Estimands in Randomized Clinical Trials with Complex Life History Processes

by

Alexandra Bühler

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics (Biostatistics)

Waterloo, Ontario, Canada, 2024

© Alexandra Bühler 2024

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Jong-Hyeon Jeong
Mathematical Statistician
National Cancer Institute, Rockville

Supervisors: Richard J. Cook
University Professor
Dept. of Statistics and Actuarial Science, University of Waterloo

Jerald F. Lawless
Distinguished Professor Emeritus
Dept. of Statistics and Actuarial Science, University of Waterloo

Internal Members: Leilei Zeng
Professor
Dept. of Statistics and Actuarial Science, University of Waterloo

Michael Wallace
Associate Professor
Dept. of Statistics and Actuarial Science, University of Waterloo

Internal-External Member: William W. L. Wong
Associate Professor
School of Pharmacy, University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The chapters of this doctoral thesis describe and contain Alexandra Bühler's work, under the supervision of Professors Richard J. Cook and Jerald F. Lawless, that has been published or submitted for publication in the following peer-reviewed journals.

Ch. 2 Bühler, A., Cook, R. J., and Lawless, J. F. (2023). Multistate models as a framework for estimand specification in clinical trials of complex processes. *Statistics in Medicine*, 42(9):1368–1397; https://doi.org/10.1002/sim.9675.

Ch. 3 Bühler, A., Cook, R. J., and Lawless, J. F. (2024a). Estimands and cumulative incidence function regression in clinical trials: some new results on interpretability and robustness. *Under minor revision for Statistics in Medicine*; arXiv: https://arxiv.org/abs/2401.04863.

Ch. 4 Bühler, A., Cook, R. J., and Lawless, J. F. (2024b). Reconsidering mean function regression in clinical trials with recurrent and terminal events: concerns regarding the interpretability of estimands. *Manuscript under preparation.*

Ch. 5 Bühler, A., Cook, R. J., and Lawless, J. F. (2024c). Combining multistate models and utility functions for the specification of estimands in complex disease processes. *Manuscript under preparation.*

Abstract

Clinical trials in oncology, cardiovascular disease and many other settings are dealing with complex outcomes involving multiple endpoints, competing or semi-competing risks, loss to follow-up and cointerventions related to the management and care of patients. These all complicate the design, analysis and interpretation of randomized trials. In such settings, traditional analyses of the time to some event are not sufficient for assessing new treatments.

This thesis discusses the issues involved in the specification of estimands within a comprehensive multistate model framework. Intensity-based multistate models are used to (a) conceptualize the event-generating process involving primary and post-randomization outcomes, to (b) define and interpret causal estimands based on observable marginal features of the process, and to (c) conduct secondary analyses of marginal treatment effects. For a broad range of disease process settings, we investigate how marginal estimands depend on the full intensity-based process. Using large sample theory, factors influencing the limiting values of estimators of treatment effect in generalized linear models for marginal process features are studied. Rejection rates of a variety of hypothesis tests based on marginal regression models are also examined in terms of the true intensity-based process; based on these findings robustness properties are established. Such numerical investigations give insights into the interpretation and use of marginal estimands in randomized trials. We discuss in detail estimands based on cumulative incidence function regression for semi-competing risks processes and mean function regression for processes involving recurrent and terminal events.

Specification of utilities for different disease-related outcomes, rescue interventions and other post-randomization events facilitate synthesis of information on complex disease processes, enabling simple causal treatment comparisons. Derivations are provided of an infinitesimal jackknife variance estimator for utility-based estimands to facilitate robust methods for causal inference of randomized trials.

Illustrations for cancer, cardiovascular and platelet transfusion trials are provided to demonstrate the usefulness of the multistate model framework for primary and secondary analyses of randomized trials.

Acknowledgments

I would like to express my gratitude to all those who encouraged, inspired, accompanied and helped me throughout my doctoral studies in Biostatistics at the University of Waterloo.

First and foremost, I feel exceptionally fortunate to have had *Dr. Richard J Cook* and *Dr. Jerry F. Lawless* as my PhD supervisors. Their immense knowledge, enthusiasm and extensive experience in academic research have significantly shaped my research skills over the past four years. I sincerely thank Jerry and Richard for their tremendous commitment to guiding me through my PhD studies. Their supervision was perfect just as it was - many thanks for making this thesis possible!

I would like to extend my thanks to my committee members *Dr. Leilei Zeng*, *Dr. Michael Wallace*, *Dr. Jong-Hyeon Jeong* and *Dr. William Wong* for taking the time to review my thesis and for providing helpful suggestions. Thanks to *Dr. Terry Therneau* for the fruitful correspondence regarding the R survival package. I would like to thank Professors *Kevin Thorpe* and *Wayne Taylor* for permission to use the NASCET data, and Professor *Nancy Heddle* from McMaster University for permission to use the PREPAReS trial data. I would also like to thank Novartis for permission to use the data from the trial of patients with bone metastases.

I want to thank the entire *SAS department* for providing an excellent research and learning environment - being involved with this department has been a great pleasure! Thanks to *Mary Lou Dufton, Carla Daniels, Greg Preston* and *Carlos Mendes* for their help in administrative issues. I will always be grateful to *Ker-Ai Lee* for sharing her experience with statistical computing.

A special thanks goes to *Dr. Audrey Boruvka* for connecting me to Richard and Jerry. It was Audrey who encouraged me in 2019 to pursue a PhD studies at the University of Waterloo. I am also thankful for the insights *Dr. Oana Danila* and *Dr. Adrian Waddell* provided about their PhD experience at Waterloo, which contributed to my decision to pursue this path as well. *Dr. Jan Beyersmann* also deserves thanks for the great support I received throughout my Bachelor's and Master's studies at the University of Ulm, and for introducing me to the analysis of survival and event history data.

A big thanks to *Mahsa Panahi* and *Jianchu Chen* for making our office, M4101, a fun place to be. I also want to acknowledge *Chris Salahub*, *Kecheng Li*, *Xianwei Li* and *Laura Bumbulis* for adding another layer of fun to our office building.

Anita Santin and *Way Tang* deserve special recognition for their incredible guidance and support since my arrival in Canada. I am very grateful for their friendship. I cannot thank Anita enough for her warm hospitality, always offering me a lovely place to stay in Waterloo whenever I needed it.

I appreciate the support I have received from my long-time friends in Germany: *Pia Hach, Isabelle Bührle, Lisa Zeifang, Lisa Siegel, Tanja Frank* and *Caro Barth*. I am grateful to have met *Felice Yeung*, who enriched my life in Toronto with her friendship - thanks for our work sessions!

I am incredibly indebted to my family, especially my parents, *Anita* and *Hans-Jürgen Bühler*, who supported me in every way possible and gave me the strength to keep going in times of need. Many thanks to my siblings, *Katrin* and *Stefan*, for providing an additional layer of support.

I would like to extend a heartfelt thanks to *Heike* and *Alex Pohl* not only for always being invested in my PhD progress but especially for their several visits to Canada, which brought me joy and a sense of home.

Finally, I would like to express my greatest appreciation to *Dennis Pohl*. There are literally no words that could sufficiently describe the love, support, encouragement, patience and flexibility he has shown me during the most challenging times of my PhD. Without his endless care, I doubt I could have made it over the final stretch. Thanks for spending four and a half wonderful years with me in Canada and for the beautiful memories we have created together!

Dedication

To my wonderful parents.

Table of Contents

Examining Committee Membership **ii**

Author's Declaration **iii**

Statement of Contributions **iv**

Abstract **v**

Acknowledgments **vi**

Dedication **viii**

List of Figures **xvi**

List of Tables **xxiii**

List of Abbreviations **xxviii**

1 Introduction **1**

1.1 Statistical challenges in late stage clinical trials 2

1.1.1 Trials with (disease-related) death as primary event of interest (I) 3

1.1.2 Trials with a non-fatal primary event of interest (II) 4
1.1.3 Intercurrent events in type (I) and (II) trial settings 4
1.1.4 The utility of multistate modeling in analyzing clinical trials 5
1.2 A brief introduction to the estimand framework based on multistate models . . . 6
1.2.1 Multistate model formulation for disease processes 6
1.2.2 Marginal process features 8
1.2.3 Frameworks for the specification of marginal estimands 10
1.3 Motivating studies 11
1.3.1 Efficacy of liraglutide for cardiovascular outcomes in type II diabetes . . 11
1.3.2 Assessing carotid endarterectomy in stroke prevention 13
1.3.3 Prevention of skeletal complications in metastatic breast cancer 15
1.3.4 Quality of life in a therapeutic breast cancer trial 16
1.3.5 Assessing the burden of bleeding in thrombocytopenia 17
1.4 Research topics and outline of thesis 19
1.4.1 Theme 1: Principles for estimand specification 19
1.4.2 Theme 2: Estimands and tests based on common marginal features 20
1.4.3 Theme 3: Utility-based analyses for complex outcomes 22
1.4.4 Thesis structure 23

2 Multistate models as a framework for estimand specification in clinical trials of complex disease processes 24

2.1 Introduction 24
2.2 Estimands for interventions in event history processes 26
2.2.1 Multistate models, intensities and associated functionals 27
2.2.2 Principles for defining estimands 28

2.2.3 Problems with conditioning on features of the process history 32
2.2.4 General frameworks for expressing marginal treatment effects 37
2.3 Time-to-event framework: estimands for illness-death processes based on cumulative incidence functions 38
2.3.1 Generalized linear models for cumulative incidence functions 39
2.3.2 Illustrative calculations for cumulative incidence function regression . . . 41
2.4 Count framework: estimands for recurrent and terminal event processes based on marginal rate functions 46
2.5 Intercurrent events 47
2.5.1 Illustrative examples involving intercurrent events 50
2.5.2 Surgical prevention of stroke-related events in the NASCET trial 55
2.6 Remarks on potential outcomes 60
2.7 Discussion and recommendations 62

3 Cumulative incidence function regression in clinical trials: results on interpretability of estimands and robustness of tests 64
3.1 Introduction 64
3.2 Model formulation and notation 66
3.3 Cumulative incidence function regression under misspecification 68
3.3.1 Estimation methods and large sample results 68
3.3.2 Limiting values of FG and DB estimators under intensity-based processes 74
3.3.3 Goodness of fit 77
3.4 Robustness of tests 78
3.4.1 Formulation of testing framework and robustness property for Wald tests 78
3.4.2 Simulation studies 82
3.5 Prevention of skeletal complications in metastatic breast cancer 92
3.6 Discussion and recommendations 96

4 Mean function regression in clinical trials involving recurrent and terminal event processes: interpretability of estimands **99**

4.1 Introduction 99

4.2 Notation and marginal models for mean functions 102

4.2.1 Marginal rate functions in recurrent event processes terminated by death . 103

4.2.2 Marginal rate functions with recurrent-terminal composite events and competing risks 105

4.3 Mean function regression under misspecification 106

4.3.1 Ghosh-Lin and Mao-Lin methods of estimation 107

4.3.2 Estimation based on weighted pseudo-Poisson estimating functions . . . 109

4.3.3 Empirical studies of finite sample properties 112

4.4 Limiting values of estimator of treatment effect under particular data-generating mechanisms 115

4.4.1 The Ghosh-Lin estimator 120

4.4.2 The Mao-Lin estimator 135

4.5 Prevention of recurrent skeletal complications in metastatic breast cancer 140

4.6 Discussion and recommendations 145

5 Methods based on utility functions and ranked outcomes **148**

5.1 Introduction 148

5.2 Notation and utility functions 150

5.3 A critical discussion of estimands based on utility functions and ranked outcomes 153

5.3.1 Win ratio 154

5.3.2 Restricted mean time in favor of treatment 162

5.3.3 Unification of estimands through utilities 165

5.4 More general use of utilities . . . 167
5.4.1 Variance estimation using the jackknife procedure . . . 169
5.4.2 Empirical studies of finite sample properties . . . 176
5.5 Bleeding in chemotherapy-induced thrombocytopenia . . . 185
5.6 Discussion . . . 192

6 Conclusions and future work **196**
6.1 Overview and contributions of thesis . . . 196
6.2 Future research on cumulative incidence function (CIF) regression . . . 198
6.2.1 CIF regression and intermittent observation . . . 198
6.2.2 Multinomial logistic CIF regression . . . 200
6.3 Future research on recurrent and terminal events . . . 202
6.3.1 Mean function regression and intermittent observation . . . 202
6.3.2 Estimands for disease processes involving recurrent episodes and fatal events . . . 202
6.4 Future research on utility-based estimands . . . 204

References **207**

APPENDICES **232**

A Appendix to Chapter 2 **233**
A.1 Adequacy of the cloglog transformed CIF model under time-inhomogeneous intensities . . . 233

B Appendix to Chapter 3 **236**

B.1 Estimation of the censoring distribution and large sample results 236

B.1.1 Limiting behavior of the KM estimator under different true censoring mechanisms 237

B.1.2 Limiting behavior of the Cox estimator under misspecification 238

B.2 Limiting values of FG and DB treatment effect estimators under intensity-based processes 239

B.2.1 Supplementary results for the numerical investigation of Section 3.3.2 . . 239

B.2.2 Limiting values under time-inhomogeneous intensities 244

B.3 Proof of the robustness property 247

B.3.1 FG-based Wald test 247

B.3.2 DB-based Wald test 248

B.4 Additional results for the simulation study of Section 3.4.2 250

C Appendix to Chapter 4 **257**

C.1 Details on the pseudo-Poisson procedures of Section 4.3.2 257

C.1.1 Software: an illustrative dataframe 257

C.2 Derivation of the cumulative mean function under scenario GL-1-2 260

C.3 Limiting values of treatment effect estimators in mean function regression under different data-generating frameworks 261

C.3.1 The Ghosh-Lin estimator 261

C.3.2 The Mao-Lin estimator 261

C.4 Additional results for the application of Section 4.5 266

D Appendix to Chapter 5 **269**

D.1 Details on the win ratio approach of Section 5.3.1 269

D.1.1 Proof of equality of win probabilities 269

D.1.2 Derivation of win probabilities for illness-death processes 271

D.1.3 Simulation study 273

D.2 Reformulation of the utility-based estimand 274

D.3 Illustration of the recursive infinitesimal jackknife variance formula for $\widehat{D}(\tau)$ in illness-death processes 276

D.3.1 Expressions for $\widehat{I}_{ik}(t_{(s)})$ at $t_{(1)}$ 279

D.3.2 Expressions for $\widehat{I}_{ik}(t_{(s)})$ at $t_{(2)}$ 279

D.4 Software for the infinitesimal jackknife procedure: an illustrative example for the illness-death model 282

D.5 Additional results for the simulation study of Section 5.4.2 286

D.6 Additional results for the application of Section 5.5 291

List of Figures

1.1 Multistate diagrams for chronic disease processes. 9

1.2 A multistate model depicting the occurrence of stroke, switch from medical care to carotid endarterectomy, stroke-related death and non-stroke related death for the NASCET trial reported on by Barnett et al. (1998). 14

1.3 Timeline diagrams depicting the occurrence of different skeletal complications and survival or censoring times in breast cancer patients with bone metastases from Theriault et al. (1999). 15

1.4 A multistate model for bleeding severity, recovery from thrombocytopenia and death in the PREPAReS trial reported on by van der Meer et al. (2018). 18

2.1 Asymptotic percent relative difference $100(\phi_1^* - \gamma_{1x})/\gamma_{1x}$ where ϕ_1^* is the limiting value of the estimator from a cause-specific Cox model for the $0 \longrightarrow 1$ transition when the covariate V is omitted; $C = 1, P(X = 1) = 0.5, P(V = 1) = 0.5, P(T \leq C) = 0.6, \gamma_{1x} = \log(0.75), \gamma_{2x} = \log(0.9)$. 36

2.2 Plots of $IR_{01}(t)$ when the cause-specific hazards functions $\lambda_{01}(t|X, V)$ and $\lambda_{02}(t|X, V)$ depend on both X and V; $C = 1, P(X = 1) = 0.5, P(V = 1) = 0.5, P(T \leq C) = 0.6, \gamma_{1x} = \log(0.75), \gamma_{2x} = \log(0.9)$. 37

2.3 Plots of $g(F_1(s_r|X = 1)) - g(F_1(s_r|X = 0))$ for 20 equi-spaced values of s_r in $(0, 1)$ and different values of $P(T \leq C), P(T_1 < T_2|T \leq C)$ and γ_2 under the log and cloglog transformation models; $C = 1, P(X = 1) = 0.5, \gamma_1 = \log(0.75)$. 43

2.4 Multistate diagrams for the three-state illness-death model and joint models for illness-death and type 1 and 2 intercurrent event processes. 54

2.5 Multistate models for a particular process of interest and an expanded process incorporating a post-randomization intervention for the control arm. 54

2.6 Multistate diagrams for analyses of stroke or stroke-related death and non-stroke related death in the NASCET trial (Barnett et al., 1998). 56

2.7 Nonparametric Nelson-Aalen estimates of the cumulative intensities for selected transitions based on Figure 2.6(a) and Figure 2.6(b). 57

2.8 Nonparametric Aalen-Johansen estimates for the cumulative incidence functions for stroke/stroke-death and non-stroke death in Figure 2.6(a) and Figure 2.6(b). . 59

3.1 Contours showing the limiting values $\exp(\beta^{\star}_{FG})$ and $\exp(\beta^{\star}_{DB})$ as a function of $\exp(\gamma_2)$ and $P(T_1 < T_2 | T \leq \tau, X = 0)$ when the true process has intensities $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X)$ for $\exp(\gamma_1) = 1, 0.75$ and 0.6; $\tau = 1, P(T \leq \tau | X = 0) = 0.6$, $G^{\star} = G$. 76

3.2 Relationships between different null hypotheses in a competing risks process. . . 79

3.3 Empirical type I error rates of tests T^{F_1} under $H_0^{\lambda_1 \lambda_2}$ and $H_0^{\lambda_1}$ for different values of $\exp(\gamma_2)$ and $P(T_1 < T_2 | T \leq \tau, X = 0)$, nominal test size is $\omega = 0.05; n = 1000$ individuals, $n_{sim} = 10000$, $\tau = 1, P(T \leq \tau | X = 0) = 0.6, \pi_r = 0.2$. 87

3.4 Empirical rejection rates for different statistical tests under various alternatives; $n = 1000$ individuals, $n_{sim} = 10000$, $\tau = 1$, $P(T \leq \tau | X = 0) = 0.6, P(T_1 < T_2 | T \leq \tau, X = 0) = 0.4, 0.6$ and 0.8, $\pi_r = 0.2$. 90

3.5 Comparisons of nonparametric Aalen-Johansen estimates of the cumulative incidence function for skeletal-related events and semiparametric estimates implied by the Fine-Gray method and direct binomial estimation under the cloglog transformation model, and plots of $g(\hat{F}_1^{AJ}(t|X = 1)) - g(\hat{F}_1^{AJ}(t|X = 0))$ for $g(\cdot) = \text{cloglog}(\cdot), \log(\cdot)$ and $\text{logit}(\cdot)$ along with pointwise 95% CIs. 93

3.6 Nonparametric Nelson-Aalen estimates of the cumulative intensity functions and Aalen-Johansen estimates of the cumulative incidence functions for skeletal complications and complication-free death for breast cancer patients in Hortobagyi et al. (1996). 94

4.1 *GL-1-1:* Plots of $\mu(t|X=1)/\mu(t|X=0)$ for $\exp(\gamma)=0.75$ and different values of $P(T\leq\tau|X=0), \exp(\delta)$ and $\mathbb{E}(N(\tau)|X=0)$ when the true process has intensities $\lambda(t|X)=\lambda_0\exp(\gamma X)$ and $\rho(t|X)=\rho_0\exp(\delta X)$; $\tau=1,\kappa=1$. . . . 122

4.2 *GL-1-1:* Plots of $\mu(t|X=1)/\mu(t|X=0)$ for $\exp(\gamma)=0.75$ and different values of $\kappa\geq 1, P(T\leq\tau|X=0), \exp(\delta)$ and $\mathbb{E}(N(\tau)|X=0)$ when the true process has intensities $\lambda(t|X)=\lambda_0^\kappa\kappa t^{\kappa-1}\exp(\gamma X)$ and $\rho(t|X)=\rho_0\exp(\delta X)$; $\tau=1$. . . 123

4.3 *GL-1-1:* Limiting values $\exp(\beta_{GL}^\star)$ and $\exp(\beta_P^\star)$ as a function of $P(T\leq\tau|X=0)$ for different values of $\exp(\gamma), \exp(\delta)$ and $\mathbb{E}(N(\tau)|X=0)$ when the true process has intensities $\lambda(t|X)=\lambda_0\exp(\gamma X)$ and $\rho(t|X)=\rho_0\exp(\delta X)$; $\kappa=1,\tau=1,R=6,G^\star=G$. . . . 125

4.4 *GL-1-1:* Dependency of limiting values $\beta_{GL}^\star$ and $\beta_P^\star$ on κ for different values of $\exp(\delta)$ and $P(T\leq\tau|X=0)$; $\tau=1,R=6,G^\star=G$. . . . 126

4.5 *GL-1-2:* Limiting values $\exp(\beta_{GL}^\star)$ as a function of b for different values of $P(T\leq\tau|X=0)$ and $\mathbb{E}(N(\tau)|X=0)$ when the true process has intensities $\lambda_k(t|X)=\lambda(t|X)=\lambda_0\exp(\gamma X)$ and $\rho_k(t|X)=b^k\rho_0\exp(\delta X)$ for $\exp(\gamma)=0.75$ and $\exp(\delta)=1$; $K=20,\kappa=1,\tau=1,G^\star=G$. . . . 127

