GIB in our motivating example, and an ***indirect effect***, where a treatment influences the recurrent process through survival since longer survivors tend to experience more recurrent event. The primary goal is to disentangle these two pathways and isolate the direct effect in the evaluation, while excluding the indirect survival-mediated effect (Stensrud et al. 2022).

The Separable Effects Framework (SEF) was previously introduced by Stensrud et al. (2022) for binary outcomes with competing events and later extended by Janvin et al. (2024) to recurrent event settings. Several other extensions were also proposed (Stensrud et al. 2021, 2023). Adopting this framework, we aim to develop a hypothesis testing procedure for the direct effect in the SEF paradigm. In this paper, we mainly consider a randomized trial setting, in which we focus on the relationships between the treatment, recurrent process, and terminal events, ignoring all the other confounders. In cases where the treatment was not randomized, the statistical techniques (such as matching or IPTW weighting) may be used to balance the treatment groups in order to apply this SEF method.

Let $A \in \{0, 1\}$ denote the primary treatment under investigation. Consider a discrete-time partition of the follow-up time window in the form $0 = t_0 < t_1 < \cdots < t_K = \tau$, where τ is the

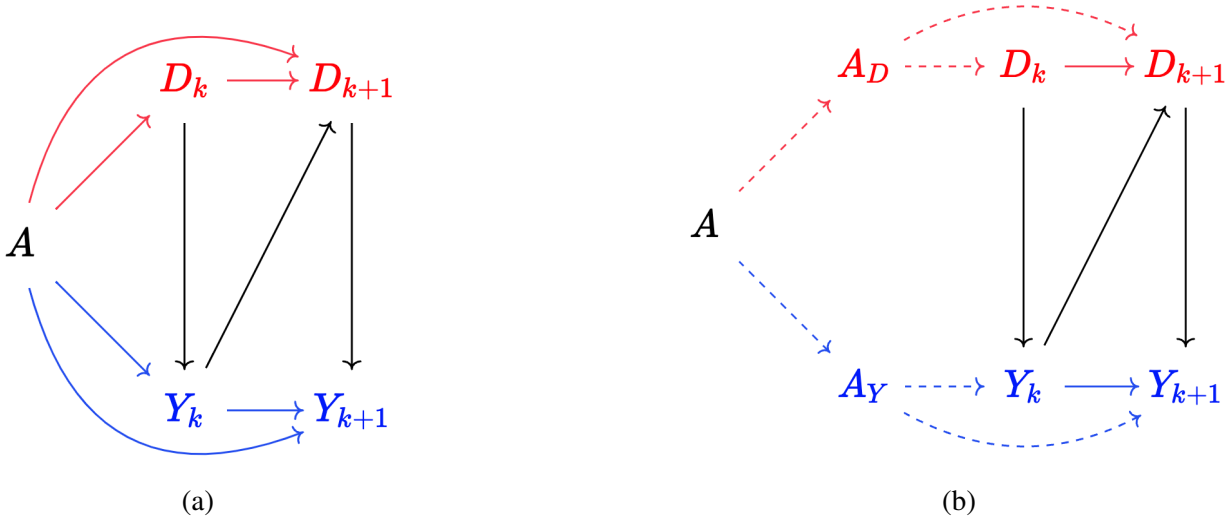
maximum follow-up time. For an interval cell $(t_{k-1}, t_k]$, let D_k be the death indicator within the time interval, and Y_k be the cumulative number of recurrent events by t_k. Without loss of generality, we define the topological order in a hypothesized Directed Acyclic Graph (DAG) shown in Figure 3.3a within the time interval $(t_{k-1}, t_k]$, with D_k occurring before Y_k. For convenience, for an arbitrary process X_k, we denote its change by $\Delta X_k := X_k - X_{k-1}$, and the history by $\bar{X}_k = \{X_j\}_{j=0}^{k}$.

In the SEF paradigm, the treatment A is supposed to be decomposed into two distinct treatment components: A_Y, which solely influences the occurrence of recurrent events, and A_D, which exclusively impacts terminal events, such that receiving $A_Y = A_D = a$ could result in the same outcomes as receiving $A = a$, $a \in \{0, 1\}$. See Figure 3.3b for the organization of these two components in the SEF framework. Although in practice, we cannot always identify the exact decomposition with two treatment components, conceptually we may leverage such a hypothetical framework to conduct a causal inference for their mechanistic pathways.

Conceptually, this decomposition separates the treatment effect into two distinct pathways: the one that directly influences the recurrent events, and the other that operates through terminal events. In the context of the INTERMACS application, this means distinguishing whether a device directly increases the frequency of bleeding events among patients who remain alive, versus whether differences in observed bleeding are driven by differential survival across device types. This distinction is critical because patients who survive longer have more opportunity to experience recurrent GIB events, even if the device itself does not directly increase bleeding risk. By fixing the survival pathway while varying the recurrent-event pathway, the separable effects framework enables a clinically interpretable comparison that isolates the direct effect of device type on recurrent GIB from the indirect effect mediated through survival.

Formally, define counterfactual recurrent outcomes as $Y_k^{a_Y, a_D}$, representing the cumulative number of recurrent events by t_k had the subject received $A_Y = a_Y$ and $A_D = a_D$. Similarly, we define the counterfactual death indicators $D_k^{a_Y, a_D}$. The counterfactual mean function is then given by

$$\mu_k^{a_Y, a_D} = E\left[Y_k^{a_Y, a_D}\right].$$



(a) (b)

Figure 3.3: Directed acyclic diagram (DAG) illustrates: (a) the causal pathways for treatment A, recurrent event Y_k, and terminal event D_k; (b) the hypothetical decomposition of treatment A into A_Y (directly affecting recurrent events) and A_D (directly affecting terminal events).

It follows that the **separable direct effect** is captured by contrasts of $\mu_k^{a_Y=1,a_D}$ versus $\mu_k^{a_Y=0,a_D}$ for fixed a_D, while the **separable indirect effect** by contrasts of $\mu_k^{a_Y,a_D=1}$ versus $\mu_k^{a_Y,a_D=0}$ for fixed a_Y (Janvin et al., 2023). Such definitions conceptually partition the overall treatment effect into distinct and interpretable pathways as shown in Figure 3.2.

To identify separable effects, we impose the following standard assumptions of causal inference:

(i) Consistency: If $A = a$, then $\bar{D}_K = \bar{D}_K^a$, $\bar{Y}_K = \bar{Y}_K^a$;

(ii) Exchangeability: $(\bar{Y}_K^a, \bar{D}_K^a) \perp A$;

(iii) Positivity: $f_{\bar{D}_{k+1},Y_k}(0, y_k) > 0$ implies $P(A = a|\bar{D}_{k+1} = 0, \bar{Y}_k = y_k) > 0$, for all $a \in \{0, 1\}$, $k \in \{0, \cdots, K\}$.

Additionally, for the SEF, the following conditions hold:

(iv) Dismissible component conditions:

$$Y_{k+1} \perp A_D | A_Y, \bar{D}_{k+1}, \bar{Y}_k;$$

$$D_{k+1} \perp A_Y | A_D, \bar{Y}_k, \bar{D}_k, \text{for all } k \in \{0, \cdots, K\}.$$

The conditions in (iv) ensure that the decomposed A_Y and A_D stay at non-overlapping pathways. Under the identification conditions, the G-formula for the counterfactual mean function is

$$\mu_k^{a_Y,a_D} = \sum_{t=1}^{k} E\left[\frac{I(A=a_Y)}{\pi_A(a_Y)} \prod_{j=0}^{t} \frac{\pi_{D_j}^{a_D}}{\pi_{D_j}^{a_Y}} \Delta Y_t\right], \quad k \in \{0, \cdots, K\},$$

where $\pi_A(\bullet) = P(A = \bullet)$ and $\pi_{D_j}^{\bullet} = f_j^{\bullet}(D_j)$, with $f_j^{\bullet}(\circ) = P(D_j = \circ | \bar{Y}_{j-1}, \bar{D}_{j-1}, A = \bullet)$.

For each time interval, the change in the mean function is estimated by

$$\Delta\hat{\mu}_t^{a_Y,a_D} = \hat{S}_t^{a_Y} \Delta\hat{B}_t^{Y}, \tag{3.1}$$

where $\hat{S}_t^{a_Y}$ is the discrete Kaplan–Meier estimator of the survival function for group a_Y and

$$\Delta\hat{B}_t^{Y} = \frac{\sum_{i=1}^{n} \hat{W}_{i,t}^{D}(a_Y, a_D) I(A_i = a_Y) Z_{i,t} \Delta Y_{i,t}}{\sum_{j=1}^{n} I(A_j = a_Y) Z_{j,t}},$$

where $Z_{i,t}$ is the at-risk indicator and

$$\hat{W}_{i,t}^{D}(a_Y, a_D) = \prod_{q=1}^{t} \left(\frac{1 - \Delta\hat{\Lambda}_{i,q}^{D|\mathcal{F}}(a_D)}{1 - \Delta\hat{\Lambda}_{i,q}^{D|\mathcal{F}}(a_Y)}\right)^{1-D_{i,q}} \left(\frac{\Delta\hat{\Lambda}_{i,q}^{D|\mathcal{F}}(a_D)}{\Delta\hat{\Lambda}_{i,q}^{D|\mathcal{F}}(a_Y)}\right)^{D_{i,q}},$$

with $\Delta\hat{\Lambda}_{i,q}^{D|\mathcal{F}}(a) = \hat{P}(D_{i,q} = 1 | \bar{D}_{i,q-1} = 0, \bar{Y}_{i,q-1}, A_i = a)$. The cumulative estimator is then $\hat{\mu}_k^{a_Y,a_D} = \sum_{t=1}^{k} \Delta\hat{\mu}_t^{a_Y,a_D}$ (Janvin et al. 2024).

The weighting process $\hat{W}_{i,t}^{D}(a_Y, a_D)$ attempts to mimic a hypothetical randomized trial on which A_Y and A_D may be implemented separately. Specifically, it adjusts the observed data distribution so that the counterfactual survival process under a_D is comparable to that under a_Y, taking effect only when $a_D \neq a_Y$. Intuitively, this calibration via the weights is used to eliminate the imbalance between groups even the treatment A is randomized, allowing us to recover the distribution of recurrent events had A_Y and A_D been independently randomized. The weights scale down those individuals who are more likely to die under the counterfactual a_D than under their observed

assignment, and scale up those who are less likely to die, effectively reshaping the risk set to reflect the distribution of recurrent events under the counterfactual intervention. In this way, the weighting aligns the survival experience across groups so that differences in the estimated recurrent event process can be attributed solely to the direct effect of A_Y, rather than being biased by differential survival induced by A_D.

The above review of the SEF introduces a framework in which the treatment can be rigorously decomposed into two components acting exclusively on recurrent and terminal events. The direct effect in the SEF, therefore, represents the contrast in the recurrent event process attributable solely to A_Y, while holding A_D fixed. By separating this key pathway of interest from the indirect effect, we can test hypotheses about the direct treatment effects on recurrent events while mitigating the distortions observed with the WA method under time-varying recurrent event rates.

3.4 Proportional Rate Marginal Structural Model-assisted Test (PR-MSMaT)

Building on the SEF, we develop a hypothesis test for the direct treatment effect on recurrent events. The logic is analogous to conventional survival analysis: in a similar spirit of the Kaplan–Meier estimator motivating both the log-rank test and the Cox proportional hazards model with treatment as the sole covariate, the separable estimators proposed by Janvin et al. (2024) motivate a Proportional Rate Marginal structural model (PR-MSM) along the lines of a score-based test.

Let $\mu^{a_Y,a_D}(t)$ denote the counterfactual marginal mean number of recurrent events by time t had the treatment components been set at (a_Y, a_D). The PR-MSM takes the form:

$$\mu^{a_Y,a_D}(t) = \mu^{a_Y=0,a_D}(t) \exp\left(\beta^{a_D} a_Y\right), \quad a_D \in \{0,1\}, \tag{3.2}$$

suggesting that, for each fixed a_D, the counterfactual mean functions across a_Y differ by a constant multiplicative factor over time. It follows that β^{a_D} can be estimated by solving the following

estimation equation:

$$\int_0^\tau \sum_{i=1}^{n} Z_i(t) A_i \left\{ d\hat{\mu}^{a_Y=1,a_D}(t) - \exp(\beta^{a_D})\, d\hat{\mu}^{a_Y=0,a_D}(t) \right\} = 0, \tag{3.3}$$

where $\hat{\mu}^{a_Y,a_D}(t)$ is the counterfactual mean function estimator and $Z_i(t)$ is the at-risk process. See Appendix A for the detailed derivation for (3.3).

Under $H_0 : \beta^{a_D} = 0$ (no direct treatment effect), the left-hand side of (3.3) reduces to

$$U_n \;=\; \int_0^\tau \sum_{i=1}^{n} Z_i(t) A_i \left\{ d\hat{\mu}^{a_Y=1,a_D}(t) - d\hat{\mu}^{a_Y=0,a_D}(t) \right\}. \tag{3.4}$$

Let $\widehat{\mathrm{Var}}(U_n)$ denote a consistent estimator of $\mathrm{Var}(U_n)$, which can be estimated by the transforming cumulative hazard estimates based on differential equations (Ryalen et al. 2018). Our PR-MSM-assisted Test (PR-MSMaT) statistic is formed as the standardized score test statistic:

$$T_n \;=\; \frac{U_n}{\widehat{\mathrm{Var}}(U_n)^{1/2}}.$$

Suppose that (i) $\hat{W}_{i,t}^{D}(a_Y, a_D)$ and $\hat{S}_t^{a_Y}$ are uniformly consistent to their population parameters; (ii) the risk process $Z_i(t)$ is bounded away from 0 and ∞; and (iii) the standard martingale Lindeberg conditions hold. Then, under $H_0 : \beta^{a_D} = 0$, the test statistic $T_n \xrightarrow{d} N(0,1)$ as $n \to \infty$.

Theorem 3.4 shows that the proposed PR-MSMaT is a valid test for the null hypothesis of no separable direct treatment effect on recurrent events. See Appendix B for the proof of Theorem 3.4.

The PR-MSMaT extends the separable effect estimators by providing a formal hypothesis test that summarizes treatment effects across the pre-specified follow-up time period. Unlike the original estimators, which yield time-specific contrasts that may be harder to synthesize, the proposed test delivers an overall inference for the direct treatment effect. It enjoys interpretability for broader audiences in clinical practice; in particular, it provides the causal interpretability within the SEF paradigm, in a similar way to the log-rank or Cox score test widely used in practice. Simulation

studies in Section 3.5 show that our PR-MSMaT maintains a proper type I error control and achieves higher power than while-alive-based tests in scenarios where recurrent event rates vary over time, a realistic situation encountered often in practice, including our motivating example.

3.5 Simulation Experiment

Following the simulation design given in Janvin et al. (2024), we evaluated the finite-sample performance of our PR-MSMaT compared with the WA-based test under a 2-arm randomized trial for treatment A, with treatment group $A = 1$ and the control group $A = 0$. In this case, all contrasts are evaluated by treatment versus control. Outcomes were generated under an additive intensity model over $K = 1000$ intervals $(t_{k-1}, t_k]$, $k = 1, \ldots, 1000$. For each interval, we consider

$$P(\Delta Y_k = 1 \mid A_Y, D_k = 0) = \beta_{Y,0,k} + \beta_{Y,A} A_Y,$$
$$P(D_k = 1 \mid A_D, \bar{Y}_{k-1}) = \beta_{D,0} + \beta_{D,A} A_D + \beta_{Y,D} Y_{k-1},$$

with at most one recurrent event per interval. We fix $\beta_{D,A} = -0.5/K$, set the null case with $\beta_{Y,A} = 0$ to evaluate type I error scenarios, and $\beta_{Y,A} = -0.5/K$ for power evaluation. Each configuration is examined via $N = 1000$ rounds of Monte Carlo replicates with sample size $n = 1000$. The three types of baseline recurrent-event rate $\beta_{Y,0,k}$ are designed according to the following temporal patterns:

(a) **Constant:** $\beta_{Y,0,k} = 2/K$ for all k.

(b) **Decreasing:** Starting at $3/K$ and declining by $0.5/K$ every 200 intervals (down to $1/K$).

(c) **Increasing:** Starting at $1/K$ and rising by $0.5/K$ every 200 intervals (up to $3/K$).

We examine three types of hypotheses: **Two-sided:** $H_0 : \theta = 0$ vs. $H_1 : \theta \neq 0$; **Left-sided:** $H_0 : \theta \geq 0$ vs. $H_1 : \theta < 0$; **Right-sided:** $H_0 : \theta \leq 0$ vs. $H_1 : \theta > 0$. The parameter θ represents the difference in recurrent events between treatment and control groups, which refers to β^{a_D} for our PR-MSMaT and the WALR ratio for the WA-based test.

Simulation results are summarized in Figure 3.4. When the baseline recurrent event rate is constant (i.e., scenario (a)), both our PR-MSMaT and the WA test maintain a nominal type I error rate and achieve similar power. This agreement under time-homogeneous rates confirms that both methods are well-behaved in this scenario. In contrast, when the baseline recurrent event rate varies with time, their performances diverge. Under a decreasing rate (i.e, scenario (b)), WA exhibits inflated type I errors for both the two-sided test and the left-sided test. Under an increasing rate (i.e., scenario (c)), WA again shows inflated type I errors for both the two-sided test and the right-sided test. Amazingly, our PR-MSMaT clearly controls type I errors well in all scenarios. In terms of power, our PR-MSMaT achieves comparable power to the WA test in scenario (b) and even larger power than WA in scenario (c). Note that, in both scenarios (b) and (c), WA has the disadvantage of inflated type I errors. This demonstrates that PR-MSMaT performs reliably against nonstationary baseline recurrent-event processes and clearly outperforms the WA test.

Directional test results help explain the WA behavior. The WA test summarizes the data by scaling the cumulative number of recurrent events by the restricted mean survival time, in which the "while-alive rates" are compared across arms. This procedure implicitly treats the observed total event count of each subject as if it arose from a roughly constant per-unit-time rate over that subject's observed survival. When the true baseline rate decreases over time, most events happen earlier, so a subject who dies earlier tends to have a larger fraction of their lifetime spent in the high-rate period. In this case, WA imputes a higher "average event rate while alive" to that shorter-lived profile, even if the subject is not actually more event-prone than someone who lives longer. As a result, WA can systematically take earlier deaths as evidence of a higher event intensity in the control arm; this pushes the estimated treatment effect (treatment vs. control) downward and inflates left-sided type I error. Likewise, when the baseline rate increases over time, events concentrate later, so longer survivors accumulate more events simply because they remain under observation during the high-rate period. WA handles it as an arm-specific elevation in the "while-alive rate", biasing the estimated treatment effect upward and inflating right-sided type I error.

By contrast, our PR-MSMaT directly targets the separable, pathway-specific effect of treatment

on the recurrent event process, holding the survival pathway fixed. Because it does not attribute differences driven purely by unequal follow-up windows in high- or low-rate portions of the hazard curve to a treatment effect, our PR-MSMaT avoids these directional biases. As a result, it preserves nominal type I error across all time-varying baseline patterns while still delivering competitive power.

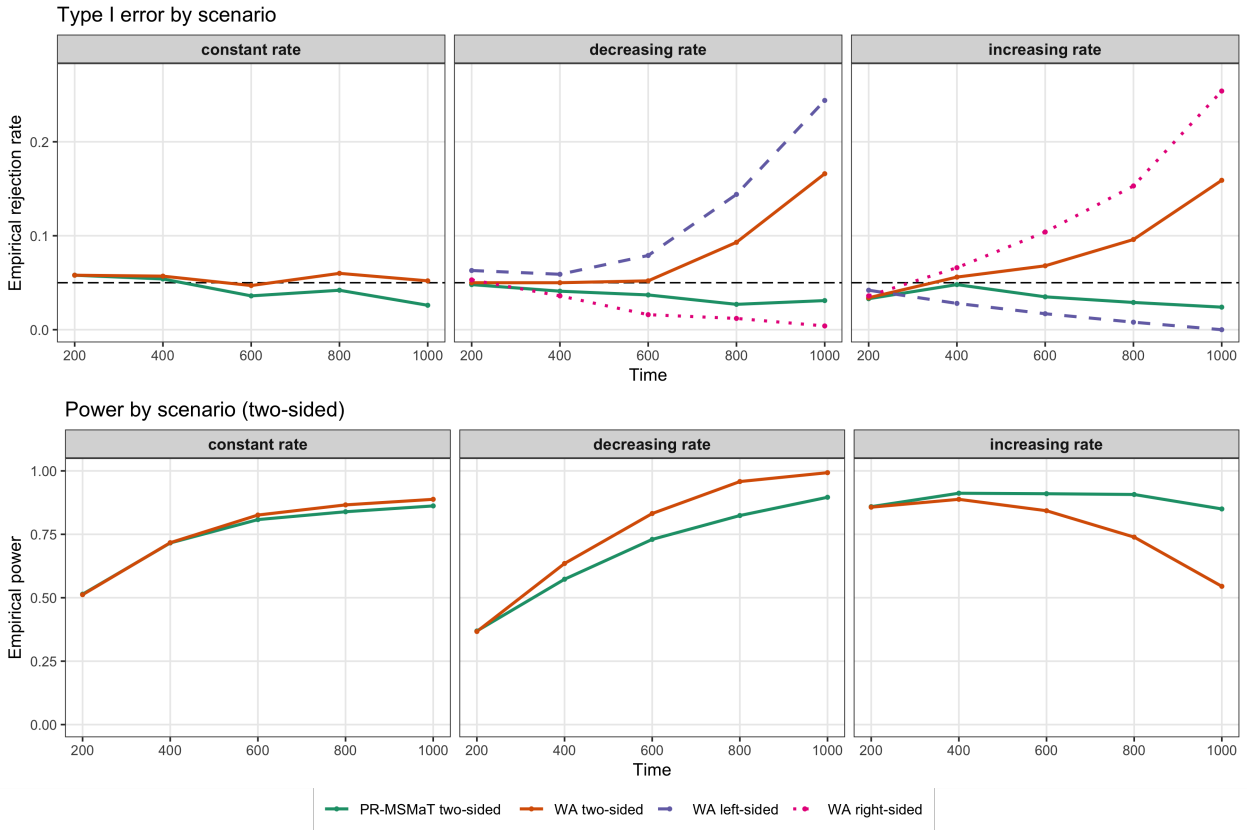


Figure 3.4: Empirical type I error rate and power across time under three scenarios designed in the simulation study: constant, decreasing, and increasing baseline recurrent-event rates. The dashed horizontal line marks the nominal significance level 0.05.

3.6 Application

We now analyze the data of the motivating example introduced in Section 4.1. Within the SEF, the question of interest is whether device type has a significant direct effect on recurrent gastrointestinal bleeding (GIB) when the survival pathway is held fixed. In this formulation, any difference in

recurrent event rates can be attributed to the device itself rather than to differences in survival. Our dataset contains adults (age $\geq$ 19 years) who underwent a primary continuous-flow LVAD implant, with or without subsequent RVAD (BiVAD), between April 1, 2006 and December 31, 2017. This study period largely precedes the widespread adoption of newer-generation devices, such as the HeartMate 3; therefore, patients receiving these newer devices are effectively excluded from the analysis. We excluded pulsatile devices, isolated RVAD, total artificial hearts, and non-primary LVAD implants. The primary outcome is a recurrent event GIB, defined as bleeding from the upper or lower GI tract or occult-positive events of unknown source. Death is treated as a terminal (or nuisance) event.

To emulate the randomized trial setting, we performed a 1:1 propensity-score matching using pre-implant demographic, clinical, laboratory, echocardiographic, and hemodynamic factors. Specifically, the variables included demographic factors (age, gender, race, and body mass index), indicators of disease severity and functional status (INTERMACS patient profile, NYHA class IV status, frailty, and primary ischemic diagnosis), vital signs (heart rate), echocardiographic measures (e.g., left ventricular and right ventricular function and dimensions), hemodynamic measurements (e.g., pulmonary pressures, cardiac index, and vascular resistance), laboratory values (e.g., hemoglobin, platelet count, renal function, and biomarkers), as well as comorbidities and preoperative conditions (e.g., diabetes, renal disease, liver dysfunction, and gastrointestinal ulcer). In addition, procedural characteristics such as concomitant surgery were incorporated. These variables were selected because they jointly characterize baseline health status and clinical decision-making factors that may influence both device selection (LVAD vs. BiVAD) and subsequent risks of recurrent gastrointestinal bleeding and death. Baseline characteristics were well balanced post-match (all p-values for univariate tests > 0.05). See Appendix C for the summary tables for baseline factors in the dataset after matching. After 1:1 matching, the analytic sample comprises $n = 1{,}051$ patients per group (BiVAD and LVAD).

We summarized recurrent gastrointestinal bleeding (GIB) events and overall survival in the matched cohort by device type (LVAD vs. BiVAD). Figure 3.5 displays two descriptive plots

for GIB burden as well as their follow-up time. In the survival comparison as seen clearly, the BiVAD group experienced markedly earlier mortality than the LVAD group. We also evaluated the treatment contrast between BiVAD and LVAD using the WA test and our PR-MSMaT at 5 time horizons $\tau \in \{1, 2, 5, 10, 15\}$ (months). Figure 3.6 reports the standardized Z values and p-values, where both methods use the contrast BiVAD − LVAD, implying that negative Z favors BiVAD.

It is noticeable that WA and PR-MSMaT in fact target different estimands, so they react differently to the survival imbalance in Figure 3.5 where people in the BiVAD group tend to die much earlier. WA divides the survival-completed cumulative frequency by the mean survival time, which reduces survivorship bias but does not fully remove it. If the event rate is time varying, especially higher soon after implant, earlier mortality in BiVAD means a larger fraction of its observed time lies in the high-risk window, which can inflate the while-alive rate and yield a positive Z, indicating more severe GIB burden in the BiVAD group, even without a higher direct propensity for events in that group. PR-MSMaT uses separable-effects weighting to align the survival pathway (A_D) across groups and tests the direct effect on recurrent events (A_Y) while holding the death process fixed. In Figure 3.6, the negative PR-MSMaT Z values at larger follow-up time (τ) indicate that, after fixing the survival pathway, BiVAD has a lower direct recurrent-event burden (fewer GIB events attributable to the direct effect) than LVAD. Thus, the sign disagreement (WA positive vs. PR-MSMaT negative) reflects an indirect survival pathway that over-weights high-risk early periods for BiVAD.

Hospitals participating in the Society of Thoracic Surgeons (STS) INTERMACS leverage standardized data forms to collect longitudinal data on this population. Use of data for this study was approved by the STS Participant User File Task Force. Patient consent to submit data to STS INTERMACS is not required by STS and is determined by local hospital policy. The opinions expressed in this manuscript do not represent those of STS.

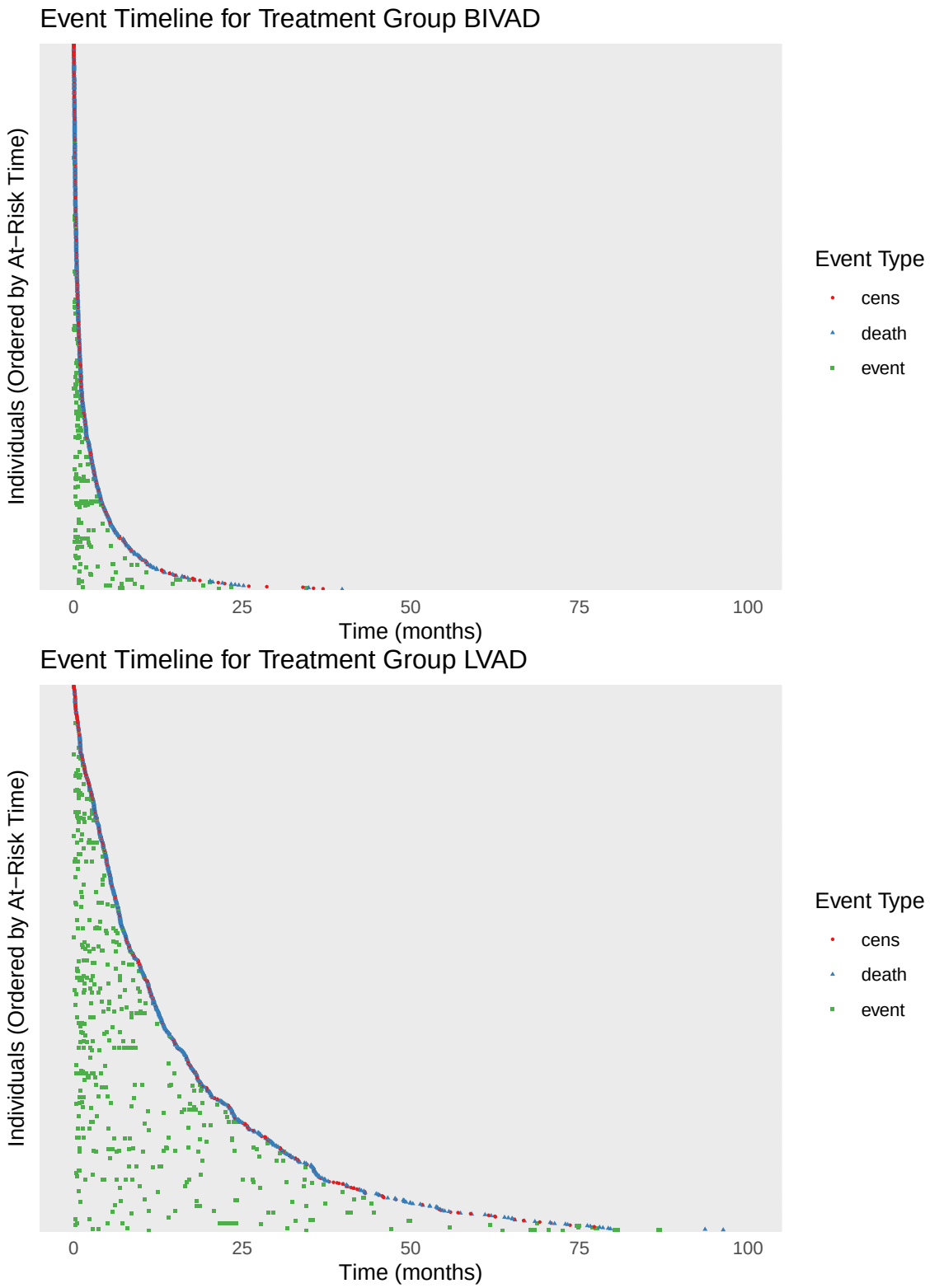


Figure 3.5: Plots of times for recurrent event of gastrointestinal bleeding (GIB) marked by green dots, death by blue dots appearing at the edge, and censoring by red dots among the matched pairs across device type (LVAD vs. BiVAD).

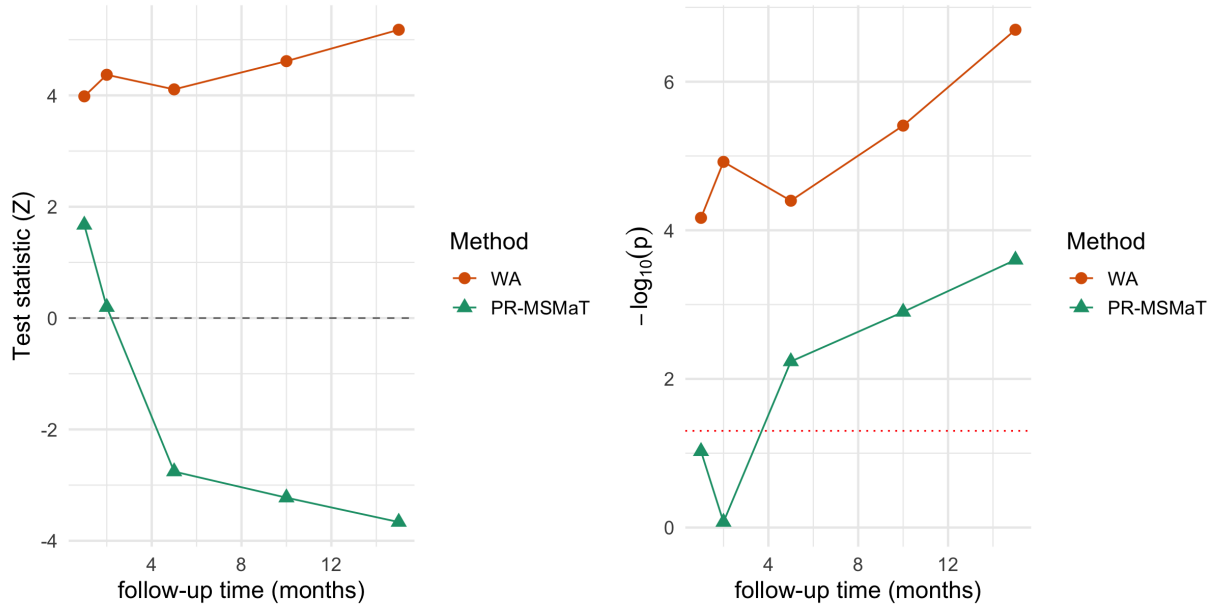


Figure 3.6: Comparison of test statistics (Z) and their $-\log_{10}(p)$ values at five selected follow-up times (months) across two test methods: existing WA (While-alive-based test) and new PR-MSMaT (Proportional Rate Marginal Structural Model–assisted Test). The horizontal dashed line in the left panel indicates $Z = 0$, and the red dotted line in the right panel corresponds to the conventional significance threshold $p = 0.05$.