4.6 *GL-2*: Plots of $\mathbb{E}(dN(t)|\tilde{\mathcal{H}}(t))$, $\mathbb{E}(dN(t)|T\geq t,X)$ and $\mathbb{E}(dN_D(t)|\tilde{\mathcal{H}}(t))$ for $X=0$ and $\text{Var}(U)=0.001,0.1,0.4$ when $P(T\leq\tau|X=0)=0.3$ and $\mathbb{E}(N(\tau)|X=0)=1.5$; $\tau=1,\kappa=1$. . . . 131

4.7 *GL-2*: Plots of $\mathbb{E}(U|\bar{N}(t),T\geq t,X=0)$ versus time t for different values of $\text{Var}(U)$; $\tau=1,\kappa=1,P(T\leq\tau|X=0)=0.3$ and $\mathbb{E}(N(\tau)|X=0)=1.5$. . . . 132

4.8 *GL-2:* Contour plots showing the dependence of the limiting values $\exp(\beta_{GL}^\star)$ and $\exp(\beta_P^\star)$ on $\exp(\delta)$ and $P(T\leq\tau|X=0)$ for $\text{Var}(U)=0.1,0.4$ when the true process has intensities $\lambda(t|X,U)=U\lambda_0(t)\exp(\gamma X)$ and $\rho(t|X,U)=U\rho_0\exp(\delta X)$ for $\exp(\gamma)=0.75$ and $\kappa=1$, $G^\star=G$. . . . 134

4.9 *ML-1-1:* Plots of $\mu^{\dagger}(t|X=1)/\mu^{\dagger}(t|X=0)$ for $\exp(\gamma)=0.85$ and different values of $P(T\leq\tau|X=0), P(T_D<T'_D|T\leq\tau,X=0), \exp(\delta)$ and $\mathbb{E}(N(\tau)|X=0)$ when the true process has intensities $\lambda(t|X)=\lambda_0\exp(\gamma X), \rho(t|X)=\rho_0\exp(\delta X)$ and $\psi(t|X)=\psi_0$; $\tau=5, \kappa=1$. 137

4.10 *ML-1-1:* Plots showing the dependency of the limiting values $\exp(\beta^{\dagger\star}_{ML})$ on treatment and the time-homogeneous process intensities $\lambda(t|X)=\lambda_0\exp(\gamma X), \rho(t|X)=\rho_0\exp(\delta X)$ and $\psi(t|X)=\psi_0$; $\kappa=1, \tau=5, P(T\leq\tau|X=0)=0.12, G^{\star}=G$. . 139

4.11 Nonparametric Nelson-Aalen estimates of the cumulative intensity functions for the occurrence of new skeletal-related events, death and censoring for placebo patients in Hortobagyi et al. (1996). 141

4.12 Unweighted and weighted nonparametric Kaplan-Meier and Aalen-Johansen estimates of the cumulative probability for death. 143

4.13 Unweighted and weighted estimates of the cumulative mean function for skeletal-related events for breast cancer patients in Hortobagyi et al. (1996). 144

5.1 Illustrative multistate diagrams for progressive disease processes. 152

5.2 Plots of the win probability $p_{ij}^{ID}(\tau)$ and its 2 (left panel) or 5 components (right panel) versus τ when the true illness-death process has time-homogeneous intensities $\lambda_{01}=1, \lambda_{02}=0.75$ and $\lambda_{01}=0.5$ (moderate mortality scenario). 159

5.3 Plots of the win probability $p_{ij}^{ID}(\tau)$ and its 2 (left panel) or 5 components (right panel) versus τ when the true illness-death process has time-homogeneous intensities $\lambda_{01}=1, \lambda_{02}=0.1$ and $\lambda_{01}=0.1$ (low mortality scenario). 160

5.4 Plots of the win probability $p_{ij}^{ID}(\tau)$ and its 2 (left panel) or 5 components (right panel) versus τ when the true illness-death process has time-homogeneous intensities $\lambda_{01}=1, \lambda_{02}=1.5$ and $\lambda_{01}=1.5$ (high mortality scenario). 161

5.5 Illustration of the RMT-IF (5.3.7) and the utility-based estimand (5.3.8) in a three-state illness-death process using sample paths of 4 hypothetical individuals. 163

5.6 Plots of $P_{0k}(0,t|X)$ and $D(t|X)$ when the true illness-death process is Markov or mixed-Markov with $\exp(\gamma_{01}) = 0.75, \exp(\gamma_{02}) = 0.8$ and $\exp(\gamma_{12}) = 1$, $k = 0, 1, 2$ and $X = 0, 1$. 177

5.7 Plots of $P_{0k}(0,t|X)$ and $D(t|X)$ when the true illness-death process is Markov or mixed-Markov with $\exp(\gamma_{01}) = 1, \exp(\gamma_{02}) = 0.8$ and $\exp(\gamma_{12}) = 0.8$, $k = 0, 1, 2$ and $X = 0, 1$. 178

5.8 Nonparametric estimates of the expected cumulative utility score $D(t)$ for different utility scores u_1 and u_2. 188

5.9 Plots of $\hat{\beta}(t) = \hat{D}(t|X = 1) - \hat{D}(t|X = 0)$ for $u_1 = 0.7, u_2 = 0.3$ and $u_1 = 0.9, u_2 = 0.1$. 189

5.10 Nonparametric Aalen-Johansen estimates of state occupancy probabilities $P_{0k}(0,t)$ and pointwise 95% CIs by transfusion type. 190

5.11 Nonparametric Nelson-Aalen estimates of the cumulative intensity functions for transitions out of the transient states 0, 1 and 2 of Figure 1.4. 191

5.12 Stacked plots of the nonparametric Aalen-Johansen estimates of the state occupancy probabilities for no or grade 1 bleeding, grade 2 bleeding, grade 3+4 bleeding and death by transfusion type. 193

6.1 Multistate diagrams for processes involving recurrent episodes and terminal events. 203

6.2 Multistate diagram for recurrent hospitalizations and death. 205

A.1 Plots of $\lambda_{01}(t|X = 0), \lambda_{02}(t|X = 0)$ and $F_1(t|X = 0)$ for different values of κ_1 when the true competing risks process has baseline intensities of Weibull form $\lambda_k(t) = \kappa_k \lambda_k (\lambda_k t)^{\kappa_k - 1}$; $\kappa_2 = 1, \tau = 1, P(T \leq \tau | X = 0) = 0.6, P(T_1 < T_2 | T \leq \tau, X = 0) = 0.6$. 234

A.2 Plots of $g(F_1(t|X = 1)) - g(F_1(t|X = 0))$ for $t \in (0, 1)$ and different values of κ_1 and $\exp(\gamma_2)$ under the cloglog transformation model; $g(u) = \log(-\log(1 - u)), \kappa_2 = 1, \tau = 1, P(T \leq \tau | X = 0) = 0.6, \gamma_1 = \log(0.75)$. 235

B.1 Limiting values $\exp(\beta^{\star}_{FG})$ and $\exp(\beta^{\star}_{DB})$ as a function of $\exp(\gamma_2)$ for different values of $\exp(\gamma_1)$ and $P(T_1 < T_2|T \leq \tau, X = 0)$ when the true process has intensities $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X)$; $\tau = 1, P(T \leq \tau|X = 0) = 0.6$, $G^{\star} = G$. . 240

B.2 Limiting values $\exp(\beta^{\star}_{FG})$ and $\exp(\beta^{\star}_{DB})$ as a function of $P(T_1 < T_2|T \leq \tau, X = 0)$ for different values of $\exp(\gamma_2)$ and $\exp(\gamma_1)$ when the true process has intensities $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X)$; $\tau = 1, P(T \leq \tau|X = 0) = 0.6, G^{\star} = G$. 241

B.3 Comparisons of $F_1(t|X)$, $F^{\star}_{1,FG}(t|X) = h(\alpha^{\star}_{FG}(t) + \beta^{\star}_{FG}X)$ and $F^{\star}_{1,DB}(s_r|X) = h(\alpha^{\star}_{rDB} + \beta^{\star}_{DB}X), X = 0, 1$ when the true process has intensities $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X)$ for $\exp(\gamma_1) = 1, \exp(\gamma_2) = 0.5, 0.8, 1, 1.5$ and $P(T_1 < T_2|T \leq \tau, X = 0) = 0.6, 0.2$; $\tau = 1, P(T \leq \tau|X = 0) = 0.6, G^{\star} = G$. 242

B.4 Comparisons of $F_1(t|X)$, $F^{\star}_{1,FG}(t|X) = h(\alpha^{\star}_{FG}(t) + \beta^{\star}_{FG}X)$ and $F^{\star}_{1,DB}(s_r|X) = h(\alpha^{\star}_{rDB} + \beta^{\star}_{DB}X), X = 0, 1$ when the true process has intensities $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X)$ for $\exp(\gamma_1) = 0.75, \exp(\gamma_2) = 0.5, 0.8, 1, 1.5$ and $P(T_1 < T_2|T \leq \tau, X = 0) = 0.6, 0.2$; $\tau = 1, P(T \leq \tau|X = 0) = 0.6, G^{\star} = G$. 243

B.5 Limiting values $\exp(\beta^{\star}_{FG})$ of the FG estimator as a function of κ_1 for different values of $\exp(\gamma_2)$ and $P(T_1 < T_2|T \leq \tau, X = 0)$ when the true process has Weibull intensities $\lambda_{0k}(t|X) = \kappa_k \lambda_k (\lambda_k t)^{\kappa_k - 1} \exp(\gamma_k X)$ with $\exp(\gamma_1) = 1, 0.75$; $\kappa_2 = 1, \tau = 1, P(T \leq \tau|X = 0) = 0.6, G^{\star} = G$. 245

B.6 Limiting values $\exp(\beta^{\star}_{DB})$ of the DB estimator as a function of κ_1 for different values of $\exp(\gamma_2)$ and $P(T_1 < T_2|T \leq \tau, X = 0)$ when the true process has Weibull intensities $\lambda_{0k}(t|X) = \kappa_k \lambda_k (\lambda_k t)^{\kappa_k - 1} \exp(\gamma_k X)$ for $\exp(\gamma_1) = 1, 0.75$; $\kappa_2 = 1, R = 6, \tau = 1, P(T \leq \tau|X = 0) = 0.6, G^{\star} = G$. 246

C.1 *GL-1-1:* Plots of $\mu(t|X = 1)/\mu(t|X = 0)$ for $\exp(\gamma) = 0.55$ and different values of $P(T \leq \tau|X = 0), \exp(\delta)$ and $\mathbb{E}(N(\tau)|X = 0)$ when the true process has intensities $\lambda(t|X) = \lambda_0 \exp(\gamma X)$ and $\rho(t|X) = \rho_0 \exp(\delta X)$; $\kappa = 1, \tau = 2$. . . . 262

C.2 *GL-1-2:* Plots of $\mu(t|X = 1)/\mu(t|X = 0)$ for $\exp(\gamma) = 0.75, \exp(\delta) = 1$ and different values of $P(T \leq \tau|X = 0), b$ and $\mathbb{E}(N(\tau)|X = 0)$ when the true process has intensities $\lambda_k(t|X) = \lambda(t|X) = \lambda_0 \exp(\gamma X)$ and $\rho_k(t|X) = b^k \rho_0 \exp(\delta X)$; $K = 20, \tau = 1, \kappa = 1$. 263

C.3 *GL-1-2:* Limiting values $\exp(\beta_P^\star)$ as a function of b for different values of $P(T \leq \tau|X=0)$ and $\mathbb{E}(N(\tau)|X=0)$ when the true process has intensities $\lambda_k(t|X) = \lambda(t|X) = \lambda_0 \exp(\gamma X)$ and $\rho_k(t|X) = b^k \rho_0 \exp(\delta X)$ for $\exp(\gamma) = 0.75$ and $\exp(\delta) = 1$; $K = 20, a = 1, \kappa = 1, \tau = 1$, $G^\star = G$. 264

C.4 *ML-1-1:* Plots showing the dependency of the limiting values $\exp(\beta_P^{\dagger\star})$ on treatment and the time-homogeneous process intensities $\lambda(t|X) = \lambda_0 \exp(\gamma X)$, $\rho(t|X) = \rho_0 \exp(\delta X)$ and $\psi(t|X) = \psi_0$; $\kappa = 1, \tau = 5, P(T \leq \tau|X=0) = 0.12, G^\star = G$. . 265

C.5 Nonparametric Nelson-Aalen estimates of the cumulative intensity functions for the occurrence of new skeletal-related events, death and censoring for pamidronate patients in Hortobagyi et al. (1996). 266

C.6 Nonparametric Nelson-Aalen estimates of the cumulative intensity functions for the occurrence of new skeletal-related events (Hortobagyi et al., 1996). 267

C.7 Nonparametric Nelson-Aalen estimates of the cumulative intensity functions for the occurrence of death (Hortobagyi et al., 1996). 268

D.1 Multistate diagram of a four-state illness-death process. 277

D.2 Nonparametric Nelson-Aalen estimates of the cumulative intensity functions for transitions out of state 0 in Figure 1.4 (van der Meer et al., 2018). 291

List of Tables

2.1 Asymptotic and empirical properties of Fine-Gray and direct binomial regression estimators for β in cloglog transformation models for $F_1(t|X)$; $C = 1, P(X = 1) = 0.5, \gamma_1 = \log(0.75), \gamma_2 = \log(0.9), n = 1000$ individuals, $n_{sim} = 2000$. . . . 45

3.1 Empirical rejection rates for different statistical tests of nominal size $\omega = 0.05$ when the true process has intensities $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X)$, along with the limiting values $\beta^{\star}_{FG}$ and $\beta^{\star}_{DB}$; $n = 1000$ individuals, $n_{sim} = 10000$, $\tau = 1$, $P(T \leq \tau|X = 0) = 0.6, P(T_1 < T_2|T \leq \tau, X = 0) = 0.6, \pi_r = 0.2$. 88

3.2 Empirical rejection rates for different statistical tests of nominal size $\omega = 0.05$ when the true process is implied by models (3.4.7)-(3.4.8) and (3.4.9)-(3.4.10), along with the limiting values $\beta^{\star}_{FG}$ and $\beta^{\star}_{DB}$; $n = 1000$ individuals, $n_{sim} = 10000$, $\tau = 1, F_1(\tau|X = 0) = 0.36, F_2(\tau|X = 0) = 0.27, \pi_r = 0.2$. 89

3.3 Goodness-of-fit-tests for checking the assumption of a time-constant treatment effect in the cloglog transformed CIF model (Hortobagyi et al., 1996). 95

4.1 Empirical properties of treatment effect estimators $\hat{\beta}$ from a Ghosh-Lin model fitted under different estimation procedures; $\tau = 1, \mathbb{E}(N(\tau)|X = 0) = 0.8, \pi_D = 0.15$ and $0.3, \exp(\gamma) = 0.75, \exp(\delta) = 1, \exp(\gamma_r) = 1, \hat{G} \longrightarrow G, n = 500$ individuals, $n_{sim} = 2000$. 116

4.2 Empirical properties of treatment effect estimators $\hat{\beta}$ from a Ghosh-Lin model fitted under different estimation procedures; $\tau = 1, \mathbb{E}(N(\tau)|X = 0) = 0.8, \pi_D = 0.15$ and $0.3, \exp(\gamma) = 0.75, \exp(\delta) = 0.8, \exp(\gamma_r) = 1, \hat{G} \longrightarrow G, n = 500$ individuals, $n_{sim} = 2000$. 117

4.3 Empirical properties of treatment effect estimators $\hat{\beta}$ from a Ghosh-Lin model fitted under different estimation procedures; $\tau = 1, \mathbb{E}(N(\tau)|X = 0) = 0.8, \pi_D = 0.15$ and $0.3, \exp(\gamma) = 0.75, \exp(\delta) = 1, \exp(\gamma_r) = 0.85, \hat{G} \longrightarrow G, n = 500$ individuals, $n_{sim} = 2000$. 118

4.4 Empirical properties of treatment effect estimators $\hat{\beta}$ from a Ghosh-Lin model fitted under different estimation procedures; $\tau = 1, \mathbb{E}(N(\tau)|X = 0) = 0.8, \pi_D = 0.15$ and $0.3, \exp(\gamma) = 0.75, \exp(\delta) = 0.8, \exp(\gamma_r) = 0.85, \hat{G} \longrightarrow G, n = 500$ individuals, $n_{sim} = 2000$. 119

5.1 Empirical properties of the utility-based estimator $\hat{\beta}(\tau)$ and associated tests of $H_0 : \beta(\tau) \leq 0$ vs. $H_1 : \beta(\tau) > 0$ when the illness-death process is Markov with intensities $\lambda_{kl}(t|X) = \lambda_{kl0} \exp(\gamma_{kl} X)$; $\tau = 1, \pi_T = 0.6, \pi_T^1 = 0.9, \pi_D = 0.2, \pi_r = 0.2$ and $0.4, \exp(\gamma_r) = 1, u_0 = 1, u_1 = 0.5, u_2 = 0, n = 100, 250, 1000$ individuals, $n_{sim} = 10000$. 181

5.2 Empirical properties of the utility-based estimator $\hat{\beta}(\tau)$ and associated tests of $H_0 : \beta(\tau) \leq 0$ vs. $H_1 : \beta(\tau) > 0$ when the illness-death process is Markov with intensities $\lambda_{kl}(t|X) = \lambda_{kl0} \exp(\gamma_{kl} X)$; $\tau = 1, \pi_T = 0.6, \pi_T^1 = 0.9, \pi_D = 0.2, \pi_r = 0.2$ and $0.4, \exp(\gamma_r) = 0.85, u_0 = 1, u_1 = 0.5, u_2 = 0, n = 100, 250, 1000$ individuals, $n_{sim} = 10000$. 182

5.3 Empirical properties of the utility-based estimator $\hat{\beta}(\tau)$ and associated tests of $H_0 : \beta(\tau) \leq 0$ vs. $H_1 : \beta(\tau) > 0$ when the illness-death process is mixed-Markov with conditional intensities $\lambda_{kl}(t|X, U) = U \lambda_{kl0} \exp(\gamma_{kl} X)$ for $\phi = 0.5$; $\tau = 1, \pi_T = 0.6, \pi_T^1 = 0.9, \pi_D = 0.2, \pi_r = 0.2$ and $0.4, \exp(\gamma_r) = 1, u_0 = 1, u_1 = 0.5, u_2 = 0, n = 100, 250, 1000$ individuals, $n_{sim} = 10000$. 183

5.4 Empirical properties of the utility-based estimator $\hat{\beta}(\tau)$ and associated tests of $H_0 : \beta(\tau) \leq 0$ vs. $H_1 : \beta(\tau) > 0$ when the illness-death process is mixed-Markov with conditional intensities $\lambda_{kl}(t|X,U) = U\lambda_{kl0}\exp(\gamma_{kl}X)$ for $\phi = 1$; $\tau = 1, \pi_T = 0.6, \pi_T^1 = 0.9, \pi_D = 0.2, \pi_r = 0.2$ and $0.4, \exp(\gamma_r) = 1, u_0 = 1, u_1 = 0.5, u_2 = 0, n = 100, 250, 1000$ individuals, $n_{sim} = 10000$. 184

5.5 Nonparametric Aalen-Johansen-based estimates of $D(\tau|X = 0), D(\tau|X = 1)$ and $\beta(\tau) = D(\tau|X = 1) - D(\tau|X = 0)$ for different utility scores u_1 and u_2, along with SEs obtained by the infinitesimal jackknife procedure (Therneau, 2024). 187

5.6 Nonparametric estimates of the restricted mean times spent in states 0, 1, 2 and 3 by transfusion type. 193

B.1 Empirical rejection rates for different statistical tests of nominal size $\omega = 0.05$ when the true process has intensities $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X)$, along with the limiting values $\beta_{FG}^\star$ and $\beta_{DB}^\star$; $n = 1000$ individuals, $n_{sim} = 10000$, $\tau = 1$, $P(T \leq \tau|X = 0) = 0.6, P(T_1 < T_2|T \leq \tau, X = 0) = 0.4, \pi_r = 0.2$. 251

B.2 Empirical rejection rates for different statistical tests of nominal size $\omega = 0.05$ when the true process has intensities $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X)$, along with the limiting values $\beta_{FG}^\star$ and $\beta_{DB}^\star$; $n = 1000$ individuals, $n_{sim} = 10000$, $\tau = 1$, $P(T \leq \tau|X = 0) = 0.6, P(T_1 < T_2|T \leq \tau, X = 0) = 0.8, \pi_r = 0.2$. 252

B.3 Empirical rejection rates for different statistical tests of nominal size $\omega = 0.05$ when the true process is implied by models (3.4.7)-(3.4.8) and (3.4.9)-(3.4.10), along with the limiting values $\beta_{FG}^\star$ and $\beta_{DB}^\star$; $n = 1000$ individuals, $n_{sim} = 10000$, $\tau = 1, F_1(\tau|X = 0) = 0.24$ or 0.48, $F_2(\tau|X = 0) = 0.39$ or $0.15, \pi_r = 0.2$. . . . 253

B.4 Empirical rejection rates for different statistical tests of nominal size $\omega = 0.05$ when the true process has intensities $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X)$, along with the limiting values $\beta_{FG}^\star$ and $\beta_{DB}^\star$; $n = 500$ individuals, $n_{sim} = 10000$, $\tau = 1$, $P(T \leq \tau|X = 0) = 0.6, P(T_1 < T_2|T \leq \tau, X = 0) = 0.6, \pi_r = 0.2$. 254

B.5 Empirical rejection rates for different statistical tests of nominal size $\omega = 0.05$ when the true process is implied by models (3.4.7)-(3.4.8) and (3.4.9)-(3.4.10), along with the limiting values $\beta^{\star}_{FG}$ and $\beta^{\star}_{DB}$; $n = 500$ individuals, $n_{sim} = 10000$, $\tau = 1, F_1(\tau|X = 0) = 0.36, F_2(\tau|X = 0) = 0.27, \pi_r = 0.2$. 255