3.7 Discussion

We developed an inference method, termed PR-MSMaT, a score-type test that is constructed under the separable effects framework (SEF) to evaluate the direct treatment effect on recurrent events while explicitly holding the survival pathway fixed. Conceptually, PR-MSMaT evaluates the separable effect estimands of Stensrud et al. (2022) and Janvin et al. (2023) by a practical score test analogous to a log-rank test or Cox score test. It targets a pathway-specific contrast rather than a mixture of mechanisms driven by both recurrent and terminal processes, aggregating information over the entire follow-up to yield a single, interpretable summary statistic with a well-defined causal estimand. In extensive simulations across diverse data-generating regimes, PR-MSMaT has been shown to maintain nominal type I error and demonstrated robustness to time-varying baseline hazards. Notably, it preserves validity in scenarios where the existing while-alive (WA) test exhibits systematic type I error inflation. In effect, the SEF effectively decouples recurrent-event dynamics from differential survival and reduces spurious associations induced by varying survival lengths.

In our empirical study, we observe that fixing the survival pathway enables us to isolate the component of recurrent gastrointestinal bleeding (GIB) burden attributable solely to the device choice itself, independent of mortality differences. Such finding is clinically relevant; it may help make clinical decisions on device selection to minimize bleeding risk, conditional on patients' expected survival. Furthermore, because the SEF provides causal interpretability under a well-executed 1:1 matching scheme, our comparison between two types of devices shows appealing insights useful to improve patients' post-implant quality of life. Technically, our method PR-MSMaT allows the time horizon τ and baseline rate function to vary, an important flexibility to analyze clinical data collected in other settings. These features position PR-MSMaT as a useful, well-justified analysis tool to evaluate treatments beyond implant devices in either randomized trials or registry-based observational studies.

Several directions warrant further development of PR-MSMaT. Extending the procedure to simultaneously accommodate multiple recurrent endpoints (e.g., bleeding, infection, and device

malfunction) and center-level heterogeneity would enhance its utility for multicenter registries and composite safety monitoring. Developing formal sensitivity analyses for potential violations of the dismissible component conditions would strengthen its application to observational data, where pathway isolation may be imperfect. Finally, incorporating covariate-adaptive weighting and sample-size formulas calibrated to pathway-specific alternatives could improve trial efficiency and power, particularly for registry-embedded or platform studies in which evolving practice patterns interact with event dynamics.

CHAPTER 4

Causal Inference for Recurrent Events in the Presence of Terminal Events with Covariates and Informative Censoring: An Extension of PR-MSMaT

4.1 Introduction

Heart transplantation remains the definitive therapy for end-stage heart failure but is severely limited by donor scarcity. Durable mechanical circulatory support (MCS) therefore plays a central role in clinical practice, with thousands of ventricular assist devices (VADs) implanted annually in the Interagency Registry for Mechanically Assisted Circulatory Support (INTERMACS) . Most patients receive left ventricular assist devices (LVADs), whereas those with substantial right ventricular dysfunction require biventricular devices (BiVADs). Although these therapies prolong survival, serious adverse events remain common. In particular, gastrointestinal bleeding (GIB) is a major contributor to morbidity, hospital readmission, and mortality. A clinically important question is whether the choice between BiVAD and LVAD implantation leads to differential risks of recurrent GIB events among patients with comparable baseline and evolving clinical profiles, while appropriately accounting for the fact that death terminates the observation of future bleeding events.

From a statistical perspective, this question involves comparing recurrent event processes in the presence of a terminal event under complex dependence induced by treatment and longitudinal

patient characteristics. Standard approaches that analyze recurrent events or survival separately are generally inadequate in this setting, as they fail to capture the interplay between event recurrence, survival, and evolving covariate processes. Several methodological challenges arise, even under ideal randomization. First, death acts as a competing terminal event that precludes future recurrent events, and naïve comparisons that ignore this mechanism may lead to biased inference. Second, time-varying risk factors, such as infections, may be influenced by device choice and subsequently affect both survival and GIB, thereby lying along pathways through which treatment impacts recurrent outcomes. Third, censoring may depend on treatment assignment, the evolving recurrent-event process, and baseline or time-varying covariates; in such settings, it cannot be assumed to be non-informative and must be appropriately modeled. Finally, even in randomized trials, imbalances in baseline risk factors may arise by chance, and standard guidance recommends covariate adjustment to improve both validity and efficiency (Vickers & Altman 2001, Kahan et al. 2014).

Among these challenges, appropriately handling the terminal event mechanism is particularly fundamental, as it directly affects both the definition and interpretation of recurrent event outcomes. Our previous work (Huang et al. 2026) addressed this aspect by developing a model-assisted hypothesis testing procedure within the separable effects framework (SEF), providing an alternative to the while-alive (WA)–based test. Historically, recurrent events have often been summarized using the mean cumulative function (MCF) (Ghosh & Lin 2000); however, such approaches implicitly disadvantage groups with longer survival, as more events can accumulate over a longer time at risk. The WA-based approach (Mao 2023) instead compares event accumulation per unit time alive, yielding a more interpretable estimand when survival differs between groups. Despite this advantage, the WA-based test may exhibit inflated type I error under the null when the baseline recurrent event rate varies over time. To address this limitation, Huang et al. (2026) introduced the Proportional-Rate Marginal Structural Model–assisted test (PR-MSMaT), a separable-effects–based score test that decomposes treatment into components acting on recurrent-event and terminal-event pathways and evaluates pathway-specific contrasts. By isolating these pathways, the procedure stabilizes

inference under time-varying baseline rates and restores valid type I error control.

However, Huang et al. (2026) focused on settings without incorporating baseline covariates, time-varying covariate processes, or informative censoring, and therefore does not directly extend to observational data settings such as INTERMACS, where these features are intrinsic.

The remaining challenges involve appropriately accounting for baseline covariates, longitudinal covariate processes affected by treatment, and informative censoring. A substantial literature has addressed related issues using marginal structural models (MSMs) with inverse probability weighting (IPW), which provide a general framework for handling time-dependent covariates and informative censoring. These methods have been extended to recurrent event settings (Jensen et al. 2016) and joint recurrent–terminal event settings (Gregorio et al. 2022), and have also been used to construct standardized nonparametric estimands under competing risks (Gaber et al. 2023).

Within the separable effects framework, Janvin et al. (2024) established identification results and corresponding weighting schemes that account for baseline covariates, time-varying covariate processes, and informative censoring. These developments provide a rigorous foundation for estimation under complex longitudinal data structures. However, they do not directly address hypothesis testing for pathway-specific effects, nor do they resolve the type I error inflation of while-alive–based tests under non-constant baseline recurrent rates identified in Huang et al. (2026).

In this paper, we extend PR-MSMaT by directly building on the weighting framework of Janvin et al. (2024) to account for baseline covariates, time-varying covariate processes, and informative censoring. Leveraging these identification results, we construct a model-assisted, score-type hypothesis test that preserves the separable-effects interpretation and the global, time-integrated structure of PR-MSMaT. The proposed procedure enables valid inference for pathway-specific treatment effects in observational registry settings such as INTERMACS under standard identification conditions.

The remainder of the chapter is organized as follows. Section 4.2 reviews the separable-effects framework and consolidates the weighting ingredients required for identification. Section 4.3 presents the proposed testing procedure. Section 4.4 reports simulation studies in settings with

baseline covariate, time-varying covariate process and censoring mechanisms. Section 4.5 applies the method to INTERMACS data. Section 4.6 concludes with discussion and future directions.

4.2 Separable Effect Framework (SEF)

We formalize the problem using a discrete-time longitudinal framework motivated by the INTERMACS registry. Let $A \in \{0, 1\}$ denote treatment, where $A = 1$ corresponds to implantation of a biventricular assist device (BiVAD) and $A = 0$ corresponds to a left ventricular assist device (LVAD). Let D_k denote the indicator of death by time t_k, and Y_k denote the cumulative number of recurrent gastrointestinal bleeding (GIB) events by t_k. Let L_0 denote a vector of baseline covariates measured at implantation, and let L_k denote time-varying covariates measured at time t_k, such as post-implant clinical states (e.g., infections) that evolve over time and influence both bleeding risk and survival. The follow-up window is $0 = t_0 < t_1 < \cdots < t_K = \tau$. For a generic stochastic process X_k, define the increment $\Delta X_k = X_k - X_{k-1}$, the history $\bar{X}_k = \{X_0, \ldots, X_k\}$, and the future $\underline{X}_k = \{X_k, \ldots, X_K\}$.

Following the separable effects framework, we conceptualize a hypothetical decomposition of treatment A into components (A_Y, A_D) that act along distinct causal pathways. The component A_Y operates on the recurrent-event pathway, while A_D operates on the survival pathway. Let $(Y_k^{a_Y,a_D}, D_k^{a_Y,a_D})$ denote the counterfactual outcomes under $(A_Y = a_Y, A_D = a_D)$. Our target is the pathway-specific mean number of recurrent events,

$$\mu_k^{a_Y,a_D} \;=\; E\{Y_k^{a_Y,a_D}\},$$

which represents the expected cumulative number of GIB events by time t_k under a hypothetical intervention that sets the recurrent-event pathway to a_Y while fixing the survival pathway to a_D. Separable effects are summarized by contrasts of $\mu_k^{a_Y=1,a_D}$ versus $\mu_k^{a_Y=0,a_D}$ for fixed a_D.

Let L_k denote time-varying covariates measured at time t_k that may be affected by treatment

and influence subsequent outcomes. We partition

$$L_k = \big(L_{Y,k},\, L_{D,k}\big),$$

where $L_{Y,k}$ contains variables whose causal effects enter the recurrent-event pathway and $L_{D,k}$ contains variables whose causal effects enter the survival pathway. This partition is required to define pathway-specific interventions under the separable effects framework: it ensures that the components (A_Y, A_D) can be interpreted as acting exclusively on the recurrent-event and survival pathways, respectively.

Formally, the partition must satisfy a causal separation condition: any causal path from A_Y to D_{k+1} or $L_{D,k+1}$ must pass through Y or $\bar{L}_{Y,k}$, and any causal path from A_D to Y_{k+1} or $L_{Y,k+1}$ must pass through D or $\bar{L}_{D,k}$ (Stensrud et al. 2021). This condition guarantees that the time-varying covariate process can be decomposed in a manner consistent with the separable decomposition of treatment, thereby enabling identification of pathway-specific counterfactual outcomes.

In our application, we specialize to $L_{Y,k} = \emptyset$ (for all k), so that $L_k = L_{D,k}$ (for all k). This specification reflects an asymmetric modeling assumption on how treatment components propagate through the longitudinal covariate process. In particular, time-varying covariates such as post-implant infections are allowed to transmit effects from the survival component A_D to the recurrent outcome Y_{k+1}, representing clinically plausible pathways through which device-related complications influence bleeding risk via overall deterioration or treatment modification.

In contrast, the recurrent-event component A_Y is assumed to act directly on the recurrent-event process and not to affect the terminal event through the longitudinal covariate process. That is, A_Y does not induce changes in L_k that subsequently influence D_{k+1}. Under this restriction, any causal path from A_Y to D_{k+1} must pass through the recurrent-event process Y, which is consistent with the separable-effects framework and supports the validity of the partition with $L_{Y,k} = \emptyset$.

Identification of $\mu_k^{a_Y,a_D}$ relies on the following standard assumptions (IA.1-IA.3) in causal inference, as well as the dismissible component assumptions (IA.4) required in SEF.

IA.1 Exchangeability:

$$\left(\bar{Y}_{K+1}^{a,\bar{c}=0},\, \bar{D}_{K+1}^{a,\bar{c}=0},\, \bar{L}_{K+1}^{a,\bar{c}=0}\right) \perp A \mid L_0,$$

and, for each $k = 0, \ldots, K$,

$$\left(\underline{Y}_{k+1}^{a,c=0},\, \underline{D}_{k+1}^{a,\bar{c}=0},\, \underline{L}_{k+1}^{a,\bar{c}=0}\right) \perp C_{k+1}^{a,\bar{c}=0} \mid \bar{Y}_k^{a,\bar{c}=0},\, \bar{D}_k^{a,\bar{c}=0},\, \bar{C}_k^{a,\bar{c}=0},\, \bar{L}_k^{a,\bar{c}=0},\, A.$$

IA.2 Positivity:

$$P(L_0 = l_0) > 0 \implies P(A = a \mid L_0 = l_0) > 0,$$

$$f_{\bar{L}_k,\bar{D}_{k+1},\bar{C}_{k+1},\bar{Y}_k}(\bar{l}_k, 0, 0, \bar{y}_k) > 0 \implies P(A = a \mid \bar{D}_{k+1} = 0,\, \bar{C}_{k+1} = 0,\, \bar{Y}_k = \bar{y}_k,\, \bar{L}_k = \bar{l}_k) > 0,$$

$$f_{A,\bar{L}_k,\bar{D}_k,\bar{C}_k,\bar{Y}_k}(a, \bar{l}_k, 0, 0, \bar{y}_k) > 0 \implies P(C_{k+1} = 0 \mid \bar{L}_k = \bar{l}_k,\, \bar{D}_k = 0,\, \bar{C}_k = 0,\, \bar{Y}_k = \bar{y}_k) > 0,$$

for all relevant a, k, and histories.

IA.3 Consistency: If $A = a$ and $\bar{C}_{k+1} = 0$, then

$$\bar{L}_{k+1} = \bar{L}_{k+1}^{a,\bar{c}=0}, \qquad \bar{D}_{k+1} = \bar{D}_{k+1}^{a,\bar{c}=0}, \qquad \bar{Y}^{k+1} = \bar{Y}^{a,\bar{c}=0,k+1}.$$

IA.4 Dismissible component conditions: For each $k = 0, \ldots, K$,

$$Y_{k+1}^{\bar{c}=0} \perp A_D \mid A_Y,\, \bar{D}_{k+1}^{\bar{c}=0},\, \bar{Y}_k^{\bar{c}=0},\, \bar{L}_k^{\bar{c}=0},$$

$$D_{k+1}^{\bar{c}=0} \perp A_Y \mid A_D,\, \bar{D}_k^{\bar{c}=0},\, \bar{Y}_k^{\bar{c}=0},\, \bar{L}_k^{\bar{c}=0},$$

$$L_{Y,k}^{\bar{c}=0} \perp A_D \mid A_Y,\, \bar{Y}_k^{\bar{c}=0},\, \bar{D}_k^{\bar{c}=0},\, \bar{L}_{k-1}^{\bar{c}=0},\, L_{D,k}^{\bar{c}=0},$$

$$L_{D,k}^{\bar{c}=0} \perp A_Y \mid A_D,\, \bar{D}_k^{\bar{c}=0},\, \bar{Y}_k^{\bar{c}=0},\, \bar{L}_{k-1}^{\bar{c}=0}.$$

The conditions (IA.4) formalize the separable effects assumption that the treatment A can be decomposed into components (A_Y, A_D) acting exclusively on the recurrent-event and survival pathways, respectively. In particular, they require that A_D does not directly affect the recurrent

outcome beyond what is mediated through the survival process and the covariate history, and that A_Y does not directly affect the terminal event beyond pathways through the recurrent-event process. The latter two conditions ensure that the longitudinal covariate process can be partitioned consistently with this decomposition.

Under the restriction $L_{Y,k} = \emptyset$, the third condition in (IA.4) is vacuous.

Under assumptions (IA.1)–(IA.4), the counterfactual mean $\mu_k^{a_Y,a_D} = \mathbb{E}[Y_k^{a_Y,a_D}]$ is identified and admits the inverse probability weighting representation

$$\mu_k^{a_Y,a_D} = \sum_{t=1}^{k} E\left\{I(A = a_Y)\, W_t(a_Y, a_D)\, \Delta Y_t\right\},$$

where $W_t(a_Y, a_D)$ reweights the observed data to emulate the intervention $(A_Y = a_Y, A_D = a_D)$.

$W_k(a_Y, a_D)$ is constructed as a product of stabilized components corresponding to baseline treatment assignment, the time-varying covariate process, the terminal event process, and censoring:

$$W_k(a_Y, a_D) = W_A \cdot W_{L,k}^{a_Y,a_D} \cdot W_{D,k}^{a_Y,a_D} \cdot W_{C,k},$$

where

$$W_A \;=\; \frac{I(A = a_Y)}{P(A \mid L_0)},$$

$$W_{L,k}^{a_Y,a_D} \;=\; \prod_{j=1}^{k} \frac{P\big(L_{D,j} \mid A = a_D,\, L_0, \bar{L}_{j-1}, \bar{Y}_{j-1}, \bar{D}_{j-1}, \bar{C}_j\big)}{P\big(L_{D,j} \mid A = a_Y,\, L_0, \bar{L}_{j-1}, \bar{Y}_{j-1}, \bar{D}_{j-1}, \bar{C}_j\big)},$$

$$W_{D,k}^{a_Y,a_D} \;=\; \prod_{j=1}^{k} \frac{P\big(D_j \mid A = a_D,\, L_0, \bar{L}_j, \bar{Y}_{j-1}, \bar{D}_{j-1}, \bar{C}_j\big)}{P\big(D_j \mid A = a_Y,\, L_0, \bar{L}_j, \bar{Y}_{j-1}, \bar{D}_{j-1}, \bar{C}_j\big)},$$

$$W_{C,k} \;=\; \prod_{j=1}^{k} \frac{I(C_j = 0)}{P\big(C_j \mid L_0, \bar{L}_{j-1}, \bar{Y}_{j-1}, \bar{D}_{j-1}, \bar{C}_{j-1}\big)}.$$

The weight $W_k(a_Y, a_D)$ transforms the observed data under the two-arm treatment A into a

pseudo-population corresponding to the separable intervention (a_Y, a_D).

Specifically, W_A standardizes baseline covariate distributions across treatment groups. The components $W_{D,k}^{a_Y,a_D}$ and $W_{L,k}^{a_Y,a_D}$ take effect only when $a_Y \neq a_D$ and follow a common principle: they adjust for differences in the conditional distributions of the terminal event process and the time-varying covariate process under the observed treatment levels $A = a_Y$ and $A = a_D$. The weight $W_{D,k}^{a_Y,a_D}$ uses the ratio of conditional probabilities of the terminal event to reshape the risk set under $A = a_Y$ so that the survival experience aligns with that under $A = a_D$, while $W_{L,k}^{a_Y,a_D}$ analogously calibrates the distribution of the longitudinal covariate process under $A = a_Y$ to match that under $A = a_D$. This reweighting is driven by ratios of conditional distributions rather than the magnitude of individual-level differences, thereby removing distortions arising from discrepancies in the distributions of terminal event and time-varying covariate process across treatment levels and ensuring a common risk set for evaluating recurrent event processes.

Finally, $W_{C,k}$ adjusts for informative censoring by reweighting individuals who remain under observation to represent those who would have remained uncensored.

Taken together, these weights reconfigure the observed data so that recurrent event contributions are evaluated relative to a_Y, while the associated survival and covariate processes are aligned with those under a_D. This construction mirrors the separable effects framework and forms the basis of the proposed extended PR-MSMaT test statistic.

Importantly, the role of these weighting components is not merely to refine the representation of the observed data, but to ensure that the resulting inference targets the intended separable effect under a properly aligned pseudo-population. In particular, the baseline weight addresses residual imbalances that may arise by chance even after randomization or matching, while the remaining components align the survival and time-varying processes across treatment levels so that comparisons of recurrent events are not driven by differences in follow-up time or evolving covariate trajectories. This alignment is essential for valid hypothesis testing, as failure to properly account for these features can lead to biased test statistics and inflated Type I error, rather than simply reduced or elevated efficiency.

The separable mean function estimator is

$$\widehat{\mu}_{a_Y,a_D}(t) = \sum_{k=1}^{K} \frac{\sum_i I(A_i = a_Y)\, Z_{i,k}\, \widehat{S}_{a_Y}(t_k)\, \widehat{W}_i(a_Y, a_D)\, \Delta Y_{i,k}}{\sum_i I(A_i = a_Y)\, Z_{i,k}},$$

where $Z_{i,k}$ is the at-risk indicator, $\widehat{S}_{a_Y}(t_k)$ denotes the estimated survival probability under $A = a_Y$, and $\hat{W}_i(a_Y, a_D)$ is the consistent estimator of $W_i(a_Y, a_D)$. See Janvin et al. (2024) and Ryalen et al. (2018) for details of the estimaton methods.

4.3 Extended PR-MSMaT (ext-PR-MSMaT)

We build on the separable-effects estimators and the stabilized weighting framework summarized in Section 4.2 to construct a score-type test assisted by a proportional-rate marginal structural model (PR-MSM). The weighting components are treated as nuisance functions estimated from the observed data, following Janvin et al. (2024). Our contribution is to extend the PR-MSMaT proposed in Huang et al. (2026) to incorporate these weights within the hypothesis testing framework.

Following Huang et al. (2026), we consider the proportional rate marginal structural model

$$\mu^{a_Y,a_D}(t) = \mu^{a_Y=0,a_D}(t) \exp\left(\beta^{a_D} a_Y\right), \quad a_D \in \{0, 1\}, \tag{4.1}$$

which posits that, for each fixed a_D, the mean functions across a_Y differ by a constant multiplicative factor over time.

Using the same derivation as in Huang et al. (2026), under the null hypothesis $H_0 : \beta^{a_D} = 0$, the score function reduces to

$$U_n \;=\; \int_0^{\tau} \sum_{i=1}^{n} Z_i(t) A_i \left\{ d\widehat{\mu}^{a_Y=1,a_D}(t) - d\widehat{\mu}^{a_Y=0,a_D}(t) \right\}, \tag{4.2}$$

where $\widehat{\mu}^{a_Y,a_D}(t)$ is the weighted separable mean estimator defined in Section 4.2. The corresponding

test statistic is

$$T_n \;=\; \frac{U_n}{\widehat{\mathrm{Var}}(U_n)^{1/2}}.$$

The variance estimator $\widehat{\mathrm{Var}}(U_n)$ follows the same martingale-based construction as in Huang et al. (2026). The asymptotic result in Theorem 1 of Huang et al. (2026) continues to hold provided that the estimated weights $\widehat{W}_i(a_Y, a_D)$ are consistent for $W_i(a_Y, a_D)$, which is satisfied under standard regularity conditions for the estimation of the nuisance functions within the transforming cumulative hazard framework of Ryalen et al. (2018). Consequently, under $H_0 : \beta^{a_D} = 0$,

$$T_n \;\xrightarrow{d}\; N(0,1), \quad \text{as } n \to \infty.$$

The proposed testing procedure extends the original PR-MSMaT by incorporating adjustments for baseline covariates, time-varying covariate processes, and informative censoring through stabilized weights. While the original test is developed under idealized settings where these factors are negligible, the extended version enables valid inference in observational settings such as the INTERMACS program in our motivating example. Importantly, the test retains the same score-type structure and interpretation: it provides a global assessment of the separable direct effect of treatment on recurrent events over the follow-up period, analogous to the log-rank or Cox score test in classical survival analysis.

4.4 Simulation Study

We conducted simulation experiments to evaluate (i) type I error control of the proposed testing procedures under the null hypothesis of no direct treatment effect on recurrent events and (ii) statistical power under alternative scenarios. The proposed ext-PR-MSMaT was compared with the original PR-MSMaT of Huang et al. (2026).

The simulation setting is motivated by the recurrent event framework described in Section 4.1, and is constructed to capture key features relevant to the proposed methodology, including recurrent

events, a terminal event, time-varying covariates, and informative censoring. In this setting, subjects may experience recurrent events over time, while death acts as a terminal event that precludes further recurrences. A time-varying covariate is incorporated to reflect longitudinal processes that may be associated with treatment, the recurrent and terminal event processes. In addition, baseline covariates are included to allow for heterogeneity in both event types. The baseline recurrent event rate is specified to vary over time in order to evaluate performance in the setting that originally motivated the development of PR-MSMaT, namely when non-constant recurrent event rates can compromise the validity of while-alive-based methods, including through type I error inflation. These components together define a data-generating mechanism that reflects the structural complexities addressed by the proposed method.

Formally, we considered a discrete-time grid $0 = t_0 < t_1 < \cdots < t_K = \tau$ with $K = 1000$ intervals, providing a fine approximation to continuous-time processes. For subject i and interval $(t_{j-1}, t_j]$, we generated $\Delta Y_{i,j} \in \{0, 1\}$ (recurrent event indicator), $D_{i,j} \in \{0, 1\}$ (death indicator), $C_{i,j} \in \{0, 1\}$ (censoring indicator), the time-varying covariate $L_{D,i,j} \in \{0, 1\}$, and baseline covariates $L_{0,i}$ consisting of two continuous biomarkers and one Bernoulli comorbidity indicator.

The data-generating mechanism was constructed to reflect the causal structure described in Section 4.2, where the time-varying covariate primarily operates through the survival pathway instead of recurrent-event pathway, i.e., $L_{i,j} = L_{D,i,j}$ and $L_{Y,i,j} = \emptyset$ for all j. We used additive increment models on the probability scale for the recurrent event, terminal event, and censoring processes, truncated to $[0, 1]$ via $[x]_{[0,1]} := \min\{\max(x, 0), 1\}$. A logistic transition model was used to generate the time-varying covariate process. Let $\lambda_{Y0}(t)$, $\lambda_{D0}(t)$, and $\lambda_{C0}(t)$ denote baseline

hazard functions. For $j = 1, \ldots, K$, we specified:

$$
\begin{aligned}
P(\Delta Y_{i,j} = 1 \mid \cdot) &= \Big[\lambda_{Y0}(t_j) + \beta_Y A_{Y,i} + \gamma_Y Y_{i,j-1} + \gamma_{L_0}^\top L_{0,i} + \gamma_{L_D} L_{D,i,j}\Big]_{[0,1]}, \\
P(D_{i,j} = 1 \mid \cdot) &= \Big[\lambda_{D0}(t_j) + \beta_D A_{D,i} + \eta_Y Y_{i,j-1} + \eta_{L_0}^\top L_{0,i} + \eta_{L_D} L_{D,i,j}\Big]_{[0,1]}, \\
P(L_{D,i,j} = 1 \mid \cdot) &= \text{logit}^{-1}\Big(\alpha_0 + \alpha_{L_D} L_{D,i,j-1} + \alpha_Y Y_{i,j} + \alpha_{A_D} A_{D,i}\Big), \\
P(C_{i,j} = 1 \mid \cdot) &= \Big[\lambda_{C0}(t_j) + \delta_{A_Y} A_{Y,i} + \delta_{A_D} A_{D,i} + \delta_{L_0}^\top L_{0,i} + \delta_{LD} L_{D,i,j-1} + \delta_Y \Delta Y_{i,j-1}\Big]_{[0,1]}.
\end{aligned}
$$

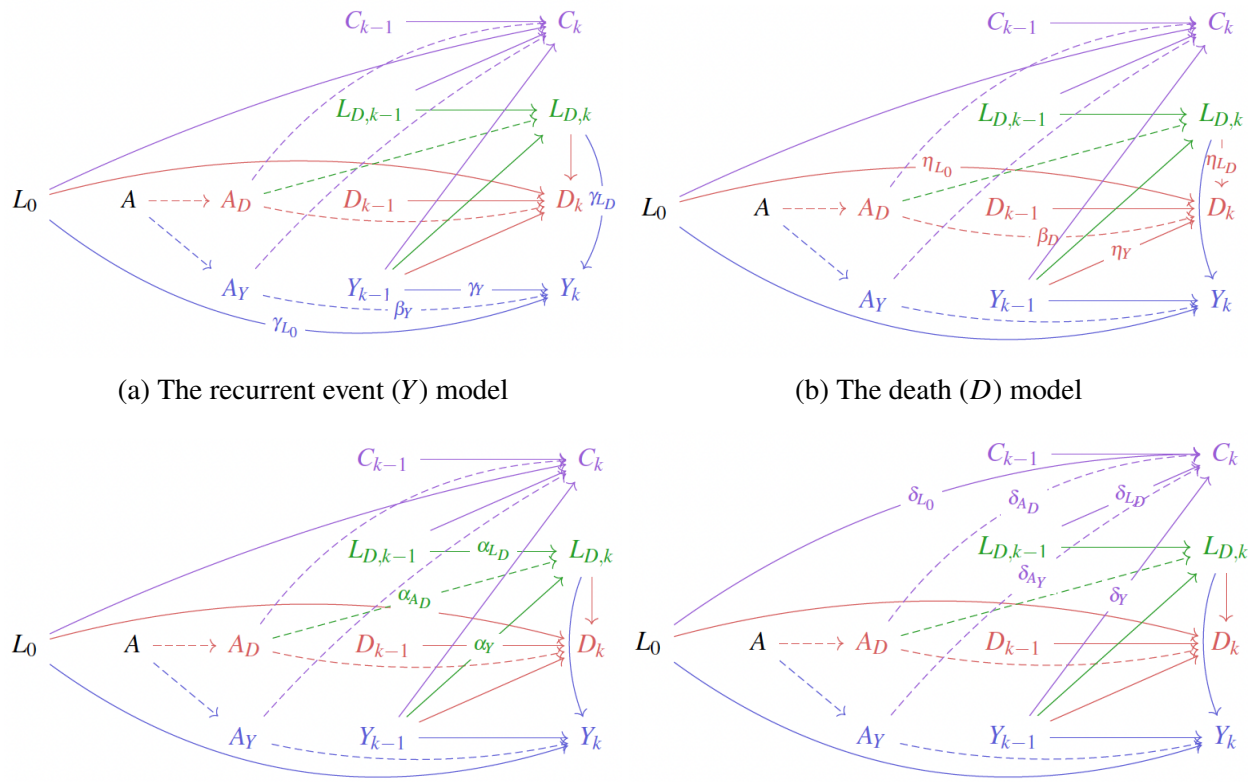
This specification incorporates several key features relevant to the methodological setting. Both recurrent events and the terminal event depend on baseline covariates and prior event history, inducing dependence over time. The time-varying covariate evolves dynamically and is affected by prior outcomes and treatment, thereby introducing time-varying confounding. In addition, the censoring mechanism depends on both baseline and time-varying variables, resulting in informative censoring.

To explore scenarios where event rates vary over time, we specified an *increasing* baseline recurrent event rate by setting $\lambda_{Y0}(t_1) = 1/K$ and increasing it by $0.5/K$ every 200 intervals. The baseline death rate was fixed at $\lambda_{D0}(t) = 0.8/K$, and the baseline censoring rate was set to $\lambda_{C0}(t) = 0.1/K$ to yield moderate censoring.

Figures 4.1a–4.1d depict the conditional dependencies used to simulate Y, D, L_D, and C, respectively, and align with the notation above.

We begin with a benchmark scenario without time-varying covariate or censoring effects, and then introduce additional complexities by varying selected parameters. Across all scenarios, we fix the treatment effect on death $\beta_D = -0.3$ and evaluate both type I error (when treatment effect on recurrent event $\beta_Y = 0$) and power (when $\beta_Y = -0.3$) at two sample sizes, $n \in \{500, 1000\}$. Each configuration uses 1000 Monte Carlo replicates.