B.6 Empirical rejection rates for different statistical tests of nominal size $\omega = 0.05$ when the true process is implied by models (3.4.7)-(3.4.8) and (3.4.9)-(3.4.10), along with the limiting values $\beta^{\star}_{FG}$ and $\beta^{\star}_{DB}$; $n = 250$ individuals, $n_{sim} = 10000$, $\tau = 1, F_1(\tau|X = 0) = 0.36, F_2(\tau|X = 0) = 0.27, \pi_r = 0.2$. 255

B.7 Empirical rejection rates for different statistical tests of nominal size $\omega = 0.05$ when the true process has intensities $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X)$, along with the limiting values $\beta^{\star}_{FG}$ and $\beta^{\star}_{DB}$; $n = 250$ individuals, $n_{sim} = 10000$, $\tau = 1$, $P(T \leq \tau|X = 0) = 0.6, P(T_1 < T_2|T \leq \tau, X = 0) = 0.6, \pi_r = 0.2$. 256

C.1 Standard dataframe with start-stop notation for four hypothetical individuals experiencing the recurrent and terminal event process in Figure 1.1(e). 258

C.2 Illustration of a stacked dataframe for fitting the pseudo-Poisson procedures at time points $s_1, s_2, \ldots, s_R$. 259

D.1 True values for $p_{ij}(\tau), p_{ji}(\tau)$ and $\beta^{\dagger}(\tau)$ when $\tau = 1$. 274

D.2 Empirical average of $\hat{p}_{ij}(\tau), \hat{p}_{ji}(\tau)$ and $\hat{\beta}^{\dagger}(\tau)$ for different values of $\pi_r = 0, 0.2, 0.4$ and 0.6 when the true illness-death process is Markov with intensities $\lambda_{jl}(t|X) = \lambda_{jl0} \exp(\gamma_{jl} X)$; $\tau = 1, \pi_T = 0.6, \pi^1_T = 0.7, \pi_D = 0.25, n = 250$ individuals, $n_{sim} = 2000$. 275

D.3 Empirical properties of the utility-based estimator $\hat{\beta}(\tau)$ and associated tests of $H_0 : \beta(\tau) \leq 0$ vs. $H_1 : \beta(\tau) > 0$ when the illness-death process is Markov with intensities $\lambda_{kl}(t|X) = \lambda_{kl0} \exp(\gamma_{kl} X)$; $\tau = 1, \pi_T = 0.6, \pi^1_T = 0.7, \pi_D = 0.4, \pi_r = 0.2$ and $0.4, \exp(\gamma_r) = 1, u_0 = 1, u_1 = 0.5, u_2 = 0, n = 100, 250, 1000$ individuals, $n_{sim} = 10000$. 287

D.4 Empirical properties of the utility-based estimator $\hat{\beta}(\tau)$ and associated tests of $H_0 : \beta(\tau) \leq 0$ vs. $H_1 : \beta(\tau) > 0$ when the illness-death process is Markov with intensities $\lambda_{kl}(t|X) = \lambda_{kl0}\exp(\gamma_{kl}X)$; $\tau = 1, \pi_T = 0.6, \pi_T^1 = 0.7, \pi_D = 0.4, \pi_r = 0.2$ and $0.4, \exp(\gamma_r) = 0.85, u_0 = 1, u_1 = 0.5, u_2 = 0, n = 100, 250, 1000$ individuals, $n_{sim} = 10000$. 288

D.5 Empirical properties of the utility-based estimator $\hat{\beta}(\tau)$ and associated tests of $H_0 : \beta(\tau) \leq 0$ vs. $H_1 : \beta(\tau) > 0$ when the illness-death process is mixed-Markov with conditional intensities $\lambda_{kl}(t|X, U) = U\lambda_{kl0}\exp(\gamma_{kl}X)$ for $\phi = 0.5$; $\tau = 1, \pi_T = 0.6, \pi_T^1 = 0.7, \pi_D = 0.4, \pi_r = 0.2$ and $0.4, \exp(\gamma_r) = 1, u_0 = 1, u_1 = 0.5, u_2 = 0, n = 100, 250, 1000$ individuals, $n_{sim} = 10000$. 289

D.6 Empirical properties of the utility-based estimator $\hat{\beta}(\tau)$ and associated tests of $H_0 : \beta(\tau) \leq 0$ vs. $H_1 : \beta(\tau) > 0$ when the illness-death process is mixed-Markov with conditional intensities $\lambda_{kl}(t|X, U) = U\lambda_{kl0}\exp(\gamma_{kl}X)$ for $\phi = 1$; $\tau = 1, \pi_T = 0.6, \pi_T^1 = 0.7, \pi_D = 0.4, \pi_r = 0.2$ and $0.4, \exp(\gamma_r) = 1, u_0 = 1, u_1 = 0.5, u_2 = 0, n = 100, 250, 1000$ individuals, $n_{sim} = 10000$. 290

List of Abbreviations

AG Andersen-Gill

AJ Aalen-Johansen

ASE average standard error

CI confidence interval

CIF cumulative incidence function

CL Cook-Lawless

CV cardiovascular

DB direct binomial

ESE empirical standard error

FDA Food and Drug Administration

FG Fine-Gray

GL Ghosh-Lin

HF heart failure

IBCSG International Breast Cancer Study Group

IE intercurrent event

IJK infinitesimal jackknife

ITT intention-to-treat

KM Kaplan-Meier

MACEs major adverse cardiovascular events

MI myocardial infarction

ML Mao-Lin

NA Nelson-Aalen

NASCET North American Symptomatic Carotid Endarterectomy Trial

OJK ordinary jackknife

PREPAReS Pathogen Reduction Evaluation and Predictive Analytical Rating Score

RCT randomized controlled trial

RMST restricted mean survival time

RMT-IF restricted mean time in favor of treatment

SREs skeletal-related events

WR win ratio

Chapter 1

Introduction

Phase III trials are an essential component in the clinical development of new interventions and treatments, following preclinical research and dose assessments in early phase trials (Moyé, 2003; Friedman et al., 2015). The overall goal of conducting phase III trials is to evaluate the benefits and risks of an experimental treatment within a "controlled" environment, characterized by randomized treatment allocation (Fisher, 1925), clear inclusion and exclusion criteria, and regular monitoring of trial participants. A randomized controlled trial (RCT), conducted by the United Kingdom Medical Research Council in 1947 to assess the efficacy of streptomycin for the treatment of pulmonary tuberculosis (Marshall et al., 1948), marks the beginning of randomization-based evidence in medicine - used earlier in agriculture, for example. Since then, RCTs have become the gold standard design for causal treatment comparisons in clinical research (Peto et al., 1976, 1977; Feinstein and Horwitz, 1982). However, even with well-designed randomized trials, major challenges can arise in the assessment of treatments and interpretation of related estimands. This is particularly true in trial settings in which complex disease processes such as cancer (Hortobagyi et al., 1996; Rittmeyer et al., 2017; Powles et al., 2018), cardiovascular (CV) diseases (Barnett et al., 1998; O'Connor et al., 2009; Marso et al., 2016), or neurological diseases (Hauser et al., 2017; Bateman et al., 2023) are studied. We motivate the research projects of this thesis by introducing statistical issues that commonly arise in the design, conduct and analysis of randomized clinical trials with chronic disease processes.

1.1 Statistical challenges in late stage clinical trials

Randomization is fundamental to clinical phase III trials. It can be traced back to Fisher (1925) and describes the process of randomly assigning individuals to receive either a new experimental treatment or the current standard treatment/placebo. While there are several other rationales for the use of randomization in health research (Sprott and Farewell, 1993), it is the comparability of treatment groups at time zero with respect to baseline characteristics and the resulting facilitation of causal inference that makes randomization crucial in clinical trials. Hernán and Robins (2020) emphasize the important distinction between association and causation in statistical analyses. Clinical trials aim for making causal statements about the effects of a treatment on an outcome.

Chronic disease processes are by nature dynamic, so in many clinical trials interest lies in assessing the causal effect of treatment on the time to some undesirable adverse event observed subject to censoring. Right censoring is a common phenomenon in studies involving times to events (Klein and Moeschberger, 2003). It may occur either due to premature loss to follow-up or end-of-study censoring. Censored failure time data can be analyzed using hazard-based techniques (Lawless, 2002), with the hazard function representing the instantaneous probability of an event occurrence at time t, conditional on being alive and event-free at t^-. While hazard-based, or more generally intensity-based, analyses are needed to study the dynamics of disease processes (Aalen, 2012), they do not benefit from randomization and impose challenges for causal treatment comparisons in clinical trials.

This is further complicated by the fact that individuals are typically at risk of several types of adverse events, so it is often unclear how best to select a primary event and to specify a sufficiently informative "estimand". An estimand is a parameter that quantifies the effect of treatment in an assumed model. There has been much discussion about the properties estimands should possess and how they should be defined in various situations; these issues are at the heart of this thesis. Desirable properties of an estimand include clinical relevance, transparency, interpretability and robustness. Additionally, an estimand should remain estimable when there is loss to follow-up and facilitate design aspects such as sample size and power calculations.

Throughout this thesis we discuss estimands in the context of two different types of clinical trial settings: (I) randomized trials where primary interest lies in survival (or death due to a particular cause) but other relevant disease-related events may arise as well, and (II) randomized trials where primary interest lies in the reduction or prevention of non-fatal events in the presence of death.

1.1.1 Trials with (disease-related) death as primary event of interest (I)

In cardiovascular (CV) trials, interest lies in prolonging overall survival but with recent advances in medical therapies and surgical interventions mortality rates are now relatively low. A trial sufficiently powered to detect a clinically relevant treatment effect on survival requires prohibitively long follow-up or large sample sizes, and is thus difficult to conduct within typical time and cost constraints. This has lead to increased interest in using composite outcomes based on recurrent myocardial infarction (MI), stroke, and hospitalization, in addition to time to CV-related death. Although composite endpoints can lead to more efficient designs (Freemantle et al., 2003), they target different estimands which can be difficult to interpret. Moreover, they fail to reflect the hierarchical ordering of the component event types based on their clinical importance. Here CV death is much worse than stroke, and treatment may affect each component in different ways. The risk of dying from a cause unrelated to CV disease gives rise to a further competing risks problem.

In cancer trials, patients are at risk of cancer progression and death, and there has been considerable discussion about the merits of responses based on progression alone, progression-free survival (PFS) and overall survival (OS) (Broglio and Berry, 2009; Booth and Eisenhauer, 2012). OS is defined as the time from randomization until death from any cause, whereas PFS is the time from randomization to the earliest of progression, progression-free death or death due to reasons unrelated to cancer (FDA, 2007). While OS remains the preferred endpoint due to its significant clinical relevance, PFS has also been accepted by regulatory agencies as an alternative endpoint. Specifically, PFS is being increasingly used as a surrogate for OS. Surrogate endpoints are often motivated by the desire to accelerate treatment approval through earlier assessments of the disease burden, but a beneficial effect of treatment on the surrogate event (e.g. prolonged PFS) may not directly imply prolonged survival. Despite the fact that primary analyses based on PFS or progression alone can enhance shorter duration trials, they fail to capture the treatment effect on patients' overall disease experience and do not always employ clinically meaningful estimands.

In analyses of the time to progression, death from any cause should not be treated as a censoring event but rather as a competing risk.

1.1.2 Trials with a non-fatal primary event of interest (II)

Termination by death is also a common phenomenon in palliative trials of individuals with life-threatening illnesses. Palliative treatments are typically intended to improve patients' quality of life by preventing major adverse complications rather than prolonging their survival. A trial, for example, may then be designed with the primary analysis based on a, possibly recurrent, non-fatal event endpoint. However, death may occur before the end of follow-up, and this terminates the non-fatal event process of interest. This is in contrast to censoring where observation of the event process is terminated due to early withdrawal from study or administrative closure of study (Cook and Lawless, 2007, Section 2.6); in that case the event of interest can still be experienced beyond censoring but it is not observable. Other examples include trials in rheumatic disease, where the objective is to reduce the risk of joint damage or disability, with death as a (infrequent) terminal event. Although there are various ways to quantify damage and disease burden, specification of an optimal estimand is still open to discussion (Ogdie and Coates, 2017).

1.1.3 Intercurrent events in type (I) and (II) trial settings

Additional issues complicating the selection and interpretation of estimands arise when so-called intercurrent events occur. The ICH E9-R1 addendum on estimands and sensitivity analyses defines intercurrent event as "events occurring after treatment initiation that affect either the interpretation or the existence of the measurements associated with the clinical question of interest" (ICH E9 Steering Committee, 2019). We classify intercurrent events which preclude observation of the clinical event of interest as *type 1* intercurrent events, and intercurrent events which do not preclude the observation but change the interpretation of the clinical events as *type 2* intercurrent events. A type 1 intercurrent event (IE) can be further sub-classified as a *type 1A* IE if it simply impacts the observation of the clinical event of interest because of loss to follow-up or a *type 1B* IE that precludes the occurrence of the clinical event. Death in palliative trials (II) constitutes a type 1B IE, whereas early withdrawal from study due to inefficacy or adverse reactions to treatment

in either trial setting are type 1A IEs. In oncology trials (I), where primary interest lies in the effects of treatment on overall survival, introduction of rescue medication or treatment switching following cancer progression are examples of type 2 IEs.

1.1.4 The utility of multistate modeling in analyzing clinical trials

Disease processes and patient management are by nature dynamic. Aalen (2012) among others has highlighted the importance of modeling data in ways that recognize the evolutionary nature of disease processes. Time is a basic element and so multistate models offer an intuitive framework for characterizing and modeling disease processes (Andersen et al., 1993; Cook and Lawless, 2018), with occurrences of non-fatal and fatal events modelled as transitions between different disease states. Keiding et al. (2001), for example, used multistate modeling for the analysis of graft-versus-host disease in bone marrow transplants from registry data. Bacchetti et al. (2010) proposed a multistate approach to model progression of liver fibrosis in patients with hepatitis C. Cook and Lawless (2014) formulated different multistate models to jointly analyze the onset and progression of damage in sacroiliac joints among patients in a psoriatic arthritis registry. Multistate models are not only appealing for the analysis of observational data from disease registries, electronic health records or other administrative databases (Cook and Lawless, 2022), but also for causal inferences from clinical trial data. A major contribution of this thesis is to show the utility of multistate models for primary and secondary analyses of clinical trials with complex disease processes. We demonstrate how the statistical challenges described in the previous sections can be addressed within a comprehensive multistate model framework.

1.2 A brief introduction to the estimand framework based on multistate models

1.2.1 Multistate model formulation for disease processes

Multistate models have wide applicability to many disease processes. Some models that are highly relevant to the design and analysis of randomized trials are displayed in Figure 1.1, including competing and semi-competing risks processes (Beyersmann et al., 2012; Putter et al., 2007), some particular progressive processes (Cook et al., 2003; Cook and Lawless, 2018), and recurrent and terminal event processes (Cook and Lawless, 2007).

We consider a general multistate process with state space $\mathcal{S} = \{\mathcal{A}, \mathcal{D}\}$, where $\mathcal{A} = \{0, 1, \ldots, K\}$ is the set of transient states and $\mathcal{D}$ the set of absorbing states. State 0 is the initial event-free state occupied by individuals at the time of randomization ($t = 0$), and states $k = 1, \ldots, K$ may represent non-fatal disease conditions of increasing severity. The absorbing states in $\mathcal{D}$ are entered upon the occurrence of a fatal event (death), with the elements of $\mathcal{D}$ distinguishing between different causes of death, for example. We note that the terms "*non-fatal*" (or "*fatal*") and "*non-terminal*" (or "*terminal*") will be used interchangeably throughout this thesis. We focus for most of the thesis on progressive processes, in which case individuals in a particular transient state can only move to the next highest transient state or to an absorbing state. In some progressive processes, such as in Figure 1.1(b) for instance, entry to the absorbing state is only possible through state K. Reversible (or non-progressive) processes are useful in clinical trials for therapies designed to reduce episodes of increased disease activity. Examples include exacerbations in chronic obstructive pulmonary disease (Grossman et al., 1998) or relapse activity in multiple sclerosis (Hauser et al., 2017); we briefly comment on reversible processes in the context of utility function methods in Chapter 5.

We let $Z(t) \in \mathcal{S}$ denote the state occupied by a generic individual at time t, and let $\{Z(t), t > 0\}$ represent the associated process with $P(Z(0) = 0) = 1$. The entry time to state $k \in \mathcal{S} \setminus \{0\}$ is denoted by $T_k = \inf\{t > 0 : Z(t) = k\}$. Within the setting of a two-arm clinical trial, we let X be an indicator equal to 1 for individuals randomized to the experimental treatment group and 0

otherwise. The intensity function for a $k \longrightarrow l$ transition is defined as

$$\lambda_{kl}(t \mid \mathcal{H}(t)) = \lim_{\Delta t \downarrow 0} \frac{P(Z(t + \Delta t^-) = l \mid Z(t^-) = k, \mathcal{H}(t))}{\Delta t} , \quad k \in \mathcal{A}, l \in \mathcal{S}, k \neq l , \quad (1.2.1)$$

where $\mathcal{H}(t) = \{Z(u) : 0 < u < t, X\}$ is the process history up to time t^- plus the treatment indicator X. In counting process notation, the intensities can be equivalently defined as

$$\lambda_{kl}(t \mid \mathcal{H}(t)) = \lim_{\Delta t \downarrow 0} \frac{P(\Delta N_{kl}(t) = 1 \mid Z(t^-) = k, \mathcal{H}(t))}{\Delta t} , \quad k \in \mathcal{A}, l \in \mathcal{S}, k \neq l ,$$

with $\{N_{kl}(t), t > 0\}$ the right-continuous counting process for $k \longrightarrow l$ transitions and $\Delta N_{kl}(t) = N_{kl}(t + \Delta t^-) - N_{kl}(t^-)$. The stochastic nature of a multistate process is fully defined by the specification of all transition intensity functions (Andersen and Keiding, 2012). In what follows, we use the term *process feature* to refer to any functional of the set of intensities.

Treatment effects, or more generally, covariate effects on the intensities can be assessed through a variety of different models; semiparametric Cox-type regression models with multiplicative effects (Cox, 1972) or additive hazards models (Aalen, 1980) are popular. Software for fitting and checking such models is widely available (Cook and Lawless, 2018), facilitating comprehensive secondary intensity-based analyses of clinical trials.

1.2.1.1 Multiplicative Markov models

We primarily focus on multiplicative Markov regression models. A process $\{Z(t), t > 0\}$ is Markov if the transition intensities depend only on time t and the state occupied at t^-:

$$\lambda_{kl}(t \mid \mathcal{H}(t)) = Y_k(t)\lambda_{kl}(t \mid X) , \quad k \in \mathcal{A}, l \in \mathcal{S}, k \neq l .$$

We let $d\mathbb{A}(t|X) = \left[d\Lambda_{kl}(t|X)\right]_{k,l\in\mathcal{S}}$ denote the Markov transition intensity matrix with entries $d\Lambda_{kl}(t|X) = \lambda_{kl}(t|X)dt$ for $k \neq l$ and $d\Lambda_{kk}(t|X) = -\sum_{l\neq k} \lambda_{kl}(t|X)dt$ on the diagonal (Cox and Miller, 1965). The transition probability matrix $\mathbb{P}(s, t|X)$ with (k, l) entry

$$P_{kl}(s, t|X) = P(Z(t) = l|Z(s) = k, X)$$

is subject to $\sum_{l,l\neq k} P_{kl}(s, t|X) = 1$, and can be obtained by product integration as

$$\mathbb{P}(s, t|X) = \left[P_{kl}(s, t|X)\right]_{k,l\in\mathcal{S}} = \prod_{(s,t]} \left(\mathbb{I} + d\mathbb{A}(u|X)\right) ,$$

where $\mathbb{I}$ denotes the identity matrix (Andersen et al., 1993).

For a general multiplicative Markov regression model the transition intensities take the form

$$\lambda_{kl}(t|V(t), X) = \lambda_{kl0}(t) \exp(\widetilde{\gamma}'_{kl} V(t) + \gamma_{kl} X) \ , \tag{1.2.2}$$

where $V(t) = (V_1(t), \ldots, V_p(t))'$ is a $(p \times 1)$ vector of additional time-varying and/or baseline covariates (i.e. $V_j(t) := V_j, j = 1, \ldots, p$), $(\widetilde{\gamma}'_{kl}, \gamma_{kl})'$ is a $((p+1) \times 1)$ vector of regression coefficients, and $\lambda_{kl0}(t)$ is the baseline intensity function of individuals for whom $X = 0$ and $V(t) = 0$, respectively. Semiparametric models (1.2.2) are often preferred for assessing covariate effects as they do not depend on parametric assumptions about the baseline functions. The time-varying components in $V(t)$ may be either external or internal covariates (Kalbfleisch and Prentice, 2002).