(1) Benchmark scenario We remove all baseline covariate effects ($(\eta_{L_D}, \gamma_{L_D}) = 0$), remove all L_D-related effects ($(\alpha_{A_D}, \alpha_Y, \eta_{L_D}, \gamma_{L_D}) = 0$), remove censoring-related effects



(a) The recurrent event (Y) model

(b) The death (D) model

(c) The time-varying covariate (L_D) model

(d) The censoring (C) model

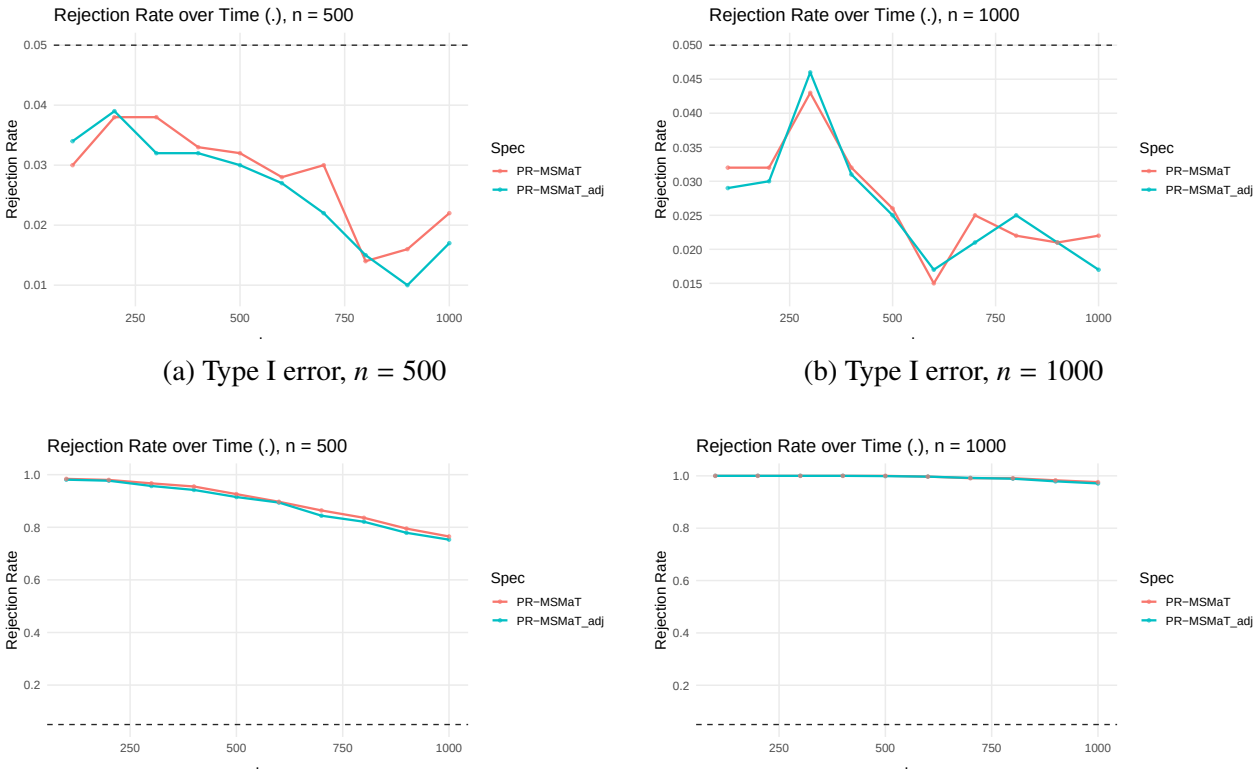
Figure 4.1: **Data-generating Models.** Panels: (a) recurrent events Y_k, (b) deaths D_k, (c) time-varying covariate $L_{D,k}$, and (d) censoring C_k. Other components: treatment A, hypothetical treatment components A_Y and A_D, baseline confounders L_0.

$((\delta_{A_Y}, \delta_{A_D}, \delta_Y, \delta_{L_0}, \delta_{L_D}) = 0)$, and set $Y \perp D$ $(\eta_Y = 0)$.

Figure 4.2 shows that the two methods are *virtually indistinguishable* in both type I error (Figs. 4.2a, 4.2b) and power (Figs. 4.2c, 4.2d). This confirms that ext-PR-MSMaT does not incur any efficiency loss when adjustment is unnecessary.

(2) Time-varying covariate and censoring scenario The L_D-related effects $(\alpha_{A_D}, \alpha_Y, \eta_{L_D}, \gamma_{L_D})$ have magnitude 0.5, the censoring-related effects $(\delta_{A_Y}, \delta_{A_D}, \delta_Y, \delta_{L_0}, \delta_{L_D})$ have magnitude 0.1, and $Y \perp D$ $(\eta_Y = 0)$.

Figure 4.3 shows that **ext-PR-MSMaT** maintains type I error close to the nominal 0.05 level across τ, whereas the **PR-MSMaT** exhibits noticeable type I error inflation. This pattern is consistent for both $n = 500$ and $n = 1000$ (Figs. 4.3a and 4.3b).



(a) Type I error, $n = 500$

(b) Type I error, $n = 1000$

(c) Power, $n = 500$

(d) Power, $n = 1000$

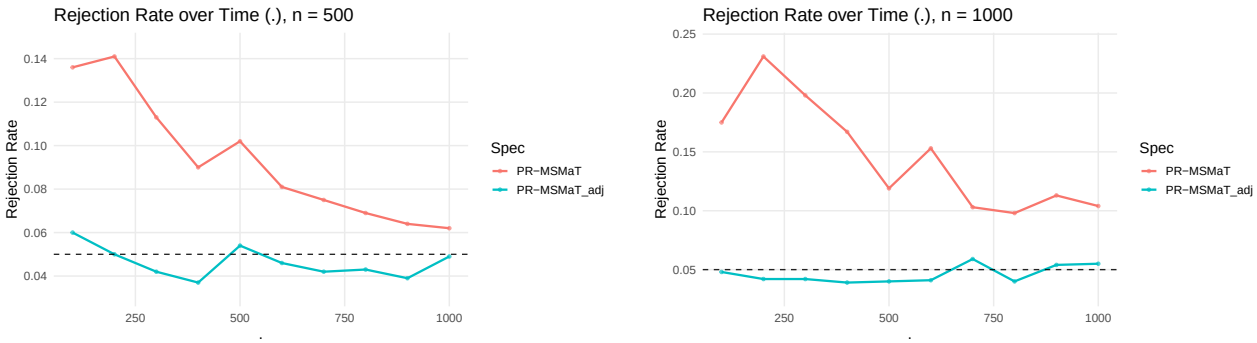
Figure 4.2: **Scenario (1)** No covariates, no censoring effects, and $Y \perp D$. Panels show Type I error and power at $n = 500, 1000$.

For power, **ext-PR-MSMaT** uniformly outperforms **PR-MSMaT** across the range of τ, with gains that become more pronounced as follow-up time increases (Figs. 4.3c and 4.3d).

(3) Recurrent events affecting death All parameters are as in scenario (2), except that recurrent events affect death ($\eta_Y = 0.5$).

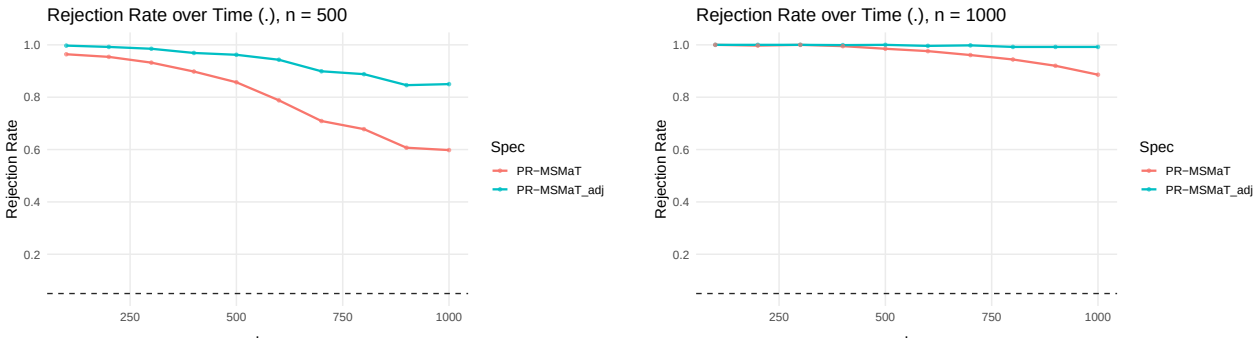
Figure 4.4 shows that **ext-PR-MSMaT** continues to control type I error appropriately. At $n = 500$ (Fig. 4.4a), the test is slightly conservative, while at $n = 1000$ (Fig. 4.4b) it is close to the nominal level.

In terms of power, **ext-PR-MSMaT** consistently dominates **PR-MSMaT** at both sample sizes (Figs. 4.4c and 4.4d), indicating robustness to additional dependence between recurrent and terminal events.



(a) Type I error, $n = 500$

(b) Type I error, $n = 1000$



(c) Power, $n = 500$

(d) Power, $n = 1000$

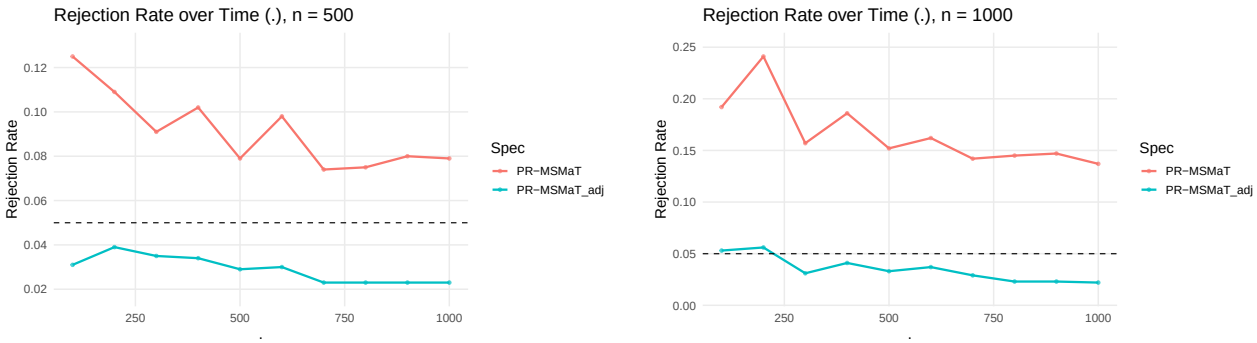
Figure 4.3: **Scenario (2)** L_D-related effects $(\alpha_{A_D}, \alpha_Y, \eta_{L_D}, \gamma_{L_D}) = 0.5$, censoring-related effects $(\delta_{A_Y}, \delta_{A_D}, \delta_Y, \delta_{L_0}, \delta_{L_D}) = 0.1$, and $Y \perp D$ $(\eta_Y = 0)$. Panels show Type I error and power at $n = 500, 1000$.

(4) Strong censoring scenario We increase censoring effects to 0.3 for $(\delta_{A_Y}, \delta_{A_D}, \delta_Y, \delta_{L_0}, \delta_{L_D})$, leading to heavy censoring (exceeding 70% in the treatment arm by design).

Figure 4.5 shows that **ext-PR-MSMaT** exhibits type I error inflation for larger τ (approximately $\tau \geq 700$), likely due to extreme censoring leading to instability in the weighting procedure and practical violations of positivity. However, for earlier follow-up times ($\tau < 700$), **ext-PR-MSMaT** maintains appropriate type I error control, whereas **PR-MSMaT** already shows inflation.

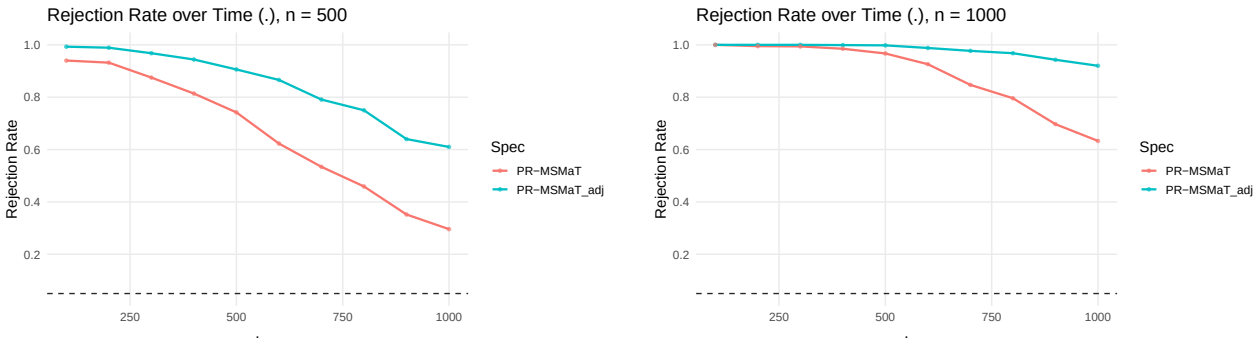
For power, **ext-PR-MSMaT** generally outperforms **PR-MSMaT** up to moderate τ (approximately $\tau \leq 700$), after which performance becomes more variable due to the increasing influence of censoring (Figs. 4.5c and 4.5d).

Overall, the results demonstrate three key findings. First, in the absence of time-varying covariate and censoring effects, ext-PR-MSMaT performs comparably to PR-MSMaT, indicating



(a) Type I error, $n = 500$

(b) Type I error, $n = 1000$



(c) Power, $n = 500$

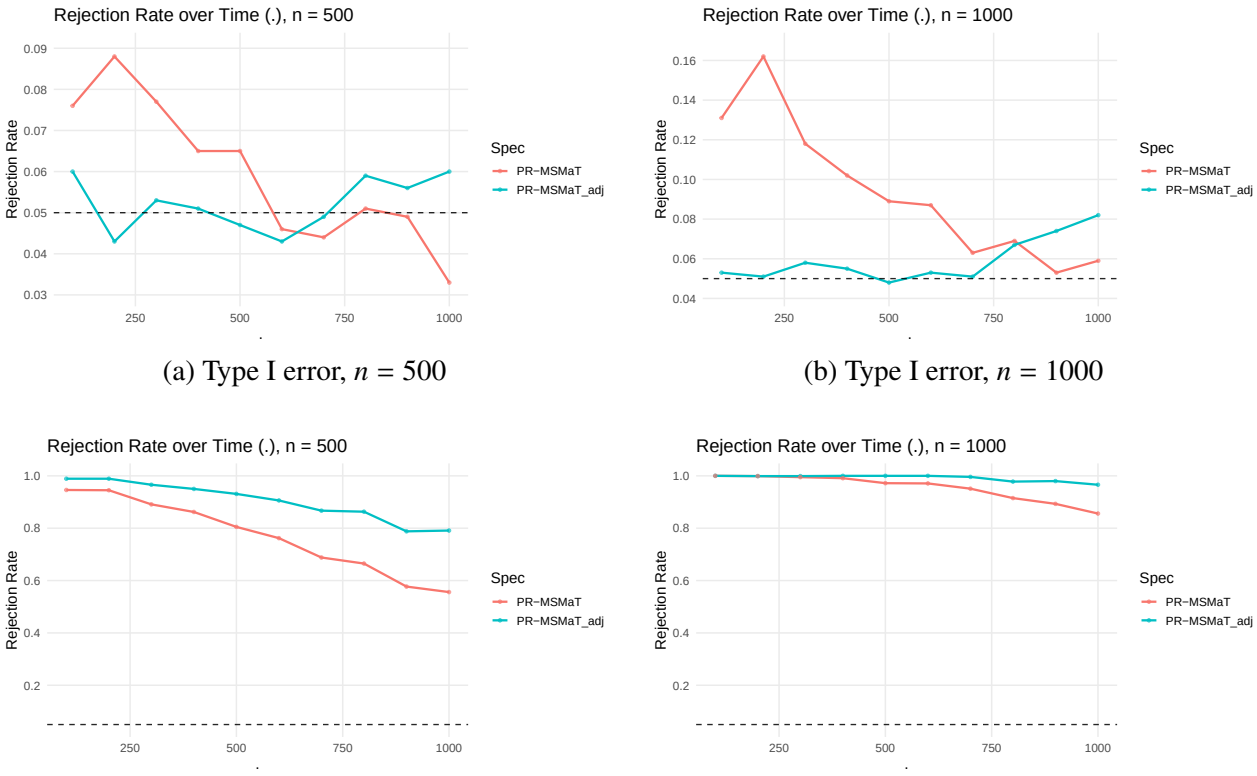
(d) Power, $n = 1000$

Figure 4.4: **Scenario (3)** L_D-related and censoring-related effects are the same as scenario (2), while $\eta_Y = 0.5$ (recurrent events affect death). Panels show Type I error and power at $n = 500, 1000$.

no efficiency loss from incorporating adjustment. Second, when the longitudinal covariate process and informative censoring jointly influence the outcome processes, ext-PR-MSMaT achieves both valid type I error control and improved power, whereas PR-MSMaT can exhibit substantial type I error inflation. Third, under extreme censoring, ext-PR-MSMaT remains preferable at earlier follow-up times, although its performance may deteriorate at later times due to instability in the weighting procedure. Together, these results highlight the practical advantages of the proposed method in settings with longitudinal covariate processes and informative censoring.

4.5 Application

We applied the proposed method to the INTERMACS registry to assess whether the choice of Bi-VAD and LVAD differ in the direct risk of recurrent gastrointestinal bleeding (GIB) after accounting



(a) Type I error, $n = 500$

(b) Type I error, $n = 1000$

(c) Power, $n = 500$

(d) Power, $n = 1000$

Figure 4.5: **Scenario (4)** L_D-related effects are same as scenario (2), censoring-related effects increased to 0.3; $Y \perp D$. Panels show Type I error and power at $n = 500, 1000$.

for baseline characteristics and the time-varying infection process, informative censoring, as well as acknowledging the fact that deaths terminate the GIB process. To facilitate a fair comparison with the original PR-MSMaT in Huang et al. (2026), we used the same analytic cohort constructed via propensity score matching (PSM), yielding $n = 2102$ patients with balanced baseline covariates across treatment groups. While the application of the previous method relied on matched pairs, the current analysis further incorporates the stabilized weighting framework to account for longitudinal covariate evolution and informative censoring.

In the registry data, infection is recorded as a time point. To represent its effect as a time-varying process, we constructed an infection window following each recorded infection event. Specifically, Each infection episode was represented using a short time window following its recorded onset. Specifically, we generated a duration from an exponential distribution with mean 0.25 months

(approximately one week), during which the infection indicator remains active. This specification reflects the clinical understanding that infections in MCS patients typically have transient effects over days to weeks rather than acting as instantaneous events (Hannan et al. 2011, Nienaber et al. 2013).

Figure 4.6 summarizes the standardized test statistics and corresponding $-\log_{10}(p)$ values across follow-up horizons for the extended Proportional Rate Marginal Structural Model-assisted Test (ext-PR-MSMaT) and the original Proportional Rate Marginal Structural Model-assisted Test (PR-MSMaT).

The results show a clear divergence between the two methods over time. The extended Proportional Rate Marginal Structural Model-assisted Test (ext-PR-MSMaT) produces consistently negative test statistics across follow-up, indicating a lower estimated recurrent GIB burden for Bi-VAD relative to LVAD. In contrast, the original PR-MSMaT exhibits a markedly different pattern: the test statistics are close to zero in early follow-up and only begin to deviate from the null after approximately 5 months, suggesting delayed and less stable evidence of a treatment difference.

This contrast is also reflected in the significance scale. The ext-PR-MSMaT demonstrates clear statistical significance at earlier follow-up times, with $-\log_{10}(p)$ values indicating strong evidence initially. Although the strength of evidence attenuates somewhat at later stages, the results remain statistically significant even up to 25 months. By comparison, the original PR-MSMaT shows little evidence of significance in the early period, with statistical significance emerging only after around 5 months and remaining less consistent overall.

Taken together, these results suggest that accounting for the longitudinal covariate process and informative censoring yields more stable and interpretable inference on recurrent event outcomes, particularly when survival profiles differ across treatment groups. By incorporating these adjustments, the ext-PR-MSMaT provides earlier and more confirmatory evidence of a treatment effect, while maintaining statistical significance over an extended follow-up period. In contrast, the original PR-MSMaT produces delayed and less consistent signals, which may be sensitive to differences in survival and covariate evolution. These findings underscore the importance of properly address-

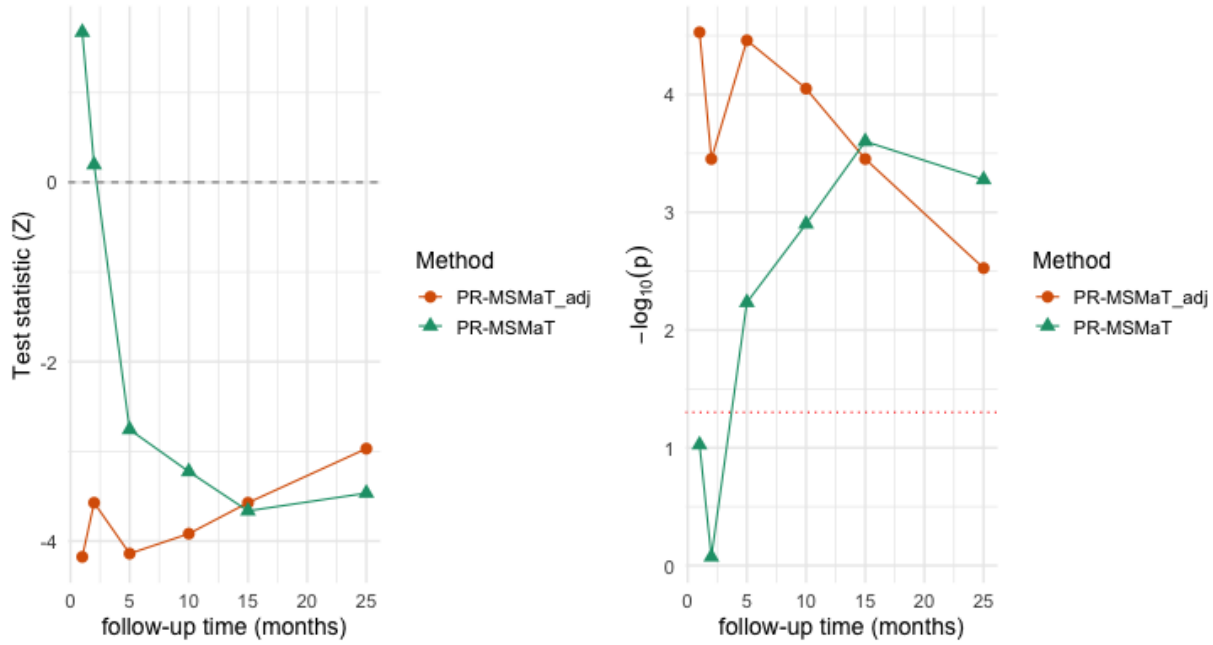


Figure 4.6: Comparison of test statistics (Z) and their $-\log_{10}(p)$ values at five selected follow-up times (months) across two test methods: our previous method PR-MSMaT (Proportional Rate Marginal Structural Model–assisted Test) and our proposed new method ext-PR-MSMaT (extended PR-MSMaT). The horizontal dashed line in the left panel indicates $Z = 0$, and the red dotted line in the right panel corresponds to the conventional significance threshold $p = 0.05$.

ing time-varying processes and differential survival when evaluating treatment effects on recurrent events.

4.6 Discussion

This paper is motivated by the INTERMACS registry, where treatment assignment, longitudinal clinical processes (e.g., infection), and loss to follow-up are intrinsically intertwined with both recurrent complications such as GIB and the terminal event such as death. In such practical settings, a central challenge is constructing a valid testing procedure that isolates pathway-specific treatment effects on recurrent events while preserving a risk set aligned with the target separable intervention. The proposed extension of PR-MSMaT addresses this gap by embedding separable effects framework within a score-type testing framework, allowing recurrent event comparisons to

be conducted under a common, intervention-aligned survival and covariate regime.

A key contribution of this work is conceptual rather than purely technical: the weighting construction is used not only for bias correction, but to redefine the target population in which the test is performed. In particular, the weights jointly align the distribution of baseline covariates, the longitudinal covariate process, and the survival process across pathway-specific hypothetical treatment levels, thereby removing distortions that arise when groups differ in follow-up time or evolving risk profiles. This perspective also clarifies why naïve or partially adjusted approaches can produce misleading inferences by comparing outcomes under incompatible risk sets.

The simulation results illustrate that the impact of these distortions can be substantial. When time-varying confounding and informative censoring are present, the original PR-MSMaT may exhibit type I error inflation because comparisons are made under differing covariate histories and survival processes across treatment groups, rather than under a common intervention-aligned regime. In contrast, the extended PR-MSMaT maintains nominal control by explicitly accounting for these mechanisms, without sacrifice efficiency in settings where the original method works well, indicating that the additional structure is only active when needed. The INTERMACS analysis further highlights that these adjustments are not merely theoretical: aligning longitudinal processes leads to earlier and more stable detection of treatment effects, suggesting that part of the variability observed in analyses that do not account for time-varying covariates and informative censoring may be driven by uncontrolled heterogeneity rather than the true signal.

Several limitations should be noted. First, the validity of the method relies on the correct specification of the nuisance models used to construct the weights, and violations of positivity may arise in the presence of extreme censoring or rare covariate patterns, leading to instability. Second, the separable effects framework requires structural assumptions on how treatment components propagate through the longitudinal system, which, although clinically motivated, are not empirically testable. In addition, the current implementation focuses on a single recurrent event type and a fixed treatment assignment at baseline, which may not capture the full complexity of clinical decision-making. Finally, formal comparisons across device brands within the same generation were not

pursued, as such analyses are difficult to interpret due to potential center-level practice patterns, surgeon preference, institutional procurement, and conflicts of interest influencing device choice; addressing these concerns would require methods that account for unmeasured confounding (e.g., instrumental variable approaches), and thus remains an important limitation and direction for future research.

Future work can extend this framework in several directions. Allowing for multiple types of recurrent events would enable joint modeling of competing or correlated complications and provide a more complete assessment of disease burden. Developing doubly robust versions of the test, combining outcome modeling with weighting, would mitigate sensitivity to nuisance model misspecification and improve stability in finite samples. In addition, extending the testing procedure to dynamic treatment regimes would allow treatment strategies to evolve over time based on patient history, while integrating individualized risk prediction could support patient-specific inference on recurrent outcomes. Together, these directions would move the methodology toward more flexible and clinically actionable tools for analyzing complex longitudinal event data.

CHAPTER 5

Future Directions

This chapter outlines several important directions for future research that build on the methodological developments presented in this dissertation. Although the proposed framework provides a rigorous approach for inference on recurrent events in the presence of terminal events, many practical and methodological challenges remain open. In particular, future work should aim to broaden the framework to accommodate more complex recurrent event structures, improve robustness to nuisance model misspecification, and extend the current methodology to settings with adaptive treatment strategies and individualized clinical decision-making. These directions would further enhance the flexibility, interpretability, and practical usefulness of the proposed methods in modern biomedical studies.

5.1 Multiple Types of Recurrent Events

The current framework focuses on a single recurrent event type, whereas many clinical settings involve multiple recurrent outcomes with distinct etiologies and dependencies. Extending the proposed separable-effects testing framework to multivariate recurrent event processes would enable a more comprehensive assessment of treatment effects across competing or co-occurring events. This requires accommodating dependence across event types and defining pathway-specific effects in a multistate or marked point process setting, while preserving the causal interpretation and validity of the test.

5.2 Doubly Robust Estimation

The proposed method relies on correct specification of nuisance models for treatment, longitudinal covariates, terminal events, and censoring. Developing doubly robust estimating equations would improve robustness by ensuring consistency when either the weighting models or outcome models are correctly specified. This may be achieved by augmenting the inverse probability weighting framework with outcome regression components, potentially combined with machine learning and cross-fitting to enhance finite-sample performance while retaining valid asymptotic inference.

5.3 Dynamic Treatment Regimes and Individualized Prediction

An important extension is to dynamic treatment regimes, where treatment decisions evolve over time based on patient history. Incorporating such regimes into the separable-effects framework would require redefining pathway-specific interventions and corresponding identification strategies. In addition, extending the method to individualized risk prediction, such as patient-specific recurrent event burden under hypothetical interventions, would enable personalized decision support, bridging causal inference and predictive modeling in longitudinal settings with recurrent and terminal events.

Overall, these future directions reflect a natural progression of the work developed in this dissertation. Extending the framework to handle multiple recurrent outcomes, strengthening robustness through doubly robust estimation, and incorporating dynamic treatment strategies and individualized prediction would substantially expand both its methodological scope and its practical relevance. Collectively, these developments would move the proposed framework toward a more comprehensive and clinically actionable approach for analyzing complex longitudinal data with recurrent and terminal events.

APPENDIX A

Derivation of Estimating Equation in the Proportional Rate Marginal Structural Model

By rewriting the equation (3.1) in Section 3.3 in continuous-time format and denoting the recurrent event process by the counting process notation $N(t)$ (Fleming & Harrington 2013), the separable increment estimator is

$$d\hat{\mu}^{a_Y,a_D}(t) = \frac{\sum_{i=1}^{n} Z_i(t)\, I(A_i = a_Y)\, \hat{S}^{a_Y}(t)\, \hat{W}_i^D(a_Y, a_D; t)\, dN_i(t)}{\sum_{j=1}^{n} Z_j(t)\, I(A_j = a_Y)}.$$

Specializing to $a_Y = 1$ gives

$$d\hat{\mu}^{a_Y=1,a_D}(t) = \frac{\sum_{i=1}^{n} Z_i(t) A_i\, \hat{S}^{a_Y=1}(t)\, \hat{W}_i^D(1, a_D; t)\, dN_i(t)}{D_1(t)}, \tag{A.1}$$

where

$$D_1(t) := \sum_{j=1}^{n} Z_j(t) A_j.$$

The PR-MSM in equation (3.2) posits

$$\mu^{a_Y=1,a_D}(t) = \mu^{a_Y=0,a_D}(t) \exp(\beta^{a_D}) \quad \Longleftrightarrow \quad d\mu^{a_Y=1,a_D}(t) = \exp(\beta^{a_D})\, d\mu^{a_Y=0,a_D}(t).$$

To connect the observed data to the model for counterfactual means $\mu^{a_Y,a_D}(t)$, we replace each

$d\mu^{a_Y,a_D}(t)$ by its consistent estimator $d\hat{\mu}^{a_Y,a_D}(t)$ (Janvin et al. 2024):

$$d\hat{\mu}^{a_Y=1,a_D}(t) = \exp(\beta^{a_D})\, d\hat{\mu}^{a_Y=0,a_D}(t). \tag{A.2}$$

Plugging (A.2) on the left of (A.1) while keeping the right-hand side as the estimator yields the moment condition:

$$\exp(\beta^{a_D})\, d\hat{\mu}^{a_Y=0,a_D}(t) = \frac{\sum_{i=1}^{n} Z_i(t) A_i\, \hat{S}^{a_Y=1}(t)\, \hat{W}_i^D(1, a_D; t)\, dN_i(t)}{D_1(t)}. \tag{A.3}$$

Multiplying both sides of (A.3) by $D_1(t)$ and rearranging leads to

$$\sum_{i=1}^{n} Z_i(t) A_i\, \hat{S}^{a_Y=1}(t)\, \hat{W}_i^D(1, a_D; t)\, dN_i(t) \;-\; \exp(\beta^{a_D})\, D_1(t)\, d\hat{\mu}^{a_Y=0,a_D}(t) \;=\; 0. \tag{A.4}$$

Now replace the unknown $d\mu^{a_Y=0,a_D}(t)$ by its estimator, obtained from the same equation (3.1) in Section 3.3 with $a_Y = 0$, we have

$$d\hat{\mu}^{a_Y=0,a_D}(t) = \frac{\sum_{i=1}^{n} Z_i(t)(1 - A_i)\, \hat{S}^{a_Y=0}(t)\, \hat{W}_i^D(0, a_D; t)\, dN_i(t)}{D_0(t)}, \tag{A.5}$$

where

$$D_0(t) := \sum_{j=1}^{n} Z_j(t)(1 - A_j).$$

Substituting (A.5) into (A.4) gives the estimating equation with combined summations on both treatment groups:

$$\begin{aligned} &\sum_{i=1}^{n} Z_i(t) A_i\, \hat{S}^{a_Y=1}(t)\, \hat{W}_i^D(1, a_D; t)\, dN_i(t) \\ &- \exp(\beta^{a_D})\, \frac{D_1(t)}{D_0(t)} \sum_{i=1}^{n} Z_i(t)(1 - A_i)\, \hat{S}^{a_Y=0}(t)\, \hat{W}_i^D(0, a_D; t)\, dN_i(t) \;=\; 0. \end{aligned}$$

Integrating t over $[0, \tau]$ yields the sample estimating equation for β^{a_D}:

$$\int_0^\tau \Bigg[\sum_{i=1}^{n} Z_i(t) A_i \, \hat{S}^{a_Y=1}(t) \, \hat{W}_i^D(1, a_D; t) \, dN_i(t) \\ - \exp(\beta^{a_D}) \, \frac{D_1(t)}{D_0(t)} \, \sum_{i=1}^{n} Z_i(t)(1 - A_i) \, \hat{S}^{a_Y=0}(t) \, \hat{W}_i^D(0, a_D; t) \, dN_i(t) \Bigg] = 0. \tag{A.6}$$

Finally, noting that the first summation corresponds to $d\hat{\mu}^{a_Y=1,a_D}(t)$ and the second to $d\hat{\mu}^{a_Y=0,a_D}(t)$, we can write (A.6) compactly as

$$\int_0^\tau \sum_{j=1}^{n} Z_j(t) A_j \left\{ d\hat{\mu}^{a_Y=1,a_D}(t) - \exp(\beta^{a_D}) d\hat{\mu}^{a_Y=0,a_D}(t) \right\} = 0.$$

APPENDIX B

The Asymptotic Normality for the Test Statistic of PR-MSMaT

The proof of Theorem 3.4.

Let $N_i(t)$ denote the recurrent-event counting process for subject i on $[0, \tau]$, $Z_i(t)$ the at-risk indicator, and $A_i \in \{0, 1\}$ the randomized treatment. Write $M_i(t) = N_i(t) - \Lambda_i(t)$, where $\Lambda_i(t)$ is the compensator of $N_i(t)$ with respect to the natural filtration. Then $\{M_i(t)\}$ is a square-integrable local martingale with predictable variation $\langle M_i \rangle(t)$ (Andersen et al. 2012).

For $g \in \{0, 1\}$, define $D_g(t) = \sum_{j=1}^{n} Z_j(t)\, I(A_j = g)$ and assume $D_g(t)/n \to p_g(t)$ uniformly with $\inf_{t \le \tau} p_g(t) > 0$.