1.2.2 Marginal process features

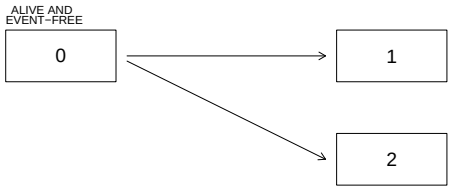
Intensity-based analyses are natural for modeling disease processes but marginal analyses are more suited for primary analysis in clinical trials. Marginal effects of treatment can be formulated based on a variety of different marginal process features. We let E represent a defined event, and let $N_E(t)$ denote the number of type E events occurring over the time period $[0, t]$; $\Delta N_E(t) = N_E(t + \Delta t^-) - N_E(t^-)$ is then the number of type E events over $[t, t + \Delta t^-)$ and $dN_E(t) = \lim_{\Delta t \downarrow 0} \Delta N_E(t)$. While the most suitable marginal feature for a particular trial depends on the scientific primary question, we restrict attention to the expected number of type E events over $[0, t]$, defined as

$$\mathbb{E}(N_E(t)) = \mathbb{E}\left(\int_0^t dN_E(u) \right) \ , \tag{1.2.3}$$

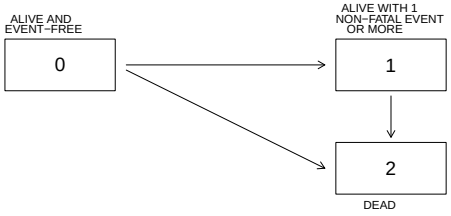
and the expected cumulative utility score over $[0, t]$ (Cook et al., 2003), given by

$$D(t) = \mathbb{E}(U(t)) = \sum_{k \in \mathcal{S}} u_k \int_0^t P(Z(u) = k) du \ , \tag{1.2.4}$$

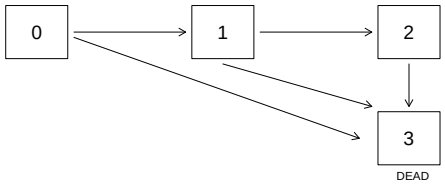
where $P(Z(u) = k) = P_{0k}(0, u)$ is the state k occupancy probability at time u and u_k is a pre-specified utility score (or quality of life score) for state $k \in \mathcal{S}$. Feature (1.2.3) covers entry time



(a) Competing risks process:
$\mathcal{A} = \{0\}, \mathcal{D} = \{1, 2\}$.



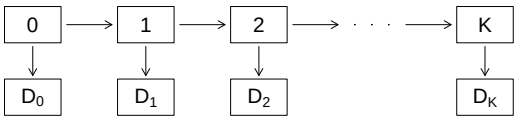
(b) Illness-death process:
$\mathcal{A} = \{0, 1\}, \mathcal{D} = \{2\}$.



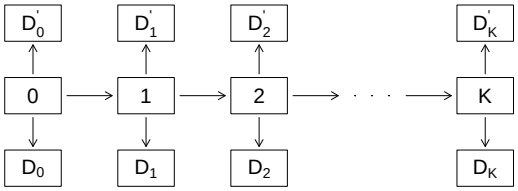
(c) Progressive process:
$\mathcal{A} = \{0, 1, 2\}, \mathcal{D} = \{3\}$.



(d) Progressive process:
$\mathcal{A} = \{0, 1, 2\}, \mathcal{D} = \{3\}$.



(e) Recurrent event process terminated by all-cause death: $\mathcal{A} = \{0, 1, \ldots, K\}$, $\mathcal{D} = \{D_0, D_1, \ldots, D_K\}$.



(f) Recurrent event process terminated by disease-related or non-disease related death: $\mathcal{A} = \{0, 1, \ldots, K\}$, $\mathcal{D} = \{D_0, D_1, \ldots, D_K, D'_0, D'_1, \ldots, D'_K\}$.

Figure 1.1: Multistate diagrams for chronic disease processes.

distributions and state occupancy probabilities as special cases; if for instance E represents the intermediate non-fatal event in a semi-competing risks process, (1.2.3) reduces to the cumulative distribution function $F_1(t) = P(T_1 \leq t)$ for the entry time T_1 to state 1, also called the cumulative incidence function for non-fatal events. More generally, the event E may be of a single type (e.g. MI, stroke, CV death), or may be a so-called composite event that consists of two or more event types, which may be recurrent or non-recurrent (e.g. MI or stroke, MI or stroke or CV death). Feature (1.2.4) includes restricted mean times spent in a particular state k over $[0, t]$ as a special case:

$$RMT_k(t) = \mathbb{E}(S_k(t)) = \int_0^t P(Z(u) = k)du, \tag{1.2.5}$$

where $S_k(t) = \int_0^t \mathbb{1}(Z(u) = k)du$ denotes the total time spent in state k over $[0, t]$.

1.2.3 Frameworks for the specification of marginal estimands

1.2.3.1 Time-to-event and count framework

In randomized trials there is a strong desire to specify one-dimensional estimands to facilitate causal inference and trial design. When (1.2.3) is the marginal feature of interest in trial settings (I) or (II), this can be addressed by adopting generalized linear models of the form

$$g\big(\mathbb{E}(N_E(t)|X)\big) = \alpha(t) + \beta X \ , \tag{1.2.6}$$

where g is a specified differentiable monotonic function on the domain of $\mathbb{E}(N_E(t))$, e.g., $g(u) = \log(u), \log(-\log(1-u))$, $\alpha(t) = g(\mathbb{E}(N_E(t)|X = 0))$ is a monotonic baseline function with $\alpha(t) \downarrow -\infty$ as $t \downarrow 0$, and

$$\beta = g(\mathbb{E}(N_E(t)|X = 1)) - g(\mathbb{E}(N_E(t)|X = 0))$$

is a one-dimensional estimand. Estimation of $\alpha(t)$ and β in marginal models (1.2.6) typically requires inverse probability of censoring weighting techniques.

Common frameworks of (1.2.6) include proportional mean models for recurrent and terminal event processes, such as those of Ghosh and Lin (2002) and Mao and Lin (2016),

$$\mathbb{E}(N_E(t)|X) = \mathbb{E}(N_E(t)|X = 0)\exp(\beta X)\,, \tag{1.2.7}$$

and transformation models for cumulative incidence functions of non-fatal events

$$g(F_1(t|X)) = \alpha(t) + \beta X \tag{1.2.8}$$

in semi-competing- and competing risks processes (Fine and Gray, 1999; Scheike et al., 2008).

1.2.3.2 Utility-based framework

Another framework that plays an important role in this thesis is utility-based, which has been used in quality of life comparisons in oncology trials (Gelber et al., 1989, 1995; Glasziou et al., 1990; Cook et al., 2003). In utility-based analyses, a one-dimensional estimand can be specified as the difference (or ratio) in the marginal feature (1.2.4) between the two treatment groups at a specific time $\tau > 0$:

$$\beta(\tau) = D(\tau|X = 1) - D(\tau|X = 0) = \mathbb{E}(U(\tau)|X = 1) - \mathbb{E}(U(\tau)|X = 0)\ . \tag{1.2.9}$$

A straightforward approach is to use nonparametric Aalen-Johansen estimates of state occupancy probabilities $P(Z(t) = k|X)$ for each treatment group to give robust estimates for (1.2.9). Despite the challenge of specifying the utility scores u_k (Llewellyn-Thomas et al., 1984), utility-based estimands have enormous potential in trial settings (I) where interest lies in combining multiple outcomes and synthesizing treatment effects across the different event types.

1.3 Motivating studies

1.3.1 Efficacy of liraglutide for cardiovascular outcomes in type II diabetes

Type II diabetes is a common form of diabetes where the human body is able to produce insulin itself but fails in its proper use. Individuals affected by this chronic condition experience hyperglycemia, which can ultimately cause serious cardiovascular complications over time. Liraglutide, a glucagon-like peptide-1 (GLP-1) receptor agonist, has been shown to be effective in lowering blood glucose levels (Du et al., 2014). Treatment with liraglutide can thus help manage type II diabetes, but it is as important to determine whether liraglutide has at the same time beneficial effects

on other outcomes. The multi-center, double-blind, placebo-controlled LEADER (Liraglutide Effect and Action in Diabetes: Evaluation of Cardiovascular Outcome Results) trial was initiated in 2010 to assess the long-term effects of liraglutide on major adverse cardiovascular events (MACEs), such as non-fatal stroke, non-fatal MI and CV death (Marso et al., 2016). Recruitment of eligible individuals took place between September 2010 and April 2012 from a total of 412 sites in 32 countries. Upon recruitment, 9340 type II diabetic patients at a high risk of CV disease and aged above 50 years were randomized to receive either a daily injection of liraglutide in addition to standard care (n=4668) or standard care alone (n=4672). Each patient was closely monitored for the occurrence of recurrent non-fatal CV events until the earliest of death (CV- or non-CV related), loss to follow-up or the administrative closure of study. The median follow-up time was 3.8 years in both treatment arms, with a planned administrative follow-up period between 3.5 and 5 years (Furberg et al., 2022; Andersen and Ravn, 2023). The multistate model displayed in Figure 1.1(f) can be used to describe the disease process in the present context, with fatal events classified as either disease-related (D_k) or non-disease related (D'_k).

Treatment effects in the LEADER trial were assessed by a Cox analysis of the time from randomization to the first occurrence of stroke, MI or CV death (Marso et al., 2016). Key issues with such a primary analysis are: (i) the hazard ratio as a treatment effect measure does not have a causal interpretation (Aalen et al., 2015), (ii) it may not be desirable to combine non-fatal and fatal events into a composite endpoint, (iii) death unrelated to any CV event may give rise to a competing risks problem, and (iv) a time to first event analysis does not take into account individuals' entire trajectories.

The LEADER trial motivates two research themes in this thesis. Theme 1 in Chapter 2 is inspired by the concerns outlined in (i)-(iv), which are not limited to the LEADER trial but relevant to many other settings, and addresses the need for clear guidance on how best to specify estimands in randomized trials of complex disease processes. Work related to theme 2 in Chapter 4 discusses the limitations of composite endpoint analyses.

1.3.2 Assessing carotid endarterectomy in stroke prevention

Atherosclerosis involves the development of arterial plaque and stenosis of the carotid arteries, putting individuals at increased risk of stroke. Barnett et al. (1998) report on the multi-center North American Symptomatic Carotid Endarterectomy Trial (NASCET) designed to evaluate the effect of a surgical intervention called carotid endarterectomy, which involves surgical removal of plaque to increase the diameter of the carotid artery and enhance blood flow, compared to usual medical therapy (e.g. use of platelet inhibitors and antihypertensive drugs). The goal of both treatment strategies is the prevention of stroke and stroke-related death in patients with carotid stenosis. To be eligible for recruitment, individuals must have experienced at least one transient ischemic attack or minor stroke and have at least a 30% narrowing of the carotid artery on the same side as (ipsilateral to) the event, as determined by central angiographic examination. Two subgroups of patients were of particular interest, defined by whether the degree of stenosis at randomization involved a $< 70\%$ narrowing of the carotid artery (= moderate stenosis) or $\geq 70\%$ narrowing (= severe stenosis). Consenting patients in both strata were randomized to either carotid endarterectomy and best medical care, or best medical care alone. As part of best medical care, aspirin was recommended for all patients at a dose of 1300 mg/day and blood pressure was carefully controlled through regular monitoring. The endpoints of interest included (i) any stroke arising from the same side as the one designated at the time of recruitment as a possible site for surgery, (ii) any stroke, (iii) the composite event of stroke or stroke-related death, and (iv) the composite event of any stroke or death. For (iii), deaths unrelated to stroke are a competing risk which should be addressed when assessing the effect of surgery.

An interim analysis led to early termination of recruitment to the severe stenosis stratum ($n = 659$) and publication of the finding that carotid endarterectomy was highly effective in preventing stroke among these patients (North American Symptomatic Carotid Endarterectomy Trial Collaborators, 1991). The recruitment and follow-up of the moderate stenosis strata continued; a total of 2226 patients with stenosis of less than 70% were assigned to either the medical (n =1118) or surgical arm (n =1108). Upon dissemination of the beneficial effect of carotid endarterectomy in severe stenosis, many patients in the moderate stenosis stratum elected to switch from the medical to the surgical therapy and underwent surgery. In addition, physicians treating patients experiencing non-fatal stroke recommended that they switch arms and undergo carotid endarterectomy to reduce

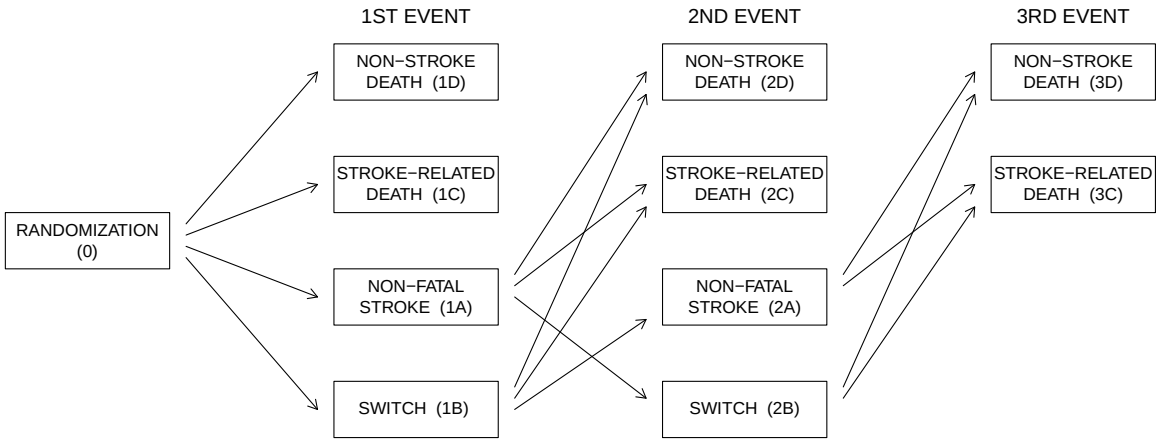


Figure 1.2: A multistate model depicting the occurrence of stroke, switch from medical care to carotid endarterectomy, stroke-related death and non-stroke related death for the NASCET trial reported on by Barnett et al. (1998).

the risk of future stroke. Figure 1.2 contains a multistate diagram depicting the occurrence of the first event among stroke, stroke-related death, death unrelated to stroke, and crossover to surgery; treatment switching can only occur among patients randomized to medical care. Additional states can be entered upon the occurrence of subsequent events of the particular types depicted under the headings of second and third events.

Treatment switching in this motivating study constitutes a type 2 intercurrent event that affects the analysis and interpretation of the NASCET data. Treatment comparisons in the presence of intercurrent events have received a great deal of attention in recent years (Ster et al., 2020; Qu et al., 2021; Manitz et al., 2022), given the importance of this topic. However, the majority of existing approaches specify estimands in ways that do not reflect real world trials. This motivates considerations of strategies that employ clear real world interpretation of estimands; in Chapter 2, we explore this within an expanded multistate model framework based on primary and intercurrent event processes.

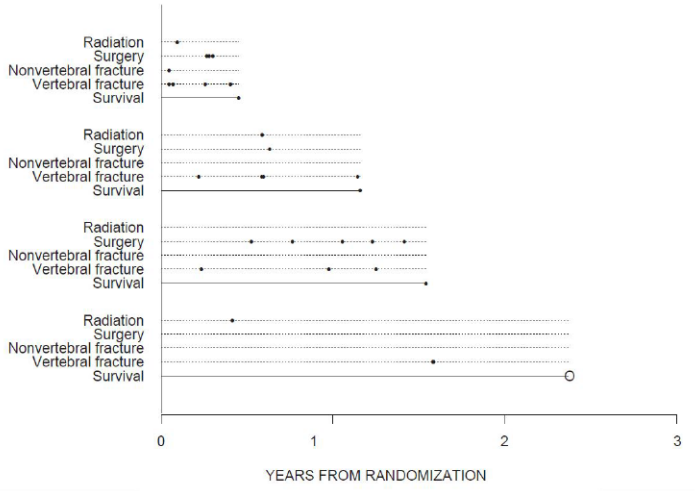


Figure 1.3: Timeline diagrams depicting the occurrence of different skeletal complications and survival or censoring times in four selected breast cancer patients with skeletal metastases from Theriault et al. (1999); black lines represent the time from randomization to the end of follow-up (due to death, early withdrawal or administrative censoring), • = skeletal-related event or death, ○=censoring.

1.3.3 Prevention of skeletal complications in metastatic breast cancer

Cancer patients with bone metastases are at an increased risk of skeletal-related adverse events such as fractures, spinal cord compression with vertebral fractures, and acute bone pain requiring radiotherapy. Therapeutic or surgical interventions are typically required to treat or prevent these fractures (Coleman, 2006). Cancer metastatic to bone cannot be cured, but prophylactic treatment with bisphosphonates (e.g. pamidronate) may strengthen bone and help reduce the risk of skeletal complications, and so can be used to improve quality of life in affected individuals. A better understanding of the effects of bisphosphonates on the occurrence of skeletal-related events (SREs) is thus an important contribution to disease management.

Hortobagyi et al. (1996) report on an international multi-center randomized trial conducted to compare monthly infusions of pamidronate (90 mg) with placebo for the prevention of skeletal complications in stage IV breast cancer patients with bone metastases. To be eligible for recruitment, patients must have received cytotoxic chemotherapy and had at least two predominantly

lytic bone lesions at least one centimeter in diameter. Between January 1991 and March 1994 eighty-five sites in the United States, Canada, Australia and New Zealand recruited 380 women, with 185 randomized to receive a 90 mg infusion of pamidronate every four weeks and 195 to receive a placebo control. Skeletal complications of interests included nonvertebral and vertebral fractures, the need for surgery to treat and prevent fractures, and the need for radiation for the treatment of bone pain. Individuals with metastatic cancer are at high risk of death so that the skeletal-related event process is terminated by death. This creates a semi-competing risks problem if interest lies in an analysis of the time to the first skeletal event, or a recurrent and terminal event problem if interest lies in an analysis of recurrent skeletal events. Figure 1.1(b) displays the well-known irreversible illness-death model, which is applicable in the former setting. Figure 1.1(e) is a multistate diagram depicting the occurrence of up to K SREs, while accommodating death from any of the transient states. SREs were recorded up to the earliest of death, loss to follow-up or the administrative closure of study; see Figure 1.3 where the disease course of four selected breast cancer patients from Theriault et al. (1999) are graphically illustrated. After completion of the planned one year follow-up, observation was extended for an additional year to assess long-term effects on survival.

This study is an example of a palliative type (II) trial, in which analysis and interpretation of data is affected by a non-neglible mortality rate. Strategies for dealing with complications due to competing risks motivate research projects in Chapters 3 and 4.

1.3.4 Quality of life in a therapeutic breast cancer trial

The International Breast Cancer Study Group (IBCSG) analyzed a randomized clinical trial investigating the effectiveness of short- and long-duration adjuvant chemotherapy in 1229 breast cancer patients (Ludwig Breast Cancer Study Group, 1988). Patients experience toxicity during chemotherapy but intervals during which they are toxicity-free and relapse-free typically follow the completion of chemotherapy; patients may of course relapse and die following such an interval.

To assess the impact of long- versus short-duration chemotherapy, overall survival could be assessed through a standard time to death analysis. The therapeutic issue under investigation here is, however, whether the greater anti-tumour effect of longer duration chemotherapy outweighs the protracted periods of toxicity-related symptoms and poorer quality of life. To explore the

impact of treatment on the sojourn times in the different stages of response, Gelber et al. (1989) proposed partitioning the survival time into different stages according to the four-state model in Figure 1.1(d) with state space $\mathcal{S} = \{0, 1, 2, 3\}$. In this context, state 0 is occupied by a breast cancer patient during the period of toxicity due to chemotherapy, state 1 is occupied following chemotherapy when patients are toxicity-free and symptom-free, state 2 is entered when symptoms reemerge upon relapse, and state 3 is the absorbing state entered upon death. The challenge in analysis of such data is to summarize the effect of treatment on the sojourn times in the different states or the state entry time distributions in a robust way that facilitates simple comparisons between the treatment arms. In Gelber et al. (1989), and a series of papers that followed by Gelber and colleagues, this was achieved by assigning utilities to each day in each state and accumulating this over time; the utility scores were chosen in a way that a higher score indicates higher quality of life, i.e. $u_{k-1} > u_k, k = 1, 2, 3$ and $u_0 = 1, u_3 = 0$. The resulting summary measure was called the Quality-adjusted Time Without Symptoms of disease and Toxicity to treatment, abbreviated as Q-TWiST (Gelber et al., 1989).

Motivated by the Q-TWiST measure in oncology trials (Glasziou et al., 1990; Gelber et al., 1995), Chapter 5 explores the properties and usefulness of utility-based estimands (1.2.9) in disease processes associated with multiple types of events of interest.

1.3.5 Assessing the burden of bleeding in thrombocytopenia

Patients with chemotherapy-induced thrombocytopenia frequently experience bleeding complications due to low platelet counts (Franchini et al., 2013). Platelet transfusions are routinely used for the prevention and treatment of bleeding in such patients. Pathogen-reduced transfusion strategies have been developed to reduce the risk of bacterial contamination of stored platelet products and to prevent transfusion-transmitted infections. The purpose of the Pathogen Reduction Evaluation and Predictive Analytical Rating Score (PREPAReS) trial was to assess non-inferiority of the Mirasol pathogen reduction technology system compared to standard platelets in thrombocytopenic patients with hemato-oncology diseases (Ypma et al., 2016; van der Meer et al., 2018). Patients were eligible for recruitment if they were expected to require at least 2 transfusions during a follow-up period of 42 days. Recruitment took place between November 2010 and April 2016 from 10 different hemato-oncology centers in the Netherlands, Norway and Canada, with randomization

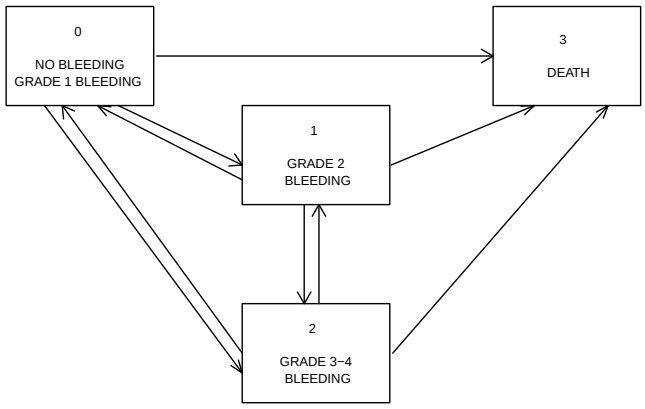


Figure 1.4: A multistate model for bleeding severity, recovery from thrombocytopenia and death in the PREPAReS trial reported on by van der Meer et al. (2018).

stratified by center. A total of 127 patients were recruited in Canada and randomized to receive either standard plasma stored platelets (arm A: plasma PC) or Mirasol-treated pathogen-reduced (PR) plasma stored platelets (arm B: PR-plasma PC).

Bleeding status was assessed daily using the World Health Organization (WHO) 4-point bleeding scale, which has also been used in other large transfusion medicine trials (Heddle et al., 2009; Stanworth et al., 2013). The WHO grade 0 represents no bleeding, grade 1 represents minor complications such as oropharyngeal bleeding or epistaxis lasting less than 30 minutes within 24 hours, and grades 2 to 4 represent more serious bleeding of increasing clinical importance. Although a grade ≥ 2 is typically considered clinically relevant, grades 3 and 4 represent the most severe complications. A bleeding requiring red cell transfusions is an example of a WHO grade 3 bleed, whereas central nervous system bleeds and fatal bleeds are examples of the WHO grade 4 bleeds. Alternatively, bleeding severity could be likewise measured using the 3-point Bleeding Severity Measurement Scale (BSMS) of Webert et al. (2012), where grade 0 represents no bleeding, grade 1 corresponds to not clinically significant bleeding, and grade 2 represents clinically significant bleeding.

Figure 1.4 represents a four-state model that can be used to assess non-inferiority of the platelet-reduced transfusion strategy on bleeding outcomes in the PREPAReS trial. Each point of the WHO bleeding scale is represented by a state in the multistate model: state 0 represents no bleeding or grade 1 bleeding, with states 1 to 3 entered upon grade 2 bleeds, grade 3-4 bleeds and death, respectively. Because individuals with grade ≥ 2 bleeds at the time of randomization were excluded from the primary analysis (Ypma et al., 2016, Figure 3), state 0 is the initial state occupied by individuals at $t = 0$. We consider the multistate process in continuous time, but bleeding scores are actually assigned daily by clinicians according to the individual's history for that day. For some individuals follow-up was terminated earlier due to platelet recovery, death, non-compliance, serious adverse transfusion reactions or immunological platelet refractoriness.

We discuss a utility-based approach for analysis of this trial in Section 5.5.

1.4 Research topics and outline of thesis

The issues introduced in Section 1.1 and in Section 1.3 further discussed in the context of examples have been discussed in the literature. In part due to the complexity of the problems, there is no clear guidance yet for many disease process settings. This thesis partially addresses this gap by offering a clear and rigorous discussion of the issues, examining the performance of available methods in these challenging areas, and offering guidance for study design and analysis. There are three thematic areas for this work.

1.4.1 Theme 1: Principles for estimand specification

The first theme is directed at statistical considerations concerning estimands in randomized clinical trials of event history processes. We use intensity-based multistate models as a basis for discussion and systematically develop guiding principles for defining primary estimands that admit causal interpretation. One of the principles emphasizes the importance of defining estimands based on marginal process features. The occurrence of "intercurrent" events, after randomization but before the primary outcome is observed, complicates the specification of a target estimand. For type (I)-(II) trial settings involving intercurrent events we show how the multistate model framework

can be extended to deal with post-randomization complications. Joint models are formulated based on multistate processes with an expanded state space encompassing both the primary and intercurrent event processes. An application to the motivating stroke prevention trial in carotid stenosis (Section 1.3.2) is provided to illustrate the proposed joint modeling approach.

1.4.2 Theme 2: Estimands and tests based on common marginal features

Intensity-based models for multistate processes are most closely aligned with how events unfold over time. When analyzing treatment effects in randomized clinical trials however, intensity-based models induce time-dependent confounding by conditioning on the complete process history at a given time. They therefore do not support conventional causal inferences regarding the effect of randomized treatments (Hernán, 2010; Aalen et al., 2015), despite playing a crucial role in gaining deeper insights into disease processes and in secondary analyses of trials. Because marginal models do not condition on post-randomization events they are more suitable for evaluating the effects of a randomized treatment in clinical trials.
The second theme of this thesis builds on the first theme, and focuses in more detail on generalized linear models (1.2.6) for marginal process features and their utility in the design, analysis and interpretation of randomized clinical trials. We assess in particular the consequences of potential model misspecification on the interpretability of treatment effects and robustness of tests. There are two research projects within this second theme.

The first research project is addressed in Chapter 3 and concerned with the use of regression models (1.2.8) based on complementary log-log transformations of cumulative incidence functions (CIFs) in clinical trial settings (II) involving semi-competing and competing risks processes. Methods for estimation of β in (1.2.8) are well-established and include the procedure of Fine and Gray (1999) (FG) based on a weighted partial likelihood function and direct binomial (DB) regression (Scheike et al., 2008). Our work investigates thoroughly the interpretability of CIF-based estimands by deriving the limiting values of the FG and DB treatment effect estimators in settings where the cloglog model is misspecified. We focus on model misspecification caused by the treatment having multiplicative effects on the intensity functions, and explore the dependence of the limiting values on the parameters of the underlying intensity models. A second goal is to assess the robustness of FG- and DB-based Wald tests of $H_0 : \beta = 0$ under some key hypotheses implied by the true

intensity-based process. Results concerning the probability limits of estimands under different hypotheses are used to study test properties and to provide guidance for the design and analysis of trials. An illustrative application is given to a palliative trial of patients with breast cancer metastatic to bone (Section 1.3.3), where the aim is to evaluate the effect of a bisphosphonate on the cumulative probability of skeletal-related events.

The second research project within the second theme is described in Chapter 4, which is devoted to mean function regression models (1.2.7) in clinical trials involving recurrent and terminal event processes. Within the multistate model framework we study the limiting values of treatment effect estimators from the proportional mean function models of Ghosh and Lin (2002) and Mao and Lin (2016) to gain insights into the interpretability of associated estimands.

Ghosh and Lin (2002) studied a mean function model based on the expected number of non-fatal events E occurring up to time t, $\mu_E(t) = \mathbb{E}(N_E(t))$, taking into account that death precludes the occurrence of further events. We consider the marginal Ghosh-Lin model as an approach to the analysis of type (II) trials aiming to prevent recurrent non-fatal events, and derive the limiting values of different treatment effect estimators in this model when the data-generating mechanism for the true process in Figure 1.1(e) is implied by a broad family of intensity-based or random effects models. We explore how the Ghosh-Lin estimand varies with factors of the true processes. As an alternative to the inverse probability of censoring weighted estimating functions considered by Ghosh and Lin (2002), we consider estimation methods based on pseudo-Poisson analyses using weighted Poisson estimating functions or weighted counts in the spirit of Scheike et al. (2008). Simulation studies are carried out to assess the validity of the pseudo-Poisson procedures in finite samples and to compare their performance to the Ghosh-Lin estimation procedure. For illustration, we provide an analysis of recurrent skeletal complications in a palliative trial involving breast cancer patients with bone metastases (Section 1.3.3).

Mao and Lin (2016) considered a mean function model for the expected total number of events (fatal or non-fatal) up to time t. This marginal model can also be used for analysis in trial settings (I) where fatal events are often classified as disease-related or non-disease related. In that case the mean function, $\mu_{\widetilde{E}}(t) = \mathbb{E}(N_{\widetilde{E}}(t))$, may address only recurrent non-fatal events or disease-related fatal events (Furberg et al., 2022). We focus on this setting, and evaluate the interpretability of Mao-Lin estimands using the large sample framework similar to that employed for the Ghosh-Lin estimands. Recommendations for randomized trials with recurrent and terminal event processes

like those in Figure 1.1(e) and Figure 1.1(f) are provided.

1.4.3 Theme 3: Utility-based analyses for complex outcomes

In type (II) trial settings experimental treatments are typically designed to target a specific non-fatal event. In many other settings treatment may also affect other non-fatal events as well as fatal events. The third theme of this thesis is devoted to estimands for type (I) clinical trial settings involving disease processes with multiple events. Treatment effects in such trials have traditionally been assessed based on analyses of the time to the fatal event, the time to the first non-fatal or fatal event, or on the cumulative number of non-fatal and fatal events (Solomon et al., 2017; Marso et al., 2016). There has however been increasing interest in methods that incorporate not only individuals' entire sample paths into the assessment of treatments, but also the severity of different event types. One such method is utility-based (Gelber et al., 1989, 1995; Cook et al., 2003).

First explored by Gelber et al. (1989) in the context of the Q-TWiST criteria, utility-based approaches offer a nuanced assessment of treatment effects by assigning a numerical utility score to each time unit spent in a particular state and accumulating scores for each individual across the duration of the trial. This approach produces marginal features and estimands that admit causal interpretations, as desired. We critically discuss other ranking and weighting approaches, including the win ratio (Pocock et al., 2012; Oakes, 2016) and the restricted mean time in favor of treatment (Mao, 2023), and demonstrate how these approaches are linked.

To facilitate the use of utility-based estimands (1.2.9) in real world practice and to enable testing, we develop a variance estimator for the expected cumulative utility score (1.2.4) based on the infinitesimal jackknife procedure (Jaeckel, 1972; Therneau, 2024). Simulation studies investigate the empirical rejection rates of the tests under different data-generating mechanisms, including Markov and mixed-Markov models. An application is given to a platelet transfusion trial in patients with chemotherapy-induced thrombocytopenia (Section 1.3.5).

1.4.4 Thesis structure

The remainder of this thesis is organized as follows. A motivating introduction to the research topic is given at the beginning of each chapter, followed by an introduction to notation, development of the required statistical methodology, and numerical investigations. Each chapter provides an illustrative application to a motivating trial and concludes with final remarks. A final discussion of this thesis and topics for future research are provided in Chapter 6.

Chapter 2

Multistate models as a framework for estimand specification in clinical trials of complex disease processes

2.1 Introduction

The randomized clinical trial is widely regarded as the best study design for evaluating an experimental intervention compared to standard care on disease processes. However, when the processes involve multiple types of events, recurrent events, or competing terminal events, it is often unclear how best to summarize treatment effects. In particular, it is challenging to specify a sufficiently informative one-dimensional estimand to summarize effects and be used as a basis for tests. In cancer trials (I), for example, patients are at risk of cancer progression and death, and there has been considerable discussion about the merits of responses based on progression alone, progression-free survival and overall survival (Booth and Eisenhauer, 2012). In cardiovascular (CV) trials (I), interest lies in prolonging overall survival but with recent advances in medical therapies and surgical interventions mortality rates are now relatively low. This has lead to increased interest in using responses based on recurrent myocardial infarction (MI), stroke, and hospitalization, in addition to time to CV-related death; non-CV deaths are typically handled as competing risks (Rufibach, 2019; Toenges et al., 2021; Furberg et al., 2022). The

conceptualization and observation of treatment effects can be further complicated by intercurrent events (Qu et al., 2021). Ways to define estimands and make treatment comparisons in the presence of intercurrent events have received a great deal of attention in recent years (Casey et al., 2021; Stensrud and Dukes, 2022). The ICH E9-R1 addendum defines an estimand in terms of five attributes (i.e. treatment, population, variable, intercurrent events and population-level summary), without linking the issues arising in clinical trials to specific models for the processes in question (ICH E9 Steering Committee, 2019). Numerous researchers and working groups have proposed estimands and guidelines for specific settings involving competing risks (Stensrud et al., 2020; Poythress et al., 2020; Young et al., 2020; Nevo and Gorfine, 2021; Deng et al., 2024), recurrent and terminal events (Andersen et al., 2019; Wei et al., 2021; Schmidli et al., 2023; Furberg et al., 2023; Janvin et al., 2024), introduction of rescue therapies (Ster et al., 2020), or treatment switching (Watkins et al., 2013; Latimer et al., 2019a,b; Manitz et al., 2022).

Much of the discussion regarding challenges in conceptualizing and specifying treatment effects often lacks detail concerning models and assumptions, which makes interpretation of the recommended estimands difficult. The challenges arise from the complexity of disease processes and related processes involved with the management of individuals in a randomized clinical trial, and from the desire to make causal statements concerning treatment effects that are based on comparisons of marginal process features, which are not dependent on post-randomization events. In addition, estimands that reflect causal effects without a meaningful interpretation in the real world in which the trial is conducted are frequently considered. The use of potential outcomes has proven powerful and popular for formalizing causal reasoning and analysis, but can lead to causal statements about effects in randomized trials that do not reflect the actual trial or patient care.

The goals of this chapter are two-fold. First, we argue for the importance of models that reflect the disease process and related events that impact the conduct of the trial in both the planning and analysis stages. Time is a basic element in trials on disease processes and so models based on stochastic processes are crucial; we discuss the particular utility of multistate models (Andersen et al., 1993; Cook and Lawless, 2018; Andersen and Ravn, 2023). The second goal is to present guiding principles for the specification of estimands for type (I)-(II) clinical trials involving complex disease processes. We consider settings involving competing risks and semi-competing risks in some detail, along with more complex processes involving multiple and possibly recurrent events. For the planning stage of a trial when primary and secondary analysis strategies are

considered, we emphasize the important role of models that account for the disease process, study protocol, and clinical interventions. In settings involving intercurrent events, we show how multistate models can be used to jointly represent disease and intercurrent event processes. The structure offered by this multistate representation can clarify the interpretation of potential target estimands, reveal what is needed to estimate them, and illuminate whether they have an interpretation in the real world.

The remainder of this chapter is organized as follows. In Section 2.2.1, we introduce notation for multistate models using illness-death and competing risks processes as illustrations, and give examples of functionals of process intensities which may serve as a basis for defining estimands. Guiding principles for the specification of estimands for clinical trials involving complex disease processes are developed in Section 2.2.2. A numerical investigation on the effects of omitted covariates in an intensity-based Cox analysis is provided in Section 2.2.3. Section 2.2.4 gives an introduction to general frameworks for the specification of marginal estimands. In Section 2.3, we discuss the time-to-event framework in more detail in the context of cumulative incidence function (CIF) regression, and provide numerical results concerning the effects of assumptions, model misspecification and the interpretation of estimands. Section 2.4 gives a brief introduction to the count framework. Section 2.5.1 illustrates the utility of multistate models for dealing with processes involving intercurrent events, followed by an illustrative application to the NASCET trial in Section 2.5.2. Section 2.6 contains additional remarks on potential outcomes, and concluding remarks are given in Section 2.7.

2.2 Estimands for interventions in event history processes

In a typical phase III clinical trial, treatment groups are formed by randomizing individuals to receive either an experimental or control treatment (Fisher, 1925). Although the disease process may be complex, a primary objective for estimation and testing of treatment efficacy is to compare some process feature in the two treatment groups. A process feature is typically related to some outcome or event, for example, the total time spent in a given state or the time of entry to a specific state. In that case a feature is based on the response distribution, for example, the mean or median time to an event or the probability it occurs by a specific time. We define, for now, an *estimand* as

a one-dimensional measure of the difference in a certain process feature in the experimental and control groups; we provide a more nuanced definition of an estimand later in Section 2.3. The central clinical challenge involves specification of the process feature of greatest relevance, and then a clinical and statistical challenge is to determine how to specify an estimand comparing the feature in the two groups.

2.2.1 Multistate models, intensities and associated functionals

For further discussion, we consider disease processes that can be represented by the illness-death process in Figure 1.1(b) (Cook and Lawless, 2018), with state space $\mathcal{S} = \{0, 1, 2\}$. The related competing risks process shown in Figure 1.1(a) is discussed later. The illness-death process is widely applicable to oncology trials, with state 0 representing a cancer-free state following initial treatment, state 1 representing non-fatal events such as recurrence, relapse or progression, and state 2 representing death; individuals may enter state 2 with or without having experienced the non-fatal event. We let $Z(t)$ denote the state occupied by a generic individual at time $t \geq 0$, and assume that the process $\{Z(t), 0 < t\}$ begins in state 0 at $t = 0$, the time of randomized treatment assignment; thus, we assume $P(Z(0) = 0) = 1$. We let $Y_k(t) = \mathbb{1}(Z(t^-) = k)$ be a state occupancy indicator that equals 1 if an individual is in state $k \in \mathcal{S}$ at time t^- and 0 otherwise, and let $T := \inf\{t > 0 : Z(t) \neq 0\}$, $T_1 := \inf\{t > 0 : Z(t) = 1\}$ and $T_2 := \inf\{t > 0 : Z(t) = 2\}$ represent the event-free survival (EFS) time (or time spent in state 0), the time to the non-fatal event and the overall survival (OS) time (or time to death), respectively. We let X denote the binary treatment indicator as before, and let V denote other observable covariates that may affect the disease process. For simplicity, we initially assume that V is fixed, but we later consider settings where it may have time-varying components. The Markov transition intensity function for a $k \longrightarrow l$ transition is defined as

$$\lim_{\Delta t \downarrow 0} \frac{P(Z(t + \Delta t^-) = l \mid Z(t^-) = k, X, V)}{\Delta t} = Y_k(t)\lambda_{kl}(t|X, V) \ , \tag{2.2.1}$$

for $(k, l) \in \{(0, 1), (0, 2), (1, 2)\}$. The stochastic nature of the illness-death process is fully determined by the specification of all three transition intensities (Andersen and Keiding, 2012). We use the term *process feature* to mean any functional of the set of intensities.

The survivor function for the EFS time is given by

$$S(t|X,V) = P(T > t|X,V) = P_{00}(0,t|X,V) = \exp\left(-\int_0^t (\lambda_{01}(u|X,V) + \lambda_{02}(u|X,V))du\right), \tag{2.2.2}$$

where $P_{kl}(s,t \mid X,V) = P(Z(t) = l \mid Z(s) = k, X, V)$ is the probability of the process being in state l at time t, given that the process has been in state k at s, $s \leq t$. Since $P(Z(0) = 0|X,V) = 1$, the state k occupancy probability can be derived as

$$P(Z(t) = k|X,V) = P_{0k}(0,t|X,V) \ .$$

In addition, the OS probability is

$$P(T_2 > t|X,V) = P_{00}(0,t|X,V) + P_{01}(0,t|X,V) \ ,$$

where

$$P_{01}(0,t|X,V) = \int_0^t P_{00}(0,u|X,V) \ \lambda_{01}(u|X,V) \exp\left(-\int_u^t \lambda_{12}(s|X,V)ds\right)du \ .$$

The distribution function for T_1 is often called the cumulative incidence function (CIF) for the non-fatal event and is given by

$$F_1(t|X,V) = P(T_1 \leq t|X,V) = \int_0^t P_{00}(0,u|X,V)\lambda_{01}(u|X,V)du \ . \tag{2.2.3}$$

It should be noted that $F_1(t|X,V)$ approaches a limit less than one as t increases.

2.2.2 Principles for defining estimands

Andersen and Keiding (2012) provide important guidance on the analysis of life history processes and the specification of meaningful functionals. This has helped to shape the views on the specification of estimands presented in this thesis. Two tenets expressed by Andersen and Keiding (2012) are: *don't condition on the future* and *stick to the real world*. These tenets are highly relevant to the discussion in this chapter. To deal with the practicalities of actual trials we use models that represent the disease process as well as features such as intercurrent events. Responses

and features refer to particular aspects of a process that are of key interest. For example, the three-state illness-death diagram of Figure 1.1(b) may represent possible disease paths, but in a study aiming to reduce the occurrence of the intermediate event the time of entry to state 1 may be the response of interest. Figure 1.2 portrays a much more complicated setting where intercurrent events may occur.

We begin with an idealized setting in which individuals in a trial experience illness-death processes, with no premature dropouts before a common end of follow-up time C. The first step in specifying an estimand involves identifying an observable process $\{Z(t), 0 \leq t\}$ and specifying a process feature on which treatment comparisons will be based. This depends on the clinical meaning of the different states and the primary goals for the experimental intervention. In randomized trials (II) of palliative therapies, treatments might aim to reduce the occurrence of an adverse non-fatal event; for example Cook et al. (2009) and Section 1.3.3 describe trials involving the reduction of fractures and other skeletal events in patients with advanced stage IV breast cancer. In this case the response of interest could be entry to state 1, with the associated feature the $0 \longrightarrow 1$ intensity function $\lambda_{01}(t)$ or the cumulative incidence function $F_1(t)$. In oncology trials (I) aiming to prolong survival in patients with advanced cancer, state 1 might represent disease progression, and the overall survival time T_2 or disease-free survival time T is often considered as the response of interest (Rittmeyer et al., 2017; Powles et al., 2018). Once a process feature has been identified, the second step is the specification of an estimand β. One option is to take a specific time $\tau \leq C$ and a feature such as $F_1(\tau|X)$ or $P(T_2 \leq \tau|X)$, and to define β as the ratio or difference in the selected process feature across the two treatment groups. Examples include

$$\beta = F_1(\tau|X=1) - F_1(\tau|X=0) \;\text{ or }\; \beta = \frac{P(T_2 \leq \tau|X=1)}{P(T_2 \leq \tau|X=0)} \,.$$

This requires minimal assumptions, but the resulting *model-free* estimands will naturally vary according to the chosen value of τ, which may be contentious. Estimands that focus on a specific time τ should be accompanied by sensitivity analyses of other time points. If the process feature is a function of time such as $F_1(t|X)$ or $\lambda_{01}(t|X)$, a second option is to adopt modeling assumptions that provide a one-dimensional estimand. In the case of entry to state 1, an estimand β is frequently

defined by assuming

$$F_1(t|X=1) = \exp(\beta) F_1(t|X=0) \ , \ \text{ or}$$
$$\lambda_{01}(t|X=1) = \exp(\beta) \lambda_{01}(t|X=0) \ .$$

This is not strictly necessary since we could, for example, define β as the maximum of $|F_1(t|X=1) - F_1(t|X=0)|$ over $(0, \tau]$. However, power calculations for tests of no treatment effect used in planning studies are most conveniently formulated in terms of *model-based* estimands, and we will focus on them.