Let $\hat{S}^{a_Y}(t)$ be a uniformly consistent estimator of $S^{a_Y}(t)$ and $\hat{W}_i^D(a_Y, a_D; t)$ uniformly consistent for the corresponding weights. Define the separable mean-increment estimator

$$d\hat{\mu}^{a_Y, a_D}(t) = \frac{\sum_{i=1}^{n} Z_i(t)\, I(A_i = a_Y)\, \hat{S}^{a_Y}(t)\, \hat{W}_i^D(a_Y, a_D; t)\, dN_i(t)}{D_{a_Y}(t)},$$

which corresponds to (3.1) in Section 3.3.

From (3.4) in Section 3.4, the PR-MSMaT score under $H_0 \colon \beta^{a_D} = 0$ is

$$U_n = \int_0^{\tau} \sum_{j=1}^{n} Z_j(t) A_j \left\{ d\hat{\mu}^{1, a_D}(t) - d\hat{\mu}^{0, a_D}(t) \right\}. \tag{A.7}$$

Fix $a_Y \in \{0, 1\}$. Writing

$$d\hat{\mu}^{a_Y,a_D}(t) = \frac{1}{D_{a_Y}(t)} \sum_{i=1}^{n} I(A_i = a_Y)\, Z_i(t)\, \hat{S}^{a_Y}(t)\, \hat{W}_i^D(a_Y, a_D; t)\, dN_i(t),$$

replace $dN_i(t)$ by $dM_i(t) + d\Lambda_i(t)$ and expand the numerator and denominator around their expectations. A standard Taylor expansion for ratios of empirical processes, combined with uniform consistency of $\hat{S}^{a_Y}$ and $\hat{W}_i^D$ and the uniform LLN for predictable bounded processes (Andersen et al. 2012), yields

$$d\hat{\mu}^{a_Y,a_D}(t) = \mu'^{a_Y,a_D}(t)\, dt + \frac{1}{n} \sum_{i=1}^{n} \phi_i^{a_Y,a_D}(t)\, dM_i(t) + r_n^{a_Y,a_D}(t), \tag{A.8}$$

with $\sup_{t \le \tau} |r_n^{a_Y,a_D}(t)| = o_p(n^{-1/2})$, and

$$\phi_i^{a_Y,a_D}(t) = \frac{n\, I(A_i = a_Y)\, Z_i(t)\, S^{a_Y}(t)\, W_i^D(a_Y, a_D; t)}{D_{a_Y}(t)}.$$

Substituting (A.8) into (A.7),

$$U_n = \frac{1}{n} \sum_{i=1}^{n} \int_0^{\tau} \psi_i(t)\, dM_i(t) + o_p(n^{1/2}),$$

where $\psi_i(t) = \sum_{j=1}^{n} Z_j(t) A_j \{\phi_i^{1,a_D}(t) - \phi_i^{0,a_D}(t)\}$.

Define

$$\xi_{n,i} := \int_0^{\tau} \frac{\psi_i(t)}{n^{3/2}}\, dM_i(t), \quad \text{so that} \quad \sum_{i=1}^{n} \xi_{n,i} = n^{-1/2} U_n + o_p(1).$$

The predictable variation satisfies

$$V_n = \sum_{i=1}^{n} \langle \xi_{n,i} \rangle = \sum_{i=1}^{n} \int_0^{\tau} \frac{\psi_i^2(t)}{n^3}\, d\langle M_i \rangle(t).$$

Assume: (i) i.i.d. subjects; (ii) bounded predictable $\phi_i^{a_Y,a_D}(t)$; (iii) $\inf_t D_g(t)/n > 0$ and $\sup_t D_g(t)/n < \infty$; (iv) the martingale Lindeberg condition (Andersen et al. 2012, Hall & Heyde

2014).

Because increments are scaled by $1/n^{3/2}$ and weights are bounded, (iv) holds automatically. Thus $V_n \xrightarrow{p} \sigma^2 \in (0, \infty)$.

By Rebolledo's martingale central limit theorem (Rebolledo 1980),

$$\sum_{i=1}^{n} \xi_{n,i} \xrightarrow{d} N(0, \sigma^2), \quad \text{equivalently} \quad n^{-1/2} U_n \xrightarrow{d} N(0, \sigma^2).$$

A consistent variance estimator is

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \int_0^{\tau} \frac{\{\hat{\psi}_i(t)\}^2}{n^3} \, dN_i(t),$$

so that $\widehat{\mathrm{Var}}(U_n) = n\,\hat{\sigma}^2$ satisfies

$$\frac{\widehat{\mathrm{Var}}(U_n)}{n} \xrightarrow{p} \sigma^2.$$

Hence the PR-MSMaT statistic

$$T_n = \frac{U_n}{\{\widehat{\mathrm{Var}}(U_n)\}^{1/2}} = \frac{n^{-1/2} U_n}{\hat{\sigma}} \xrightarrow{d} N(0, 1),$$

under $H_0 : \beta^{a_D} = 0$.

APPENDIX C

Summary Table of Baseline Covariates in Dataset after Matching

A summary table referenced in Section 4.5 is attached here. This table shows satisfactory 1:1 matching results for the baseline factors in the data to justify the application of the proposed causal inference in the empirical study.

Table C.1: Baseline characteristics by device type.

Characteristic	BIVAD (N=1051)	LVAD (N=1051)	Total (N=2102)	p-value
AGE_DEIDENT				0.308
Mean (SD)	51.999 (14.016)	52.614 (13.622)	52.306 (13.820)	
Median (Q1, Q3)	54.000 (43.000, 63.000)	55.000 (44.000, 63.000)	55.000 (43.000, 63.000)	
Min – Max	19.000 – 83.000	19.000 – 80.000	19.000 – 83.000	
GENDER				0.518
F	279 (26.5%)	266 (25.3%)	545 (25.9%)	
M	772 (73.5%)	785 (74.7%)	1557 (74.1%)	
PX_PROFILE				0.825
1	545 (51.9%)	549 (52.2%)	1094 (52.0%)	
3	355 (33.8%)	339 (32.3%)	694 (33.0%)	
5	102 (9.7%)	109 (10.4%)	211 (10.0%)	
7	32 (3.0%)	37 (3.5%)	69 (3.3%)	

Table C.1: Baseline characteristics by device type (continued).

Characteristic	BIVAD (N=1051)	LVAD (N=1051)	Total (N=2102)	P-value
9	11 (1.0%)	9 (0.9%)	20 (1.0%)	
11	5 (0.5%)	8 (0.8%)	13 (0.6%)	
13	1 (0.1%)	0 (0.0%)	1 (0.0%)	
HR_RATE				0.999
Mean (SD)	95.559 (20.447)	95.560 (18.872)	95.560 (19.670)	
Median (Q1, Q3)	94.000 (80.000, 109.000)	95.000 (81.000, 110.000)	94.000 (80.000, 109.000)	
Min – Max	46.000 – 198.000	49.000 – 154.000	46.000 – 198.000	
PUL_DIA_PRES				0.136
Mean (SD)	25.452 (7.654)	25.979 (8.526)	25.715 (8.104)	
Median (Q1, Q3)	25.740 (21.000, 30.000)	26.000 (20.000, 31.000)	25.855 (21.000, 30.000)	
Min – Max	1.000 – 60.000	2.000 – 60.000	1.000 – 60.000	
PUL_SYS_PRES				0.122
Mean (SD)	47.171 (13.522)	48.108 (14.269)	47.640 (13.905)	
Median (Q1, Q3)	48.000 (39.000, 56.000)	48.000 (39.000, 57.500)	48.000 (39.000, 56.000)	
Min – Max	4.000 – 96.000	5.000 – 95.000	4.000 – 96.000	
LVEDD				0.705
Mean (SD)	6.543 (1.118)	6.561 (1.112)	6.552 (1.115)	
Median (Q1, Q3)	6.574 (5.900, 7.169)	6.600 (5.900, 7.200)	6.600 (5.900, 7.200)	
Min – Max	2.600 – 10.100	1.100 – 10.200	1.100 – 10.200	
LVEF				0.216
1	14 (1.3%)	6 (0.6%)	20 (1.0%)	
2	8 (0.8%)	3 (0.3%)	11 (0.5%)	
3	45 (4.3%)	55 (5.2%)	100 (4.8%)	
4	186 (17.7%)	198 (18.8%)	384 (18.3%)	
5	34 (3.2%)	37 (3.5%)	71 (3.4%)	
6	764 (72.7%)	752 (71.6%)	1516 (72.1%)	

Table C.1: Baseline characteristics by device type (continued).

Characteristic	BIVAD (N=1051)	LVAD (N=1051)	Total (N=2102)	P-value
RVEF				0.896
1	109 (10.4%)	123 (11.7%)	232 (11.0%)	
2	139 (13.2%)	138 (13.1%)	277 (13.2%)	
3	282 (26.8%)	293 (27.9%)	575 (27.4%)	
4	321 (30.5%)	304 (28.9%)	625 (29.7%)	
5	199 (18.9%)	192 (18.3%)	391 (18.6%)	
6	1 (0.1%)	1 (0.1%)	2 (0.1%)	
HEMOGLOBIN_G_DL				0.115
Mean (SD)	10.379 (2.071)	10.235 (2.103)	10.307 (2.088)	
Median (Q1, Q3)	10.100 (8.900, 11.700)	10.000 (8.800, 11.500)	10.000 (8.800, 11.600)	
Min – Max	1.040 – 18.600	0.850 – 18.600	0.850 – 18.600	
PLATELET_X10_3_UL				0.534
Mean (SD)	166.573 (91.400)	168.994 (87.079)	167.784 (89.253)	
Median (Q1, Q3)	155.000 (99.000, 212.000)	161.000 (108.000, 211.500)	157.000 (103.000, 212.000)	
Min – Max	20.000 – 731.000	5.000 – 800.000	5.000 – 800.000	
WBC_X10_3_UL				0.963
Mean (SD)	10.984 (5.330)	10.971 (6.611)	10.977 (6.004)	
Median (Q1, Q3)	9.800 (7.200, 13.400)	9.400 (7.200, 12.900)	9.600 (7.200, 13.100)	
Min – Max	2.600 – 35.800	2.400 – 88.000	2.400 – 88.000	
LYMPH_CNT_X10_3_UL				0.798
Mean (SD)	1.334 (0.502)	1.329 (0.478)	1.331 (0.490)	
Median (Q1, Q3)	1.122 (0.972, 1.758)	1.138 (0.969, 1.743)	1.132 (0.970, 1.750)	
Min – Max	0.100 – 5.300	0.100 – 3.210	0.100 – 5.300	
BUN_MG_DL				0.771
Mean (SD)	34.503 (23.852)	34.212 (21.733)	34.357 (22.812)	
Median (Q1, Q3)	29.000 (19.000, 44.000)	29.000 (19.000, 42.000)	29.000 (19.000, 43.000)	
Min – Max	1.000 – 336.120	1.900 – 211.700	1.000 – 336.120	

Table C.1: Baseline characteristics by device type (continued).

Characteristic	BIVAD (N=1051)	LVAD (N=1051)	Total (N=2102)	P-value
HEMO_A1C_PERCENT				0.866
Mean (SD)	6.502 (0.847)	6.509 (0.976)	6.505 (0.914)	
Median (Q1, Q3)	6.073 (5.926, 7.249)	6.076 (5.912, 7.164)	6.074 (5.917, 7.202)	
Min – Max	4.500 – 10.700	3.718 – 13.500	3.718 – 13.500	
ALBUMIN_G_L				0.470
Mean (SD)	30.910 (6.953)	30.694 (6.765)	30.802 (6.859)	
Median (Q1, Q3)	31.000 (27.000, 35.000)	31.000 (27.000, 35.000)	31.000 (27.000, 35.000)	
Min – Max	2.700 – 67.000	1.000 – 54.000	1.000 – 67.000	
SGOT_AST				0.061
Mean (SD)	189.153 (663.323)	139.652 (543.768)	164.403 (606.860)	
Median (Q1, Q3)	43.000 (27.000, 96.500)	37.000 (25.000, 72.000)	40.000 (26.000, 85.000)	
Min – Max	8.000 – 10000.000	7.000 – 10000.000	7.000 – 10000.000	
SGPT_ALT				0.211
Mean (SD)	173.012 (496.815)	145.306 (517.153)	159.159 (507.154)	
Median (Q1, Q3)	38.000 (22.000, 95.500)	36.000 (21.000, 76.500)	37.000 (22.000, 84.000)	
Min – Max	3.000 – 8230.000	3.000 – 9930.000	3.000 – 9930.000	
CHOLESTEROL_MG_DL				0.412
Mean (SD)	113.826 (28.371)	114.917 (32.493)	114.371 (30.499)	
Median (Q1, Q3)	113.150 (98.350, 127.478)	113.000 (95.000, 131.000)	113.000 (97.000, 129.000)	
Min – Max	50.000 – 265.000	50.000 – 257.000	50.000 – 265.000	
INR				0.197
Mean (SD)	1.489 (0.577)	1.457 (0.576)	1.473 (0.577)	
Median (Q1, Q3)	1.348 (1.200, 1.600)	1.300 (1.200, 1.500)	1.300 (1.200, 1.600)	
Min – Max	0.600 – 8.100	0.800 – 7.900	0.600 – 8.100	
BNP_PG_ML				0.595

Table C.1: Baseline characteristics by device type (continued).

Characteristic	BIVAD (N=1051)	LVAD (N=1051)	Total (N=2102)	P-value
Mean (SD)	1554.408 (1099.192)	1529.098 (1082.487)	1541.753 (1090.685)	
Median (Q1, Q3)	1130.000 (782.235, 2250.370)	1178.250 (749.350, 2175.654)	1146.610 (774.075, 2200.000)	
Min – Max	5.000 – 7052.000	14.000 – 5686.000	5.000 – 7052.000	
PRO_BNP_PG_ML				0.441
Mean (SD)	9826.448 (6907.457)	9592.152 (7022.210)	9709.300 (6964.398)	
Median (Q1, Q3)	7255.670 (5354.685, 12496.680)	7388.000 (5053.205, 12265.235)	7336.500 (5247.295, 12380.118)	
Min – Max	46.000 – 60000.000	190.000 – 60000.000	46.000 – 60000.000	
RACE				0.698
Black	214 (20.4%)	208 (19.8%)	422 (20.1%)	
Others	88 (8.4%)	79 (7.5%)	167 (7.9%)	
White	749 (71.3%)	764 (72.7%)	1513 (72.0%)	
BMI				0.849
Mean (SD)	28.280 (6.608)	28.335 (6.595)	28.308 (6.600)	
Median (Q1, Q3)	27.468 (23.995, 31.786)	27.774 (24.007, 31.729)	27.673 (23.997, 31.753)	
Min – Max	9.428 – 63.867	6.691 – 68.530	6.691 – 68.530	
CARDIAC_INDEX				0.413
Mean (SD)	2.102 (0.521)	2.122 (0.608)	2.112 (0.566)	
Median (Q1, Q3)	2.060 (1.800, 2.338)	2.080 (1.722, 2.383)	2.070 (1.790, 2.360)	
Min – Max	0.750 – 5.000	0.600 – 4.900	0.600 – 5.000	
TRICUSPID_REGURG				0.730
1	92 (8.8%)	87 (8.3%)	179 (8.5%)	
2	401 (38.2%)	405 (38.5%)	806 (38.3%)	
3	315 (30.0%)	302 (28.7%)	617 (29.4%)	
4	180 (17.1%)	179 (17.0%)	359 (17.1%)	
5	63 (6.0%)	78 (7.4%)	141 (6.7%)	
Smoking				0.906

Table C.1: Baseline characteristics by device type (continued).

Characteristic	BIVAD (N=1051)	LVAD (N=1051)	Total (N=2102)	P-value
0	879 (83.6%)	877 (83.4%)	1756 (83.5%)	
1	172 (16.4%)	174 (16.6%)	346 (16.5%)	
Hypertension				0.843
0	922 (87.7%)	919 (87.4%)	1841 (87.6%)	
1	129 (12.3%)	132 (12.6%)	261 (12.4%)	
Diabetes				0.326
0	976 (92.9%)	964 (91.7%)	1940 (92.3%)	
1	75 (7.1%)	87 (8.3%)	162 (7.7%)	
Frailty				1.000
0	950 (90.4%)	950 (90.4%)	1900 (90.4%)	
1	101 (9.6%)	101 (9.6%)	202 (9.6%)	
Renal.Disease				0.500
0	900 (85.6%)	889 (84.6%)	1789 (85.1%)	
1	151 (14.4%)	162 (15.4%)	313 (14.9%)	
Liver.Dysfunciton				0.884
0	946 (90.0%)	948 (90.2%)	1894 (90.1%)	
1	105 (10.0%)	103 (9.8%)	208 (9.9%)	
GI.Ulcer				0.711
0	948 (90.2%)	953 (90.7%)	1901 (90.4%)	
1	103 (9.8%)	98 (9.3%)	201 (9.6%)	
Lung.Disease				0.487
0	984 (93.6%)	976 (92.9%)	1960 (93.2%)	
1	67 (6.4%)	75 (7.1%)	142 (6.8%)	
Peripheral.arterial.disease				0.581
0	1026 (97.6%)	1022 (97.2%)	2048 (97.4%)	
1	25 (2.4%)	29 (2.8%)	54 (2.6%)	
NYHA_IV				0.266
0	93 (8.8%)	108 (10.3%)	201 (9.6%)	

Table C.1: Baseline characteristics by device type (continued).