Process features and related estimands can be distinguished according to whether they do or do not condition on previous process history. We refer to the former as dynamic features; process intensity functions are of this type. We term features that do not condition on the process history as marginal features; the cumulative incidence function $F_1(t|X)$ is an example. We will also refer to features as having a *marginal* interpretation or a *dynamic* interpretation. Causal inference for treatment effects based on marginal features is in principle straightforward: it is facilitated by the random allocation of treatment to trial participants at time $t = 0$. Transition intensities are defined conditional on the previous process history, and estimands such as intensity ratios are geared towards process dynamics (Aalen, 2012). Intensity functions are crucial to a full understanding of a disease process, and to an understanding of specific marginal features, but for reasons we expand on below, are not suited for simple causal inference based on randomization. They are, however, important for a deeper understanding of causal mechanisms which in processes must address time-varying factors and random events. For example, in view of the relationship (2.2.3), the effects of X on both process intensities $\lambda_{01}(t|X)$ and $\lambda_{02}(t|X)$ must be considered in order to understand what has produced an observed effect in $F_1(t|X=0)$ and $F_1(t|X=1)$. Chapter 9 in Aalen et al. (2008) provides a thoughtful survey of aspects and approaches to causality in the context of event history processes.

We support the common position that primary assessment of treatment effects in randomized trials should be based on marginal process features and estimands. Beyond this, we argue that they should possess three fundamental properties:

1. *An estimand should represent the difference between treatment groups with respect to a clinically relevant marginal feature.*

2. *Features and estimands should be interpretable in the real world, meaning they are based on an observable process.*

3. *Estimands should not be sensitive to uncheckable assumptions; models on which an estimand is based should be assessed using available data.*

The *first principle* ensures that any findings will have clear relevance for treatment decisions. While this appears a straightforward objective it can be challenging to satisfy this principle in the face of intercurrent events. The *second principle*, that estimands be interpretable "in the real world", may seem self-evident but estimands that do not satisfy this condition are often adopted. For a competing risks process, the subdistribution hazard function (Fine and Gray, 1999) corresponding to $F_1(t|X)$ is a well-known functional that violates the second principle. Individuals who have previously experienced event-free death are included in the risk set at time t when defining and estimating the subdistribution hazard but such individuals are not genuinely at risk of the non-fatal event in the real world (Andersen and Keiding, 2012; Putter et al., 2020). Moreover, most of the recent work in causal inference defines estimands based on a potential outcome for each individual subject, corresponding to each of the treatments under study; in the real world of most trials, only one of the treatments is received by an individual, making the other potential outcomes "counterfactual". In this case inter-individual treatment effects such as the difference in the two potential outcomes are not observable in the real world. Many proposed solutions to challenges with intercurrent events also fail to satisfy principle 2: they involve specification of higher dimensional potential outcomes, taking analyses further away from the real world. Estimands of this type include the survivor average causal effect and other estimands in Comment et al. (2019), Stensrud et al. (2020), Xu et al. (2020) and Young et al. (2020). The *third principle* ensures that one can assess the adequacy of an underlying model and related assumptions, and their effect on the validity of conclusions. The more complex the disease and disease management processes are, the more difficult it is to select a single primary estimand that adequately summarizes treatment effects. Secondary analyses that enhance understanding of intervention effects and more thoroughly characterize the overall response to treatment are an important aspect of clinical trials.

2.2.3 Problems with conditioning on features of the process history

Before moving on to marginal estimands for specific processes, we first review problems with causal interpretations of intensity functions and other conditional process features. In particular, estimands such as intensity ratios for experimental versus control treatment groups do not have a simple causal interpretation. Although treatment X is randomly assigned at time $t = 0$ and is thus independent of other covariates V affecting the disease process, at a later time t the random allocation of X and independence of V no longer holds among those at risk for the event in question. Confounding induced by conditioning on post-randomization events that may also be responsive to the treatment is a well-known phenomenon; it has been comprehensively discussed by Hernán (2010), Aalen et al. (2015) and Martinussen et al. (2020) for the standard survival case, where the hazard function at time t conditions on being alive then. In illness-death processes, time-dependent confounding is more involved; for example the intensity function for the non-fatal event conditions on being alive and event-free, and confounding depends on the effect of V on both event types. Hazard or intensity ratios nevertheless remain popular in the analysis of randomized trials, so we illustrate the impact of time-related confounding in an illness-death process.

For simplicity, we let V be a binary covariate with $P(V = 1) = 0.5$ and independent of X due to randomization, i.e. $X \perp\!\!\!\perp V$. We let $P(X = 1) = 0.5$ and assume that the true illness-death process has intensity functions of the form $\lambda_{0k}(t|X, V) = \lambda_k \exp(\gamma_{kx}X + \gamma_{kv}V)$ for $k = 1, 2$, with time-homogeneous baseline intensities λ_k and regression coefficients γ_{kx}, γ_{kv}; thus V may affect the intensity for both the non-fatal event and event-free death, and both intensities are of proportional hazards form. An estimator $\widehat{\gamma}_{1x}$ of the conditional treatment effect on the non-fatal event intensity can be obtained by solving the Cox (1975) partial likelihood score equation for $\gamma_1 = (\gamma_{1x}, \gamma_{1v})'$ given by

$$U(\gamma_1) = \sum_{i=1}^{n} \int_0^\infty \overline{Y}_i(t) \left\{ \begin{pmatrix} X_i \\ V_i \end{pmatrix} - \frac{S^{(1)}(t, \gamma_1)}{S^{(0)}(t, \gamma_1)} \right\} dN_{i1}(t) \ ,$$

where $\overline{Y}_i(t) = Y_{i0}(t)\mathbb{1}(\min(C_{ri}, C) \geq t), dN_{i1}(t) = \mathbb{1}(T_{i1} = t)$ and

$$S^{(r)}(t, \gamma_1) = \sum_{i=1}^{n} \overline{Y}_i(t) \begin{pmatrix} X_i \\ V_i \end{pmatrix}^{\otimes r} \exp(\gamma_{1x}X_i + \gamma_{1v}V_i), \qquad r = 0, 1 \ .$$

The conditional joint distribution of X and V for individuals in state 0 at time t^- is

$$P(X,V|Z(t^-)=0) = \frac{P(Z(t^-)=0|X,V)P(X)P(V)}{\mathbb{E}_{XV}(P(Z(t^-)=0|X,V))} ,$$

and this cannot be factored into separate functions of X and V, so $X \not\perp\!\!\!\perp V|Z(t^-)=0$.

In addition, the marginal intensity function "averaged" over V has the form

$$\begin{aligned}\lambda_{01}(t|X=x,Z(t^-)=0) &= \mathbb{E}_V(\lambda_{01}(t|X=x,V)|X,Z(t^-)=0) \\ &= \lambda_1 \exp(\gamma_{1x}x)\mathbb{E}_V(\exp(\gamma_{1v}V)|X=x,Z(t^-)=0),\end{aligned}$$

where

$$P(V|X=x,Z(t^-)=0) = \frac{P(Z(t^-)=0|X=x,V)P(V)}{\mathbb{E}_V(P(Z(t^-)=0|X=x,V))}, \qquad x=0,1 .$$

As a result, the marginal intensity ratio is given by

$$\begin{aligned}IR_{01}(t) &= \frac{\lambda_{01}(t|X=1,Z(t^-)=0)}{\lambda_{01}(t|X=0,Z(t^-)=0)} \\ &= \exp(\gamma_{1x})\left[\frac{\exp(\gamma_{1v})P(V=1|X=1,Z(t^-)=0)+P(V=0|X=1,Z(t^-)=0)}{\exp(\gamma_{1v})P(V=1|X=0,Z(t^-)=0)+P(V=0|X=0,Z(t^-)=0)}\right] .\end{aligned} \tag{2.2.4}$$

If $\gamma_{1v} \neq 0$ this is a function of time t, so the intensity ratio for experimental versus control individuals is not a constant. Consequently there are two key points: (i) the marginal intensity ratio (2.2.4) at a given time $t > 0$ cannot be interpreted causally if confounders V are omitted, and (ii) if the true intensities conditional on both X and V are proportional, so that there is a scalar treatment effect γ_{1x} that applies at all times t, (2.2.4) is time-dependent and so does not yield a scalar estimand.

2.2.3.1 Impact of covariate omission in a Cox intensity-based model

To illustrate these key points numerically, we set $e^{\gamma_{1x}} = 0.75, e^{\gamma_{2x}} = 0.9$ and considered three values $0, 2$ and 3 for each of $e^{\gamma_{1v}}$ and $e^{\gamma_{2v}}$. We chose the administrative censoring time to be $C = 1$ and for each set of regression coefficients determined λ_1 and λ_2 so that $P(T \leq 1) = 0.6$ and so

that the conditional probability of entry to state 1 given $T \leq 1$ was either $0.05, 0.4, 0.6, 0.8$ or 1.0. We incorporated loss to follow-up by introducing an independent random right-censoring time C_r which followed an exponential distribution with rate ρ, with ρ set to satisfy $\pi_R = P(C_r < T_1 | T_1 < \min(T_2, 1)) = 0.2$. Thus, 20% of the non-fatal events occurring before C are censored due to early withdrawal. As is common in real-world practice of clinical trials, we omit V and let

$$\lambda_{01}(t|X) = \tilde{\lambda}_1 \exp(\phi_1 X) \tag{2.2.5}$$

be the working Cox proportional hazards model for the non-fatal event intensity, with ϕ_1 estimated by solving the partial likelihood score equation

$$U(\phi_1) = \sum_{i=1}^{n} \int_0^{\infty} \overline{Y}_i(t) \left\{ X_i - \frac{S^{(1)}(t, \phi_1)}{S^{(0)}(t, \phi_1)} \right\} dN_{i1}(t) = 0 \ ,$$

where $S^{(r)}(t, \phi_1) = n^{-1} \sum_{i=1}^{n} \overline{Y}_i(t) X_i^r \exp(\phi_1 X_i), r = 0, 1$. As we describe below, the true hazard ratio varies with time t so model (2.2.5) is misspecified. However, the maximum likelihood estimator $\widehat{\phi}_1$ converges in probability in large samples to a limit ϕ_1^* (Struthers and Kalbfleisch, 1986), obtained as the solution to

$$\mathbb{E}(U(\phi_1)) = \int_0^{\infty} \left\{ \mathbb{E}\left(\sum_{i=1}^{n} \overline{Y}_i(t) X_i dN_{i1}(t) \right) - \frac{\mathbb{E}(S^{(1)}(t, \phi_1))}{\mathbb{E}(S^{(0)}(t, \phi_1))} \mathbb{E}\left(\sum_{i=1}^{n} \overline{Y}_i(t) dN_{i1}(t) \right) \right\} = 0,$$

where expectations are taken with respect to the true illness-death, censoring and covariate processes. Under the assumptions of $\{Z(t), 0 < t\} \perp\!\!\!\perp C_r | X, V$, $P(C_r > t | X, V) = P(C_r > t)$ and $X \perp\!\!\!\perp V$ due to randomization, the expectations needed to compute ϕ_1^* are

$$\mathbb{E}(S^{(r)}(t, \phi_1)) = P(C_r > t) \sum_{x=0,1} \sum_{v=0,1} x^r \exp(\phi_1 x) P(X = x) P(V = v) P_{00}(0, t^- | X = x, V = v) \ ,$$

$$\mathbb{E}(\overline{Y}_i(t) X_i^r dN_{i1}(t)) = P(C_r > t) \cdot$$
$$\left[\sum_{x=0,1} \sum_{v=0,1} x^r P(X = x) P(V = v) \lambda_{01}(t | X = x, V = v) P_{00}(0, t^- | X = x, V = v) \right], \ r = 0, 1 \ .$$

Figure 2.1 plots $100(\phi_1^* - \gamma_{1x})/\gamma_{1x}$, the percent relative difference between ϕ_1^* and γ_{1x}, the effect of X in the true process conditional on X and V, as a function of $\exp(\gamma_{1v})$ (upper panel) and $\exp(\gamma_{2v})$ (lower panel). When V affects the fatal event only (i.e. $\gamma_{1v} = 0$), ϕ_1^* equals the

conditional treatment effect γ_{1x} in the true process; see (2.2.4) and the bottom set of panels in Figure 2.1. For all other scenarios, the limiting value ϕ_1^* differs from γ_{1x}. In settings where the non-fatal event occurs most often (right hand panels, bottom row), the relative difference is about 10%. The magnitude of the difference increases with stronger effects of V on the fatal and non-fatal event intensities. The relative difference depends on the probability of being event-free at time t^-, which is given by $(3.2.2)$, and on the proportion of non-fatal events. In the extreme case of $P(T_1 < T_2 | T \leq 1) = 1$, $\lambda_{02}(t|X, V)$ is zero; the competing risk setting coincides with the standard survival case then and results align with the points made in Aalen et al. (2015). If on the other hand the non-fatal event happens rarely, the relative difference is smaller but its dependence on γ_{2v} becomes stronger. As noted by Struthers and Kalbfleisch (1986), we also find that ϕ_1^* depends on the censoring distribution.

Figure 2.2 depicts the true marginal intensity ratio $IR_{01}(t)$, given in (2.2.4), over the time interval $[0, 1]$ for different values of γ_{1v}, γ_{2v} and $P(T_1 < T_2 | T \leq 1)$. As noted earlier, unless $\gamma_{1v} = 0$, $IR_{01}(t)$ varies with time so that the marginal proportional hazards model is misspecified; this further complicates the interpretation of the effective estimand ϕ_1^*. The magnitude of variation in the intensity ratio over time depends on the effects of V on the intensity functions for the non-fatal and fatal event; the effects on ϕ_1^* seen in Figure 2.1 are related to this.

To summarize, intensity-based comparison of treatment groups is not recommended for the specification of estimands and primary analysis. Even if models adequately describe the observed data, intensities condition on post-randomization events and do not permit randomization-based causal interpretations. In addition, the almost inevitable presence of other factors V that affect the disease process further complicates a causal interpretation of treatment effects. For more comprehensive marginal models or intensity-based models used for secondary analysis, known baseline covariates V can be included, with inferences about the effect of X now adjusted for V. Even then there may exist unknown or unobserved covariates that affect the process, so some caution should be exercised when interpreting treatment effects.

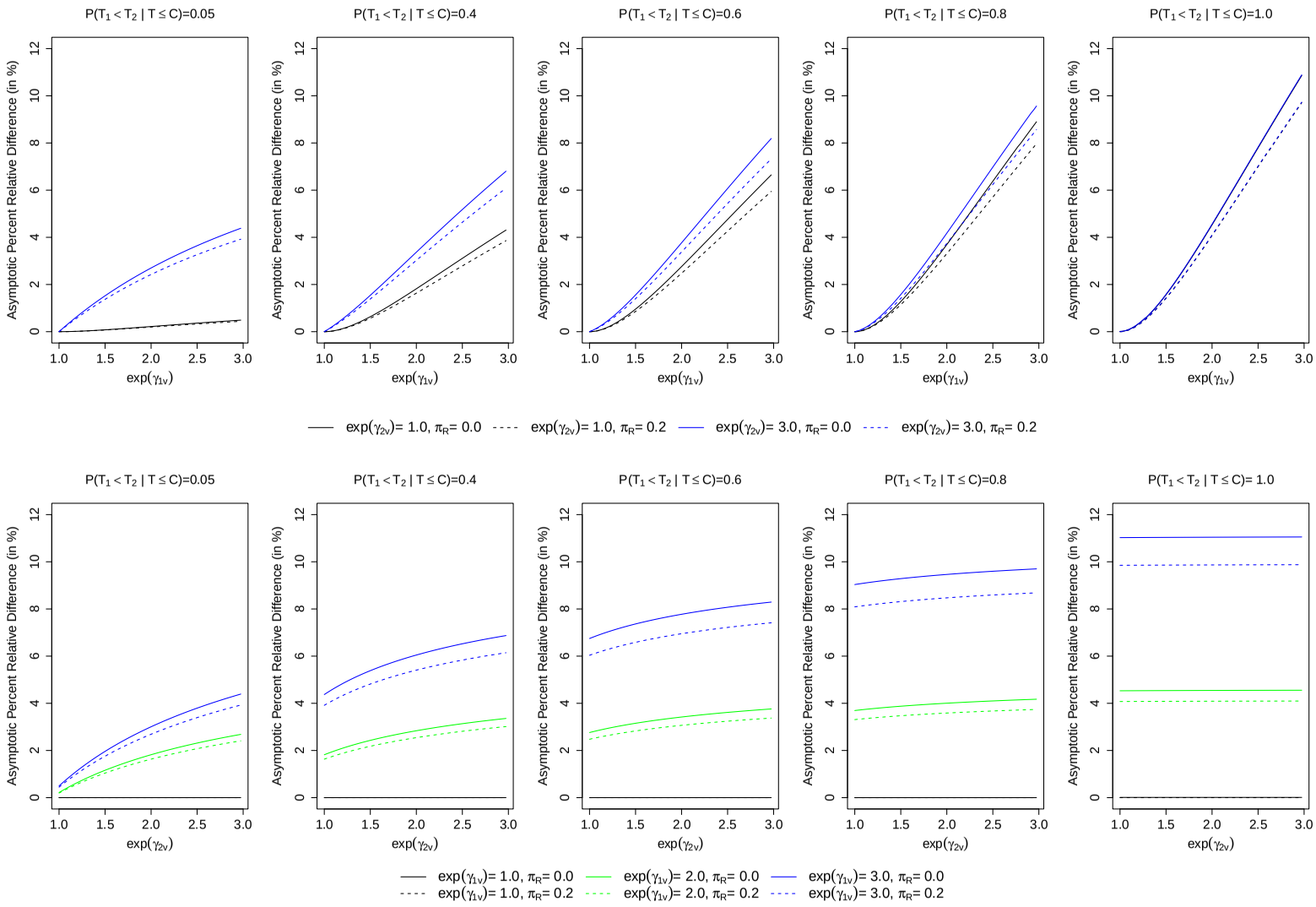


Figure 2.1: Asymptotic percent relative difference $100(\phi_1^* - \gamma_{1x})/\gamma_{1x}$ where ϕ_1^* is the limiting value of the estimator from a cause-specific Cox model for the $0 \longrightarrow 1$ transition when the covariate V is omitted; $C = 1, P(X = 1) = 0.5, P(V = 1) = 0.5, P(T \leq C) = 0.6, \gamma_{1x} = \log(0.75), \gamma_{2x} = \log(0.9)$.

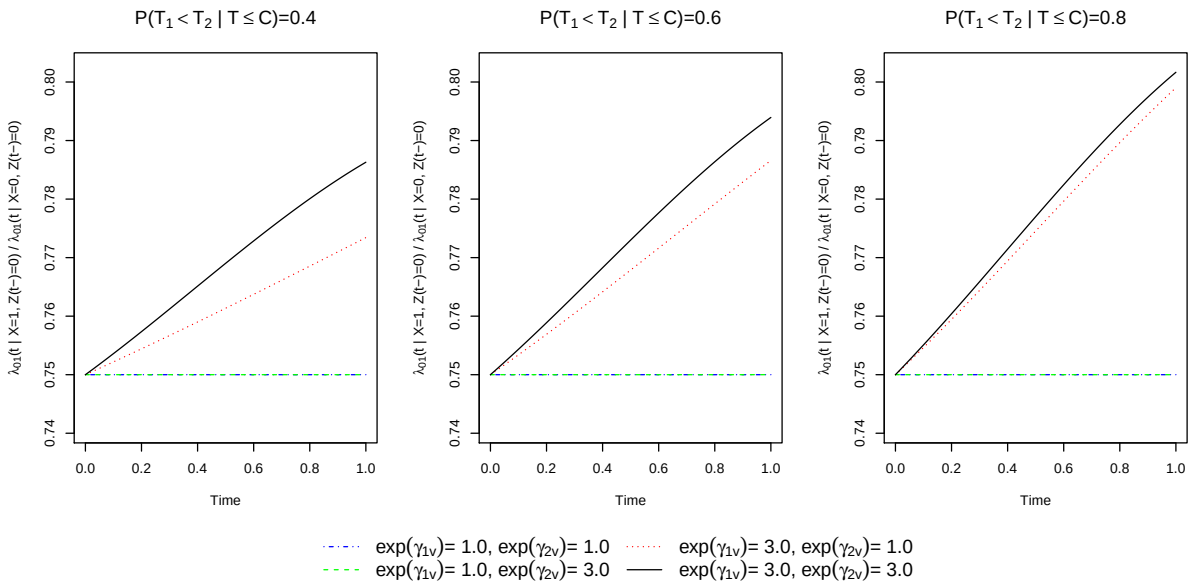


Figure 2.2: Plots of $IR_{01}(t)$ when the cause-specific hazards functions $\lambda_{01}(t|X,V)$ and $\lambda_{02}(t|X,V)$ depend on both X and V; $C = 1, P(X = 1) = 0.5, P(V = 1) = 0.5, P(T \leq C) = 0.6, \gamma_{1x} = \log(0.75), \gamma_{2x} = \log(0.9)$.

2.2.4 General frameworks for expressing marginal treatment effects

We return to marginal estimands and consider frameworks for specification of marginal effects. Most proposed estimands are based on the distribution $P(T_E \leq t)$ of the time to some event E, or more generally, on the cumulative mean function $\mathbb{E}(N_E(t))$ for type E events. For an illness-death process, the times $T_E \in \{T_1, T_2, T\}$ are all used in certain settings and related estimands can be formulated within the time-to-event framework (1.2.6) of Section 1.2.3.1 for $N_E(t) = \mathbb{1}(T_E \leq t)$, see also (1.2.8) for the special case of $T_E = T_1$. We note that the common practice of assuming a Cox proportional hazards model for time T_2 produces a model of the form (1.2.6). In particular,

$$P(T_2 > t|X) = P(T_2 > t|X = 0)^{\exp(\zeta X)} \tag{2.2.6}$$

is equivalent to a GLM for $P(T_2 \leq t|X)$ with $g(u) = \log(-\log(1-u))$. As discussed, we argue that the treatment effect ζ in such a model should be interpreted for causal purposes in terms of

the associated cloglog model (1.2.6) and not as a hazard ratio. A similar model is also often used for the EFS time T although these are not generally compatible.

Quantiles or expected values of times $T_E \in \{T_1, T_2, T\}$ in "restricted" form are also used; the restricted mean time event-free over a specified time interval $(0, \tau]$ is $RMT_0(\tau) = \mathbb{E}(\min(T, \tau))$. Estimation of state occupancy probabilities $P(Z(t) = k|X)$ by treatment group is also appealing in many cases and can be related to distributions of event times. For example, $P(T > t|X) = P(Z(t) = 0|X)$ and hence $RMT_0(\tau|X) = \int_0^\tau P(Z(u) = 0|X)du$. Similarly, $P(T_2 > t|X) = P(Z(t) \in \{0, 1\}|X)$.

A third related framework has been introduced in Section 1.2.3.2 and is utility-based: utility scores are assigned to each state $k \in \mathcal{S}$, and the average cumulative utility score over a specified time period $[0, \tau]$ is then considered as the marginal feature for estimand specification. The scalar estimand in (1.2.9), $\beta(\tau) = D(\tau|X = 1) - D(\tau|X = 0)$, can thus be given a descriptive causal interpretation. We reiterate that the utility formulation (1.2.4) covers features based on times to event T_E as special cases, where utility rates are either 0 or 1 for each state. In an illness-death process, for instance, $RMT_0(\tau|X)$ is given by (1.2.4) with $u_0 = 1$, $u_1 = u_2 = 0$; the restricted mean time spent in state 1, $RMT_1(\tau|X)$, is likewise given by (1.2.4) with $u_0 = 0, u_1 = 1, u_2 = 0$. We address utility-based estimands in detail in Chapter 5.

2.3 Time-to-event framework: estimands for illness-death processes based on cumulative incidence functions

This section takes a closer look at estimands based on marginal features of a process, and discusses estimability and model assumptions. We continue considering illness-death settings where times T_1 or T are responses of interest. In their case it is sufficient to focus on the associated competing risks process $\{Z(t), 0 < t\}$ in Figure 1.1(a), where $1 \longrightarrow 2$ transitions are not considered. The exit time from state 0 is $T = \inf\{t > 0 : Z(t) \neq 0\}$ and the indicator $\varepsilon = Z(T) \in \{1, 2\}$ records the type of event. We let X and V denote treatment indicator and other covariates, as before. The intensity functions in a competing risks process are often called cause-specific hazard functions:

$$\lim_{\Delta t \downarrow 0} \frac{P(T \in [t, t + \Delta t), \varepsilon = k|T \geq t, X, V)}{\Delta t} = \lambda_{0k}(t|X, V) \, , \qquad k = 1, 2 \, , \tag{2.3.1}$$

and they completely determine the process. Models for the intensities have to be chosen, but methodology and software for fitting and checking models is widely available (Cook and Lawless, 2018). The event time T has survivor function $S(t|X,V)$ given by (2.2.2) and the cumulative incidence function for $k=1$ is given by (2.2.3); the cumulative incidence function for event-free death (i.e. $k=2$) is defined as $F_2(t|X) = \int_0^t S(u^-|X)\lambda_{02}(u|X)du$. Each of the three features depend on both cause-specific hazard functions.

To illustrate the specification of estimands based on marginal features, we consider quantifying the difference between $F_1(t|X=1)$ and $F_1(t|X=0)$; for convenience we continue to denote the cumulative incidence function as F_1, though we condition now on X alone. One approach is to estimate the two functions nonparametrically (Cook and Lawless, 2018) and then to use some one-dimensional measure of their difference. Zhang and Fine (2008) proposed comparison at a specified time point $\tau > 0$. Such estimands do not provide a full picture of the effect of treatment on $F_1(t)$, nor do estimands such as differences in restricted mean time or the maximum of $|F_1(t|X=1) - F_1(t|X=0)|$ over a time period $(0,\tau]$.

2.3.1 Generalized linear models for cumulative incidence functions

An alternative way to obtain a one-dimensional (or scalar) estimand is by making parametric assumptions about the difference between $F_1(t|X=1)$ and $F_1(t|X=0)$. The most common approach has been to use models that are of generalized linear form, as in (1.2.8):

$$g(F_1(t|X)) = \alpha(t) + \beta X \ .$$

The monotonic baseline function $\alpha(t) = g(F_1(t|X=0))$, with $\alpha(t) \downarrow -\infty$ as $t \downarrow 0$, may be modelled parametrically (Jeong and Fine, 2006, 2007) or nonparametrically and the regression coefficient $\beta = g(F_1(t \mid X=1)) - g(F_1(t \mid X=0))$ is a marginal estimand, as desired. Several link functions $g(u)$ on $(0,1)$ have been considered in the literature (Gerds et al., 2012); common choices are $g(u) = \log(u)$, $\text{logit}(u)$ and $\log(-\log(1-u))$, often called the complementary log-log, or cloglog transform. Assumed models should of course adequately represent treatment group differences; if this is not the case the interpretation of estimates $\widehat{\beta}$ is affected, as we illustrate in Section 2.3.2.

Parametric or semiparametric estimation for arbitrary transformation models for $F_1(t|X)$ can be based on so-called direct binomial (DB) estimating functions (Scheike and Zhang, 2007) or on pseudo-value methods (Klein and Andersen, 2005). These methods avoid modeling $F_2(t|X)$, whereas maximum likelihood estimation requires such a model when some individuals are still in state 0 at the end of follow-up. The most widely used model is based on the cloglog function; it can be used in analysis based on a weighted partial likelihood estimating function targeting the hazard function for the subdistribution $F_1(t|X)$ (Fine and Gray, 1999). As noted previously, this "hazard" function does not have a real world interpretation (Andersen and Keiding, 2012). However, although the associated estimand β cannot be interpreted as a hazard ratio, it can be interpreted in terms of the cloglog transformation model, and the estimation procedure of Fine and Gray (1999) is valid in this context. These methods are reviewed in Section 3.3.1, and Cook and Lawless (2018) provide references to software. In addition to the `R` package `cmprsk` mentioned there, the `R` functions `coxph()` and `cifreg()` now handle the FG method (Therneau, 2024; Holst et al., 2016).

The adequacy of a transformation model for $F_1(t|X)$ can be checked in a given setting by plotting g-transformed nonparametric Aalen-Johansen estimates for each treatment group. We have found that the cloglog model represents the results quite well for many oncology trials, in which state 1 represents disease progression or the onset of complications (see e.g. Section 3.5). One can explore different transformations through parametric families of functions $g(t, \nu_k)$ specified by an unknown parameter ν_k; this is easier if parametric assumptions are also made about the baseline function $\alpha(t)$. Jeong and Fine (2007) considered a generalized Burr form

$$F_k(t \mid X) = 1 - \left(1 + \frac{\exp(\alpha_k(t, \phi_k)) \exp(X\beta_k)}{\nu_k}\right)^{-\nu_k}, \quad k = 1, 2 \tag{2.3.2}$$

and developed maximum likelihood procedures for estimation of $(\phi_k, \nu_k, \beta_k)'$. Model (2.3.2) includes parametric versions of the logit and cloglog transformation models ($\nu_k = 1$ and $\nu_k \longrightarrow \infty$ respectively) as special cases. Since $F_k(t|X) < 1$, we need $\alpha_k(t)$ to approach a finite limit as t increases; β_k must also be constrained to keep $F_k(t|X = 1) < 1$. In addition $F_1(t|X) + F_2(t|X)$ cannot exceed one, which may require additional constraints. In practice a transformation model should provide an adequate representation up to some maximum follow-up time τ and these constraints may not have much impact in some settings. Other families of models have also been proposed, some of which automatically constrain the CIFs to sum to one (Gerds et al., 2012).

Finally, given the importance of estimands based on marginal process features in primary estimation and testing of treatment effects, we recommend that the study planning stage include careful consideration of process intensities, and how treatment may affect them. This promotes an understanding of what types of marginal features and models to employ and informs decisions concerning sample size and duration of follow-up. In the next section we provide some numerical illustrations of the connection between intensities and marginal features, especially between cause-specific hazard functions and cumulative incidence functions in competing risks processes.

2.3.2 Illustrative calculations for cumulative incidence function regression

To provide insight into the impact of process intensities on marginal process features, some illustrative calculations are presented. A pragmatic point of view is that models only approximate reality, and when an estimand β is defined based on assumptions about a process (that is, a model), along with a method of estimating it, the true estimand is actually the limiting value β^* of the estimator $\widehat{\beta}$ as the number of subjects n becomes arbitrarily large. The value β^* is sometimes referred to as the *least false* parameter (Grambauer et al., 2010), but more precisely it is the parameter value for which the assumed model family is "closest" (in the expected log likelihood or Kullback-Leibler sense) to the true process or distribution in question.

Definition 1 *Under model misspecification the actual estimand is not β but $\beta^\star$, the value for which the estimating functions used to obtain the estimator of β (i.e. $\widehat{\beta}$) has expectation zero with respect to the true data-generating process.*

2.3.2.1 Illustrations concerning model assumptions

We present a simple illustration for the illness-death setting, with the true process assumed to have intensity functions $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X)$ for $k = 1, 2$. We again focus on the cumulative incidence function $F_1(t|X)$ for the non-fatal event, and consider transformation models (1.2.8) where $g(u) = \log(u)$ or $\log(-\log(1-u))$. The interpretation and relevance of β depends on whether the transformation model adequately approximates the difference in $F_1(t|X = 0)$ and $F_1(t|X = 1)$. It is also of interest how $\beta^\star$ is related to the baseline intensities and treatment effects

in the true process. We let $P(X = 1) = 0.5$ and considered true processes with $\exp(\gamma_1) = 0.75$ and $\exp(\gamma_2) = 0.9, 1.0$ and 1.1; thus the treatment reduces the intensity for the non-fatal event by one quarter and gives either a mild decrease, no change, or a mild increase in the intensity for the fatal event. We took the administrative censoring time to be $C = 1$ and for each value for γ_2, set λ_1 and λ_2 so that $P(T \leq 1)$ was either 0.2 or 0.6, and so that the probability of entry to state 1, given $T \leq 1$, was either $0.4, 0.6$ or 0.8. We note that the assumed form for the intensities is for illustration; as illustrated in Section 2.2.3.1, the presence of other covariates V that affect intensities can produce non-proportional intensities when only X is modeled.

Adequacy of the transformation models is assessed by computing the true values of $F_1(s_r|X)$ for 20 equi-spaced values of s_r in (0,1) and $X = 0, 1$. For each transformation, we calculated $g(F_1(s_r|X = 1)) - g(F_1(s_r|X = 0))$ for $r = 1, \ldots, 20$ and display the results in Figure 2.3. Departures from a constant line are indicative of model misspecification. In most cases, the cloglog transformation model is a fairly good approximation, especially when the treatment has no effect on the fatal event, whereas the $\log$ transformation model approximates the differences in $F_1(t|X)$ less well. It can be shown that at $t = 0$ the difference in the transformed functions is $\gamma_1 = \log(0.75) = -0.288$ under either model. For $g(u) = \log(-\log(1 - u))$, adequacy of the transformation model varies with the probability of entry to state 1 and the treatment effect γ_2 on the fatal event intensity; the model improves with increasing $P(T_1 < T_2|T \leq 1)$ and as the magnitude of γ_2 decreases. For $g(u) = \log(u)$, adequacy of the model likewise varies with the parameter setting. We have found that these results persist across various plausible settings, and we restrict our attention to $g(u) = \log(-\log(1 - u))$ in the remainder of this chapter.
Appendix A.1 shows how to determine the adequacy of the cloglog transformation model when the true process has time-inhomogeneous intensities.

2.3.2.2 Illustrations concerning estimability and interpretability

We considered estimation of β for the cloglog transformation model using the two main methods: the method of Fine and Gray (1999) (FG) and direct binomial (DB) estimation (Scheike et al., 2008). Competing risks data under the true process with intensity functions $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X)$ as specified above were simulated as follows: for each of $X = 0, 1$ the event time $T|X$ was generated from an exponential distribution with rate $\lambda_{01}(t|X) + \lambda_{02}(t|X)$, and

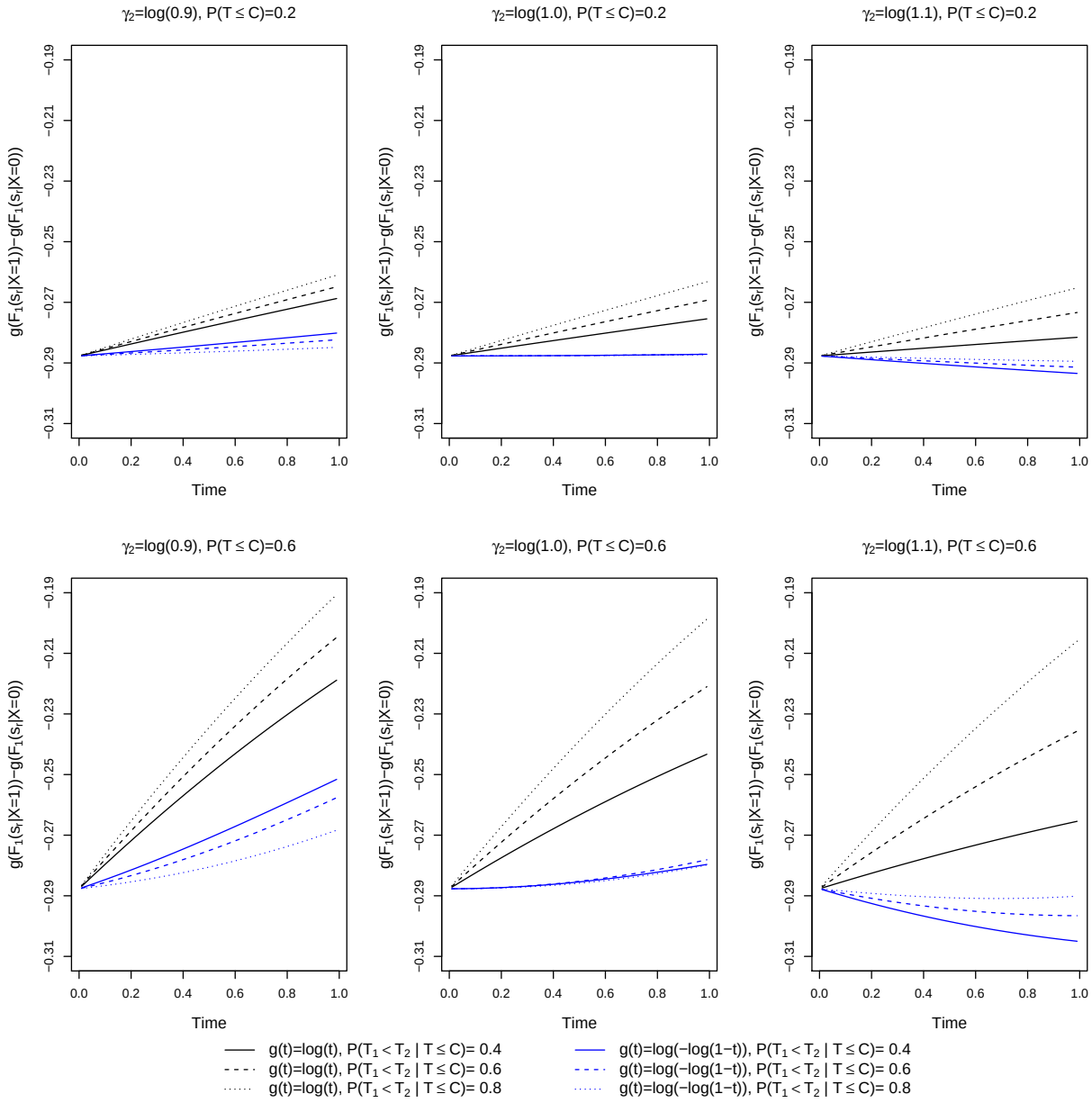


Figure 2.3: Plots of $g(F_1(s_r|X = 1)) - g(F_1(s_r|X = 0))$ for 20 equi-spaced values of s_r in $(0, 1)$ and different values of $P(T \leq C), P(T_1 < T_2|T \leq C)$ and γ_2 under the log and cloglog transformation models; $C = 1, P(X = 1) = 0.5, \gamma_1 = \log(0.75) = -0.288$.

given $T = t$, the event type was drawn from a single binomial trial with non-fatal event probability $\lambda_{01}(t|X)/(\lambda_{01}(t|X) + \lambda_{02}(t|X))$. We also considered a random withdrawal time C_r, which was generated from an exponential distribution with rate $\rho > 0$. We set ρ so that $\pi_r = P(C_r < T_1 | T_1 < \min(T_2, 1))$ was either $0.0, 0.1$ or 0.2, with $\pi_r = 0.0$ corresponding to administrative censoring only. In scenarios with $\pi_r \neq 0.0$, this gave the net censoring time $\min(C_r, C)$. We simulated data sets with $n = 1000$ observations and for each we obtained estimates of β using the FG approach with and without weight stabilization, and by DB estimation based on 6 equi-spaced time points in $(0, 1)$. We used the R functions `crr()` and `comp.risk()` to fit the models, combined with unstratified Kaplan-Meier estimation to get an estimate of $G(t) = P(C_r > t)$. Results are shown in Table 2.1 for $n_{sim} = 2000$ simulation runs. For each estimation procedure we report the limiting value $\beta^\star$, the mean of the estimates $\widehat{\beta}$, their empirical standard error (ESE), the average standard error (ASE) provided by the `crr()` or `comp.risk()` functions for the FG and DB regression models respectively, and the empirical coverage probability (ECP) for $\beta^\star$ of the nominal 95% confidence intervals. In Sections 3.3.1.1 and 3.3.1.2 we carefully outline how the limiting values β^* under either estimation procedure can be obtained. The stabilized FG estimator $\widehat{\beta}^{FG}_{stab}$ and the corresponding estimand β^{FG*}_{stab} can be found as the solutions to (3.3.5) and (3.3.8) with $w_i(t)$ replaced by $\mathbb{1}(C_i \geq \min(T_i, t))G(t)/G(\min(T_i, t))$, respectively. If the censoring model is correctly specified and thus $\widehat{G}(t) \longrightarrow G(t)$, the limiting value $\beta^{FG\star}$ does not depend on the censoring distribution. The limiting value $\beta^{FG\star}_{stab}$, on the other hand, varies with the censoring distribution due to weight stabilization.

Since $\beta^\star_{FG}$, $\beta^\star_{FG,stab}$ and $\beta^\star_{DB}$ solve different expected estimating equations, we do not expect these limiting values to be the same. However, they are in close agreement, and the size and direction of their departure from γ_1 is consistent with the departures from a horizontal line seen in Figure 2.3. In all scenarios, the average estimates under both estimation methods closely approximate the corresponding limiting values, which are close to the value $\gamma_1 = \log(0.75)$. Figure 2.3 provides a simple basis for the interpretation of estimates $\widehat{\beta}$, unlike so-called average hazard ratios. We also see close agreement between the empirical and average standard errors, and that the empirical coverage probabilities are close to the nominal level. Thus, for the scenarios considered here, the cloglog transformation model provides a reasonable approximation for the effect of X on T_1, and both FG and DB estimation of β in the cloglog model perform well.

$P(T \leq C)$	$P(T_1 < T_2 \mid T \leq C)$	π_r	***Estimation method***										
			Fine-Gray						***Direct Binomial Regression***				
			$\beta^{\star}_{FG}$	$\beta^{\star}_{FG,stab}$	mean$(\widehat{\beta}_{FG,stab})$	ESE	ASE	ECP	$\beta^{\star}_{DB}$	mean$(\widehat{\beta}_{DB})$	ESE	ASE	ECP
0.6	0.4	0.0	-0.2532	-0.2532	-0.2550	0.1308	0.1300	95.2	-0.2669	-0.2672	0.1437	0.1429	95.2
		0.1	-0.2532	-0.2551	-0.2579	0.1341	0.1369	95.6	-0.2669	-0.2693	0.1479	0.1486	95.1
		0.2	-0.2532	-0.2572	-0.2574	0.1474	0.1451	95.0	-0.2669	-0.2645	0.1567	0.1558	95.2
	0.6	0.0	-0.2601	-0.2601	-0.2594	0.1067	0.1059	94.6	-0.2714	-0.2704	0.1180	0.1161	94.4
		0.1	-0.2601	-0.2617	-0.2630	0.1113	0.1116	95.1	-0.2714	-0.2727	0.1199	0.1214	95.8
		0.2	-0.2601	-0.2635	-0.2660	0.1172	0.1182	95.3	-0.2714	-0.2737	0.1278	0.1280	95.2
	0.8	0.0	-0.2708	-0.2708	-0.2685	0.0923	0.0917	94.6	-0.2780	-0.2780	0.1014	0.1002	95.3
		0.1	-0.2708	-0.2719	-0.2721	0.0955	0.0966	95.4	-0.2780	-0.2781	0.1045	0.1055	95.5
		0.2	-0.2708	-0.2730	-0.2681	0.1020	0.1024	95.4	-0.2780	-0.2715	0.1102	0.1119	95.6

Table 2.1: Asymptotic and empirical properties of Fine-Gray and direct binomial regression estimators for β in cloglog transformation models for $F_1(t|X)$; $C = 1, P(X = 1) = 0.5, \gamma_1 = \log(0.75) = -0.288, \gamma_2 = \log(0.9), n = 1000$ individuals, $n_{sim} = 2000$ simulation runs. Weights in FG and DB approaches are based on covariate-independent censoring and under the assumption that the censoring model is correctly specified.