Characteristic	BIVAD (N=1051)	LVAD (N=1051)	Total (N=2102)	P-value
1	958 (91.2%)	943 (89.7%)	1901 (90.4%)	
PRIMARY_DGN_isch				0.302
0	659 (62.7%)	636 (60.5%)	1295 (61.6%)	
1	392 (37.3%)	415 (39.5%)	807 (38.4%)	
CONCOM_SURG				0.840
0	927 (88.2%)	924 (87.9%)	1851 (88.1%)	
1	124 (11.8%)	127 (12.1%)	251 (11.9%)	
PVR				0.854
Mean (SD)	171.214 (142.154)	169.898 (182.053)	170.556 (163.289)	
Median (Q1, Q3)	155.835 (99.042, 231.950)	145.340 (80.912, 232.306)	151.433 (89.619, 232.011)	
Min – Max	-308.943 – 1600.000	-651.852 – 1435.897	-651.852 – 1600.000	
PAPI				0.414
Mean (SD)	1.992 (1.552)	2.051 (1.754)	2.021 (1.656)	
Median (Q1, Q3)	1.833 (1.049, 2.514)	1.850 (1.143, 2.587)	1.846 (1.091, 2.556)	
Min – Max	-7.000 – 19.000	-27.000 – 15.500	-27.000 – 19.000	
RVSWI				0.134
Mean (SD)	0.431 (0.240)	0.447 (0.263)	0.439 (0.251)	
Median (Q1, Q3)	0.421 (0.275, 0.574)	0.425 (0.286, 0.582)	0.423 (0.280, 0.579)	
Min – Max	-0.324 – 1.734	-0.314 – 2.405	-0.324 – 2.405	
ratio_RA_PUL				0.514
Mean (SD)	0.631 (0.297)	0.618 (0.551)	0.624 (0.443)	
Median (Q1, Q3)	0.562 (0.466, 0.725)	0.533 (0.407, 0.690)	0.547 (0.443, 0.709)	
Min – Max	0.000 – 3.000	-0.059 – 13.000	-0.059 – 13.000	
GFR				0.516
Mean (SD)	62.103 (29.069)	62.927 (29.040)	62.515 (29.051)	
Median (Q1, Q3)	57.981 (39.656, 80.861)	57.983 (39.942, 83.911)	57.982 (39.750, 82.442)	

Table C.1: Baseline characteristics by device type (continued).

Characteristic	BIVAD (N=1051)	LVAD (N=1051)	Total (N=2102)	P-value
Min – Max	7.587 – 159.847	4.756 – 171.497	4.756 – 171.497	

Values are *n (%)* for categorical variables and mean (SD), median (Q1, Q3), and min–max for continuous variables. The p-values were from chi-square or Fisher's exact tests for categorical variables and 2-sample t tests or Mann-Whitney U tests to compare normally distributed and nonnormally distributed continuous variables across groups, respectively.

BIBLIOGRAPHY

Andersen, P. K., Borgan, O., Gill, R. D. & Keiding, N. (2012), *Statistical Models Based on Counting Processes*, Springer Science & Business Media.

Andersen, P. K. & Gill, R. D. (1982), 'Cox's regression model for counting processes: A large sample study', *The Annals of Statistics* **10**(4), 1100–1120.

Anker, S. D., Schroeder, S., Atar, D., Bax, J. J., Ceconi, C., Cowie, M. R., Crisp, A., Dominjon, F., Ford, I., Ghofrani, H.-A. et al. (2016), 'Traditional and new composite endpoints in heart failure clinical trials: facilitating comprehensive efficacy assessments and improving trial efficiency', *European Journal of Heart Failure* **18**(5), 482–489.

Bakal, J. A., Westerhout, C. M. & Armstrong, P. W. (2015), 'Impact of weighted composite compared to traditional composite endpoints for the design of randomized controlled trials', *Statistical Methods in Medical Research* **24**(6), 980–988.

Cleveland, J. C., Naftel, D. C., Reece, T. B. & et al. (2011), 'Survival after biventricular assist device implantation: An analysis of the intermacs registry', *Journal of Heart and Lung Transplantation* **30**(8), 862–869.

Emura, T., Matsui, S. & Rondeau, V. (2019), *Survival Analysis with Correlated Endpoints: Joint Frailty-Copula Models*, Springer.

Fitzpatrick, J. R., Frederick, J. R., Hiesinger, W. & et al. (2009), 'Early planned institution of biventricular mechanical circulatory support results in improved outcomes compared with delayed

conversion of a left ventricular assist device to a biventricular assist device', *Journal of Thoracic and Cardiovascular Surgery* **137**(4), 971–977.

Fleming, T. R. & Harrington, D. P. (2013), *Counting Processes and Survival Analysis*, John Wiley & Sons.

Freemantle, N., Calvert, M., Wood, J., Eastaugh, J. & Griffin, C. (2003), 'Composite outcomes in randomized trials: greater precision but with greater uncertainty?', *Journal of the American Medical Association* **289**(19), 2554–2559.

Fritsch, A., Schlömer, P., Mendolia, F., Mütze, T., Jahn-Eimermacher, A. & Consortium, R. E. Q. O. (2023), 'Efficiency comparison of analysis methods for recurrent event and time-to-first event endpoints in the presence of terminal events—application to clinical trials in chronic heart failure', *Statistics in Biopharmaceutical Research* **15**(2), 268–279.

Furberg, J. K., Rasmussen, S., Andersen, P. K. & Ravn, H. (2022), 'Methodological challenges in the analysis of recurrent events for randomised controlled trials with application to cardiovascular events in leader', *Pharmaceutical Statistics* **21**(1), 241–267.

Gaber, C. E., Edwards, J. K., Lund, J. L., Peery, A. F., Richardson, D. B. & Kinlaw, A. C. (2023), 'Inverse probability weighting to estimate exposure effects on the burden of recurrent outcomes in the presence of competing events', *American Journal of Epidemiology* **192**(5), 830–839.

Ghosh, D. & Lin, D. (2000), 'Nonparametric analysis of recurrent events and death', *Biometrics* **56**(2), 554–562.

Ghosh, D. & Lin, D. Y. (2002), 'Marginal regression models for recurrent and terminal events', *Statistica Sinica* **12**(3), 663–688.

Gregorio, C., Cappelletto, C., Romani, S., Stolfo, D., Merlo, M. & Barbati, G. (2022), 'Using marginal structural joint models to estimate the effect of a time-varying treatment on recurrent

events and survival: an application on arrhythmogenic cardiomyopathy', *Biometrical Journal* **64**(8), 1374–1388.

Hall, P. & Heyde, C. C. (2014), *Martingale Limit Theory and its Application*, Academic press.

Hammer, Y., Xie, J., Yang, G., Bitar, A., Haft, J. W., Cascino, T. M., Likosky, D. S., Pagani, F. D., Zhang, M. & Aaronson, K. D. (2024), 'Gastrointestinal bleeding following heartmate 3 left ventricular assist device implantation: The michigan bleeding risk model', *The Journal of Heart and Lung Transplantation* **43**(4), 604–614.

Hannan, M. M., Husain, S., Mattner, F., Danziger-Isakov, L., Drew, R. J., Corey, G. R., Schueler, S., Holman, W. L., Lawler, L. P., Gordon, S. M. et al. (2011), 'Working formulation for the standardization of definitions of infections in patients using ventricular assist devices', *The Journal of Heart and Lung Transplantation* **30**(4), 375–384.

Heddle, N. M. & Cook, R. J. (2010), 'Composite outcomes in clinical trials: what are they and when should they be used?', *Transfusion* **51**(1), 11–13.

Huang, C.-Y. & Wang, M.-C. (2004), 'Joint modeling and estimation for recurrent event processes and failure time data', *Journal of the American Statistical Association* **99**(468), 1153–1165.

Huang, Y., Zhou, L., Zhang, M. & Song, P. X. (2026), 'Model-assisted causal inference for the treatment effect on recurrent events in the presence of terminal events', *arXiv preprint arXiv:2601.01245* .

Janvin, M., Young, J. G., Ryalen, P. C. & Stensrud, M. J. (2024), 'Causal inference with recurrent and competing events', *Lifetime Data Analysis* **30**(1), 59–118.

Jensen, A. K., Ravn, H., Sørup, S. & Andersen, P. (2016), 'A marginal structural model for recurrent events in the presence of time-dependent confounding: non-specific effects of vaccines on child hospitalisations', *Statistics in Medicine* **35**(27), 5051–5069.

Kahan, B. C., Jairath, V., Doré, C. J. & Morris, T. P. (2014), 'The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies', *Trials* **15**(1), 139.

Kip, K. E., Hollabaugh, K., Marroquin, O. C. & Williams, D. O. (2008), 'The problem with composite end points in cardiovascular studies: the story of major adverse cardiac events and percutaneous coronary intervention', *Journal of the American College of Cardiology* **51**(7), 701–707.

Kirklin, J. K., Naftel, D. C., Pagani, F. D. & et al. (2015), 'Seventh intermacs annual report: 15,000 patients and counting', *Journal of Heart and Lung Transplantation* **34**(12), 1495–1504.

Lachin, J. M. & Bebu, I. (2015), 'Application of the wei–lachin multivariate one-directional test to multiple event-time outcomes', *Clinical Trials* **12**(6), 627–633.

Lawless, J. F. & Nadeau, C. (1995), 'Some simple robust methods for the analysis of recurrent events', *Technometrics* **37**(2), 158–168.

Lin, D. Y., Wei, L.-J., Yang, I. & Ying, Z. (2000), 'Semiparametric regression for the mean and rate functions of recurrent events', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**(4), 711–730.

Mao, L. (2023), 'Nonparametric inference of general while-alive estimands for recurrent events', *Biometrics* **79**(3), 1749–1760.

Mao, L. & Kim, K. (2021), 'Statistical models for composite endpoints of death and nonfatal events: a review', *Statistics in Biopharmaceutical Research* **13**(3), 260–269.

Mao, L., Kim, K. & Li, Y. (2022), 'On recurrent-event win ratio', *Statistical Methods in Medical Research* **31**(6), 1120–1134.

Mao, L. & Lin, D. (2016), 'Semiparametric regression for the weighted composite endpoint of recurrent and terminal events', *Biostatistics* **17**(2), 390–403.

Mao, L. & Wang, T. (2021), 'A class of proportional win-fractions regression models for composite outcomes', *Biometrics* **77**(4), 1265–1275.

Nicholls, S. J., Lincoff, A. M., Garcia, M., Bash, D., Ballantyne, C. M., Barter, P. J., Davidson, M. H., Kastelein, J. J., Koenig, W., McGuire, D. K. et al. (2020), 'Effect of high-dose omega-3 fatty acids vs corn oil on major adverse cardiovascular events in patients at high cardiovascular risk: the strength randomized clinical trial', *Journal of the American Medical Association* **324**(22), 2268–2280.

Nienaber, J. J. C., Kusne, S., Riaz, T., Walker, R. C., Baddour, L. M., Wright, A. J., Park, S. J., Vikram, H. R., Keating, M. R., Arabia, F. A. et al. (2013), 'Clinical manifestations and management of left ventricular assist device–associated infections', *Clinical Infectious Diseases* **57**(10), 1438–1448.

Ozga, A.-K. & Rauch, G. (2022), 'Weighted composite time to event endpoints with recurrent events: comparison of three analytical approaches', *BMC Medical Research Methodology* **22**(1), 1–14.

Pan, M., Vasbinder, A., Anderson, E., Catalan, T., Shadid, H. R., Berlin, H., Padalia, K., O'Hayer, P., Meloche, C., Azam, T. U. et al. (2021), 'Angiotensin-converting enzyme inhibitors, angiotensin ii receptor blockers, and outcomes in patients hospitalized for covid-19', *Journal of the American Heart Association* **10**(24), e023535.

Pepe, M. S. & Cai, J. (1993), 'Some graphical displays and marginal regression analyses for recurrent failure times and time-dependent covariates', *Journal of the American Statistical Association* **88**(423), 811–820.

Pocock, S. J., Ariti, C. A., Collier, T. J. & Wang, D. (2012), 'The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities', *European Heart Journal* **33**(2), 176–182.

Rauch, G. & Beyersmann, J. (2013), 'Planning and evaluating clinical trials with composite time-to-first-event endpoints in a competing risk framework', *Statistics in Medicine* **32**(21), 3595–3608.

Rauch, G., Kieser, M., Binder, H., Bayes-Genis, A. & Jahn-Eimermacher, A. (2018), 'Time-to-first-event versus recurrent-event analysis: points to consider for selecting a meaningful analysis strategy in clinical trials with composite endpoints', *Clinical Research in Cardiology* **107**(5), 437–443.

Rauch, G., Kunzmann, K., Kieser, M., Wegscheider, K., König, J. & Eulenburg, C. (2018), 'A weighted combined effect measure for the analysis of a composite time-to-first-event endpoint with components of different clinical relevance', *Statistics in Medicine* **37**(5), 749–767.

Rebolledo, R. (1980), 'Central limit theorems for local martingales', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **51**(3), 269–286.

Rondeau, V., Marzroui, Y. & Gonzalez, J. R. (2012), 'frailtypack: an r package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation', *Journal of Statistical Software* **47**, 1–28.

Rullier, E., Rouanet, P., Tuech, J.-J., Valverde, A., Lelong, B., Rivoire, M., Faucheron, J.-L., Jafari, M., Portier, G., Meunier, B. et al. (2017), 'Organ preservation for rectal cancer (greccar 2): a prospective, randomised, open-label, multicentre, phase 3 trial', *The Lancet* **390**(10093), 469–479.

Ryalen, P. C., Stensrud, M. J. & Røysland, K. (2018), 'Transforming cumulative hazard estimates', *Biometrika* **105**(4), 905–916.

Schmidli, H., Roger, J. H. & Akacha, M. (2023), 'Estimands for recurrent event endpoints in the presence of a terminal event', *Statistics in Biopharmaceutical Research* **15**(2), 238–248.

Stensrud, M. J., Hernán, M. A., Tchetgen Tchetgen, E. J., Robins, J. M., Didelez, V. & Young, J. G.

(2021), 'A generalized theory of separable effects in competing event settings', *Lifetime Data Analysis* **27**(4), 588–631.

Stensrud, M. J., Robins, J. M., Sarvet, A., Tchetgen Tchetgen, E. J. & Young, J. G. (2023), 'Conditional separable effects', *Journal of the American Statistical Association* **118**(544), 2671–2683.

Stensrud, M. J., Young, J. G., Didelez, V., Robins, J. M. & Hernán, M. A. (2022), 'Separable effects for causal inference in the presence of competing events', *Journal of the American Statistical Association* **117**(537), 175–183.

Toenges, G. & Jahn-Eimermacher, A. (2020), 'Computational issues in fitting joint frailty models for recurrent events with an associated terminal event', *Computer methods and programs in biomedicine* **188**, 105259.

Vasbinder, A., Anderson, E., Shadid, H., Berlin, H., Pan, M., Azam, T. U., Khaleel, I., Padalia, K., Meloche, C., O'Hayer, P. et al. (2022), 'Inflammation, hyperglycemia, and adverse outcomes in individuals with diabetes mellitus hospitalized for covid-19', *Diabetes Care* **45**(3), 692–700.

Vickers, A. J. & Altman, D. G. (2001), 'Analysing controlled trials with baseline and follow-up measurements', *British Medical Journal* **323**(7321), 1123–1124.

Wei, J., Muetze, T., Jahn-Eimermacher, A. & Roger, J. (2023), 'Properties of two while-alive estimands for recurrent events and their potential estimators', *Statistics in Biopharmaceutical Research* **15**(2), 257–267.

Zannad, F., Garcia, A. A., Anker, S. D., Armstrong, P. W., Calvo, G., Cleland, J. G., Cohn, J. N., Dickstein, K., Domanski, M. J., Ekman, I. et al. (2013), 'Clinical outcome endpoints in heart failure trials: a european society of cardiology heart failure association consensus document', *European Journal of Heart Failure* **15**(10), 1082–1094.

Zeng, D. & Cai, J. (2010), 'A semiparametric additive rate model for recurrent events with an informative terminal event', *Biometrika* **97**(3), 699–712.