2.4 Count framework: estimands for recurrent and terminal event processes based on marginal rate functions

Other disease processes involve non-fatal events that may occur repeatedly. In cardiovascular (CV) research, for example, events such as myocardial infarction (MI), stroke and admission to hospital due to heart failure may each recur (Schmidli et al., 2023; Toenges et al., 2021; Furberg et al., 2022). Competing terminal events such as death or loss to follow-up are also common. The risk of recurrent skeletal complications in the bone metastases trial in Section 1.3.3 is, for instance, terminated by death. Section 6.4 of Cook and Lawless (2007) and Cook et al. (2009) give a thorough treatment of such processes involving recurrent and terminal events. Furberg et al. (2022), Toenges et al. (2021) and Mao and Kim (2021) provide recent reviews of methods in the context of randomized trials.

In such settings, marginal estimands based on cumulative counts for recurrent non-fatal events, or counts comprised of either fatal or non-fatal events are widely used. In connection with the LEADER trial, Furberg et al. (2022) consider a recurrent composite event with components consisting of non-fatal stroke, non-fatal MI or CV death. Such a composite recurrent event is motivated by a desire to synthesize information about treatment effects across multiple types of events, but as with any composite endpoint clear interpretation requires modeling treatment effects on the component event types. If primary interest lies in the effect of treatment on non-fatal recurrent events, the proportional means model of Ghosh and Lin (2002) can be used. It is based on the marginal rate function for the non-fatal event, recognizing that further non-fatal events cannot occur after a terminal event; it is analogous to a cumulative incidence function in this way, but accommodates recurrent events. The ratio of rate or mean functions satisfies the principles for estimands given in Section 2.2.2 and has a descriptive causal interpretation. Mao and Lin (2016) proposed a multiplicative model based on the rate function and corresponding mean function for the composite counting process for fatal and non-fatal events combined. Both models can be formulated within the count-based framework (1.2.6) of Section 1.2.3.1 as

$$\log(\mu_E(t|X)) = \log(\mu_0^E(t)) + \beta X \ ,$$

where $\mu_E(t|X) = \mathbb{E}(N_E(t)|X)$ is the mean function for the counting process $\{N_E(t), 0 < t\}$ of type E events, and $\mu_0^E(t) = \mathbb{E}(N_E(t)|X = 0)$ is the corresponding baseline mean function

for individuals treated with the control treatment. Estimands based on the ratio of such mean or rate functions, $\beta = \log(\mu_E(t|X=1)/\mu_0^E(t))$, satisfy the principles of Section 2.2.2 and have a descriptive causal interpretation. By including terminal events in the composite counting process, the Mao-Lin approach may be suited to settings where treatment is needed to avoid both serious non-fatal events and terminal events. There are however the usual caveats: (i) intensity-based models should be fitted in secondary analyses to gain understanding of the mechanisms leading to any observed treatment effect; Cox models and other regression models for terminal event intensities are useful in understanding connections between non-fatal and fatal events, and (ii) the proportionality of rate or mean functions inherent to the Ghosh-Lin and Mao-Lin models should be checked. We discuss estimands of treatment effect estimators from Ghosh-Lin and Mao-Lin models and their properties in more detail in Chapter 4.

2.5 Intercurrent events

Marginal features, and hence the treatment effect, are usually conceptualized in an idealized setting where individuals under study are compliant, receive medical care according to a prescribed (deterministic or well-characterized stochastic) treatment strategy, and complete follow-up. Under such circumstances the differences between treatment arms with respect to, say, the proportion of individuals experiencing an event by some time τ can be attributed to the treatment assigned at randomization. In practice, events such as termination of follow-up due to inefficacy or adverse reactions to treatment, the introduction of rescue medication, or treatment discontinuation or switching may occur over the course of follow-up, making the resulting process incompatible with the idealized setting. Such events are examples of "intercurrent events". In Section 1.1.3 we have classified intercurrent events as type 1A, type 1B and type 2 intercurrent events depending on whether they preclude observation of the target event of interest or affect the interpretation of the target event. The core challenge with any intercurrent event is that marginal process features differ from those under the idealized setting, making the intended causal analyses challenging.

Consider a clinical trial where individuals are randomized to receive an experimental treatment ($X = 1$) or standard care ($X = 0$). To simplify discussion, we consider a trial involving an illness-death process depicted in Figure 2.4(a), where states 0, 1 and 2 represent being alive and

event-free, alive post-event, and dead, respectively. We assume individuals begin in state 0 at $t = 0$ with follow-up planned until an administrative censoring time C. We let $Z(t)$ denote the state occupied at time t with $Y_k(t) = I(Z(t^-) = k)$, $k = 0, 1$, and let $\mathcal{H}(t) = \{Z(u) : 0 < u < t, X\}$ be the process history up to time t^-. The $k \longrightarrow l$ transition intensity is denoted by $\lambda_{kl}(t|\mathcal{H}(t))$, and we also let

$$\lim_{\Delta t \downarrow 0} \frac{P(Z(t + \Delta t^-) = l | Z(t^-) = k, X)}{\Delta t} = q_{kl}(t|X) = \mathbb{E}_{\mathcal{H}(t)}\big(\lambda_{kl}(t|\mathcal{H}(t))\big) \tag{2.5.1}$$

denote the transition rate given X for $(k, l) \in \{(0, 1), (0, 2), (1, 2)\}$. The transition rate can be interpreted as the instantaneous probability of being in state l at time t given that the process has been in state k at t^-. It does not condition on any part of the history $\mathcal{H}(t)$ but averages over all possible histories prior to t^-. The target marginal estimand could be any of those discussed in Sections 2.2.4 and 2.3 including, for example,

(i) $\beta(\tau) = F_1(\tau|X = 1) - F_1(\tau|X = 0)$ given in Section 2.2.2,

(ii) $\beta(\tau) = S(\tau|X = 1) - S(\tau|X = 0)$ where $S(t|X = x) = P(Z(t) < 2|X = x)$ is the survival function for time to death, given $X = x$.

These quantities can all be estimated with Aalen-Johansen methods using separate nonparametric estimates of (2.5.1) for each treatment arm. Estimands based on transformation models of cumulative incidence functions as in (1.2.8) are also possible but we focus here on those listed above since they require minimal modeling assumptions.

Multistate models such as the illness-death process can be expanded to incorporate intercurrent events. Such a joint consideration of disease and intercurrent event processes facilitates a clear discussion of marginal features, the interpretation of possible estimands, and offers a framework for comparing analysis strategies. We represent type 2 intercurrent events by transient states in the expanded multistate diagram, and type 1 intercurrent events by absorbing states. Figure 2.4(b) depicts an expanded state space for a joint model involving the illness-death process of Figure 2.4(a) and a type 1 intercurrent event that precludes observation of transitions after it. Figure 2.4(c) represents a joint model with a type 2 intercurrent event that does not preclude further follow-up: state $0'$ is entered if the intercurrent event occurs in an individual who is event-free, while $1'$ is occupied if an individual is alive but has experienced both the non-fatal clinical event

and the intercurrent event; state $2'$ is entered upon death by individuals who have experienced the intercurrent event. We let $\{Z^\circ(t), 0 < t\}$ denote either the five-state process depicted in Figure 2.4(b) or the six-state process depicted in Figure 2.4(c), $\mathcal{H}^\circ(t) = \{Z^\circ(s), 0 < s < t, X\}$ the corresponding process history and define $Y_k^\circ(t) = I(Z^\circ(t^-) = k)$ for $k = 0, 1$ or $k = 0, 1, 0', 1'$, respectively. We let

$$\lim_{\Delta t \downarrow 0} \frac{P(Z^\circ(t + \Delta t^-) = l | Z^\circ(t^-) = k, \mathcal{H}^\circ(t))}{\Delta t} = \lambda_{kl}^\circ(t | \mathcal{H}^\circ(t)) \tag{2.5.2}$$

denote the intensity function for $k \to l$ transitions within Figure 2.4(b) or Figure 2.4(c), with the corresponding rate functions

$$\lim_{\Delta t \downarrow 0} \frac{P(Z^\circ(t + \Delta t^-) = l | Z^\circ(t^-) = k, X)}{\Delta t} = q_{kl}^\circ(t | X) \ , \tag{2.5.3}$$

for $(k, l) \in \{(0, 1), (0, 2), (0, 0'), (1, 2), (1, 1')\}$ or $(k, l) \in \{(0, 1), (0, 2), (0, 0'), (1, 2), (1, 1'), (0', 1'), (0', 2'), (1', 2')\}$.

In the next section we examine three types of scenarios involving intercurrent events: loss to follow-up (Section 2.5.1.1), discontinuation of the randomized treatment without loss to follow-up (Section 2.5.1.2), and introduction of rescue treatment without loss to follow-up (Section 2.5.1.3). In some clinical trials (e.g. NASCET trial), individuals may switch from the randomized treatment to the treatment of the other arm. For each setting we discuss issues arising from these events, including possible targets of inference (estimands) and strategies for estimating these quantities. We stress the need for careful thought about interpretation of the target estimand and the strength of assumptions required to estimate it. We focus on estimands relevant to the real world, which invariably leads to analyses within the intention-to-treat (ITT) framework (Montori and Guyatt, 2001).

2.5.1 Illustrative examples involving intercurrent events

2.5.1.1 Loss to follow-up

Premature loss to follow-up (LTF) is a type 1 IE or results from a type 1 IE. Figure 2.4(b) depicts the illness-death process with states added to represent termination due to LTF. For a more thorough analysis we need to also consider the joint model of Figure 2.4(c). In this framework, Lawless and Cook (2019) define conditionally independent LTF for multistate processes observed in cohort studies. The condition in their equation (4a) can be restated in the present notation as

$$\lim_{\Delta t \downarrow 0} \frac{P(Z^{\circ}(t+\Delta t^{-}) = l | Z^{\circ}(t^{-}) = k, \mathcal{H}^{\circ}(t))}{\Delta t} = \lambda_{kl}(t|\mathcal{H}(t)) \ , \tag{2.5.4}$$

for $(k,l) \in \{(0,1),(0,2),(1,2)\}$. This condition states that the transition intensities between states $0, 1$ and 2 are the same for the process under observation as they are for the process of interest depicted in Figure 2.4(a). If baseline covariates or marker processes governing the $j \to j'$ transitions are present, the intensities for $j \to j'$ transitions can be modeled to give insights into the types of individuals becoming lost to follow-up. These intensities can in turn be used for the construction of inverse probability of censoring weights for partially conditional rate-based analyses (Datta and Satten, 2002; Cook et al., 2009). This enables consistent estimation of the $q_{kl}(t|X)$ in (2.5.1) governing the process in Figure 2.4(a) and from this, weighted Aalen-Johansen estimates of state occupancy probabilities $P(Z(t) = k|X)$ can be obtained as described by Cook and Lawless (2018). This allows estimation of estimands such as (i)-(ii) described above.

We note that an additional more subtle requirement for conditionally independent loss to follow-up given in (4b) of Lawless and Cook (2019) is expressed here in terms of Figure 2.4(c) as

$$\lambda_{k'l'}(t|\mathcal{H}^{\circ}(t)) = \lambda_{kl}(t|\mathcal{H}(t)) \ , \tag{2.5.5}$$

for $(k,l) \in \{(0,1),(0,2),(1,2)\}$. This assumption implies that among those recruited and randomized to treatment, the intensities of the illness-death process are the same for those under follow-up as those who have been lost to follow-up. This assumption cannot be verified in the absence of data following loss to follow-up. Such data can sometimes be obtained through tracing studies (Lawless and Cook, 2019), but these are seldom done in clinical trials. Moreover, subjects in a trial receive care and treatment by protocols which do not apply after withdrawal and so

one would not expect disease dynamics to remain the same as if the person had remained in the trial. Thus, the main objective should be to protect against dependent loss to follow-up through violation of (2.5.4) by using weights as described above.

We stress that when some degree of premature LTF is a feature of a trial, estimands such as (i)-(ii) above must address LTF. In particular, consider the commonly used event times T, T_1, T_2 associated with Figure 2.4(a); here T is the exit time of state 0 and T_1, T_2 are the times of entry to states 1 and 2 respectively. Interpretation of $P(T > t|X) = P(Z(t) = 0|X)$ is clear in the setting of Figure 2.4(a), but when the observable process is that in Figure 2.4(b), $P(T > t|X) = P(Z^\circ(t) = 0|X)$ represents the probability that an individual is event-free, alive, and not lost to follow-up at time t. If LTF arises because of treatment discontinuation due to lack of efficacy or adverse effects, this represents a reasonable strategy. If LTF is driven by fixed or time-varying biomarkers associated with event occurrence, then LTF is dependent and incorporating LTF into a composite event (implicit in modeling the sojourn time in state 0 of Figure 2.4(b)) is one solution. If there is no evidence that LTF is marker-dependent, then LTF can simply be treated as independent right censoring and analyses can be based on Figure 2.4(a). Similar issues arise when considering $P(T_1 \leq t|X)$ and $P(T_2 \leq t|X)$, which in Figure 2.4(b) represent the probability that a subject has entered states 1 or 2 respectively prior to time t, while under follow-up.

2.5.1.2 Treatment discontinuation, no loss to follow-up

In some settings the intercurrent event may be a toxicity-related event leading to discontinuation of the assigned treatment, but with the individual remaining under follow-up, perhaps with a change in treatment. This constitutes a type 2 IE with observed data as in Figure 2.4(c). In this case, as in Section 2.5.1.1, we should incorporate this expanded illness-death process when defining features and estimands of interest. For EFS time T, for example, $P(T > t) = P(Z^\circ(t) \in (0, 0'))$ should be considered. For treatment group comparisons that relate to observable processes, we support ITT comparisons (Montori and Guyatt, 2001), which here might for example be based on $\beta(\tau) = P(T > \tau|X = 1) - P(T > \tau|X = 0)$, with X representing treatment assigned at randomization. This gives a more relevant estimand concerning treatment efficacy in most settings than another option that has been proposed, which is to artificially censor subjects when they

cease the initial treatment; in that case, Figure 2.4(b) would be used.

In many settings decisions about treatment switches are related to biomarkers which are used in monitoring individuals and in this case, expanded models may include a marker process. As an illustration, we consider a time-dependent marker $\{W(t), 0 < t\}$ where $W(t)$ reflects some aspect of disease severity, for example a marker of bone formation or destruction (e.g. bone alkaline phosphatase) in cancer patients with skeletal metastases. We let $\mathcal{H}(t) = \{Z(s), 0 < s < t\}$ be the history of the illness-death process in Figure 2.4(a) and $\mathcal{W}(t) = \{W(s), 0 < s < t\}$ be the history of the marker process alone. The expanded history $H(t) = \{Z(s), W(s), 0 < s < t, X\}$ includes the multistate and marker process histories along with the treatment covariate. We let $\lambda_{kl}(t|H(t))$ denote the transition intensities for (k, l) in $\mathcal{S}$ for the illness-death process of Figure 2.4(a), now accommodating dependence on the marker process. Likewise we let $\lambda^{\circ}_{kl}(t|H^{\circ}(t))$ denote the intensities for transitions among the pairs of illness-death states of Figure 2.4(c) for $(k, l) \in \{(0, 1), (0, 2), (1, 2), (0', 1'), (0', 2'), (1', 2')\}$ where $H^{\circ}(t) = \{Z^{\circ}(s), W(s), 0 < s < t, X\}$. Modeling the intensities $\lambda_{jj'}(t|H^{\circ}(t))$ for $j \to j'$ transitions corresponding to the occurrence of the IE can offer important insights concerning the types of individuals who cannot tolerate the study medication. A major difficulty in this situation, however, is the need to model the marker process in order to define and calculate marginal process features such as $P(T > t|X)$ that can be used to define estimands. A discussion of this area is beyond the scope of this thesis; Cook and Lawless (2022) provide an illustration.

2.5.1.3 Introduction of rescue treatment, no loss to follow-up

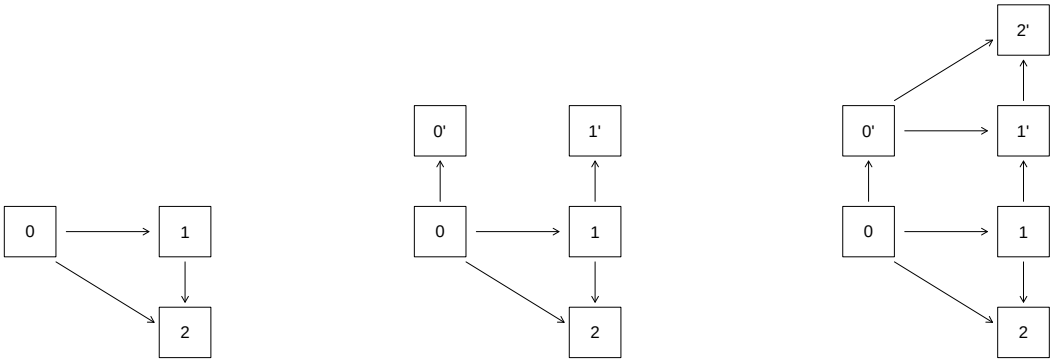
In cancer clinical trials it is common for individuals to receive rescue therapy, often after evidence of disease progression, with follow-up continuing after the rescue therapy has been introduced. The decision to prescribe rescue therapy will often be guided by marker processes, perhaps in combination with information on the illness-death process. In the case where rescue therapy is only introduced upon disease progression (entry to state 1) the $0 \to 0'$ transition intensity in Figure 2.4(c) is zero. In such a setting estimation of event-free survival probabilities or cumulative incidence of disease progression is unaffected but overall survival probabilities are affected. Modeling the $0 \to 0'$ and $1 \to 1'$ intensity functions will provide insight into the kinds of individuals who are prescribed rescue therapy. As in Section 2.5.1.2 the ITT principle considers

the full process in Figure 2.4(c) and is preferred for treatment comparison.

2.5.1.4 Treatment switching, no loss to follow-up

In some trials individuals may be switched from their assigned treatment to the treatment of the other arm. Most often individuals in the control arm are prescribed the experimental treatment; in cancer trials this is often done following cancer progression (Watkins et al., 2013). Figure 2.5(a) is a multistate diagram for an illness-death process $\{Z(t), 0 < t\}$ with separate states for two treatment groups, with overall survival the feature of interest. Henshall et al. (2016), Latimer et al. (2019a) and Latimer et al. (2019b) for example consider this setting, though not with expanded multistate models. We again prefer to formulate intensity-based models of observable event times and make clear and explicit assumptions about the disease and treatment process. In Figure 2.5(b) we add a state $1''$ which can be entered from state $1'$ (indicating progression under the control therapy) to reflect the introduction of the experimental treatment, resulting in the six-state multistate process $\{Z^\circ(t), 0 < t\}$. As in Sections 2.5.1.2 and 2.5.1.3 the intensity functions for the $1' \rightarrow 1''$ transition give insights into the kinds of individuals switching from the control to the treatment arm. In this setting we again favor use of the ITT principle in conjunction with nonparametric estimation of $P(Z^\circ(t) = 2|X)$.

We note that it is common to see analysis of individual or composite event times such as T, T_1 and T_2 based on Cox proportional hazards models, with the estimand β defined as the log hazard ratio for treatment to control groups. Even if we interpret β causally through an analogous transformation model as discussed earlier, this should be accompanied by an assessment of the model's adequacy. With either a model-based or nonparametric estimand as in (i)-(ii), secondary intensity-based analysis is crucial to an understanding of factors producing an observed marginal effect. Another type of secondary analysis is to examine conditional probabilities such as $P(Z^\circ(t) = 0|Z^\circ(t) \in (0, 1, 2), X)$ or $P(Z^\circ(t) = 2|Z^\circ(t) \in (0, 1, 2), X)$; these summarize event occurrence up to time t for persons not experiencing the IE by that time. Such probabilities can be estimated nonparametrically and although they are not suitable for causal interpretation because they condition on events that may be dependent on treatment or unobserved confounders, they and corresponding estimates of $P(Z^\circ(t) \in (0, 1, 2)|X)$ provide useful summary information. There has been discussion in the literature and in the ICH E9-R1 guidance document about the use

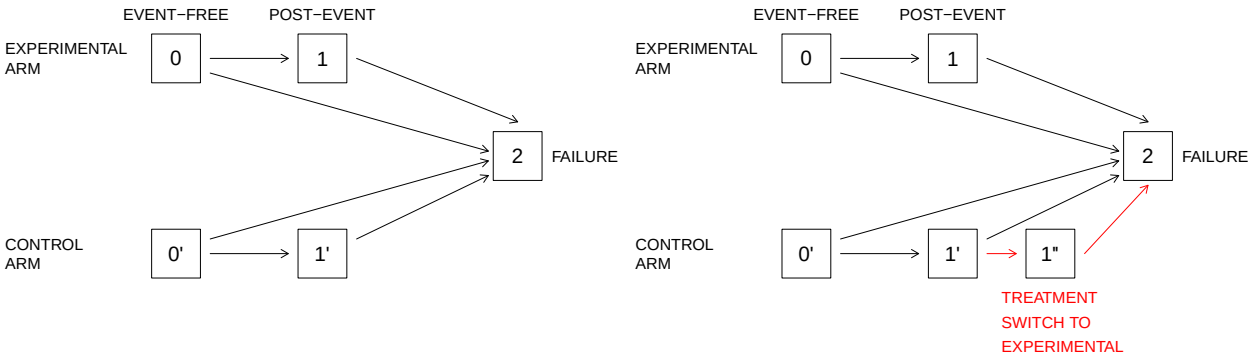


(a) An illness-death process ("idealized" setting).

(b) A joint illness-death and type 1 intercurrent event process.

(c) A joint illness-death and type 2 intercurrent event process.

Figure 2.4: Multistate diagrams for a simple illness-death process (panel (a)) and joint models for an illness-death process and type 1 (panel (b)) and type 2 (panel (c)) intercurrent events.



(a) Illness-death processes for two treatment arms.

(b) Illness-death processes with possible control to experimental arm crossover.

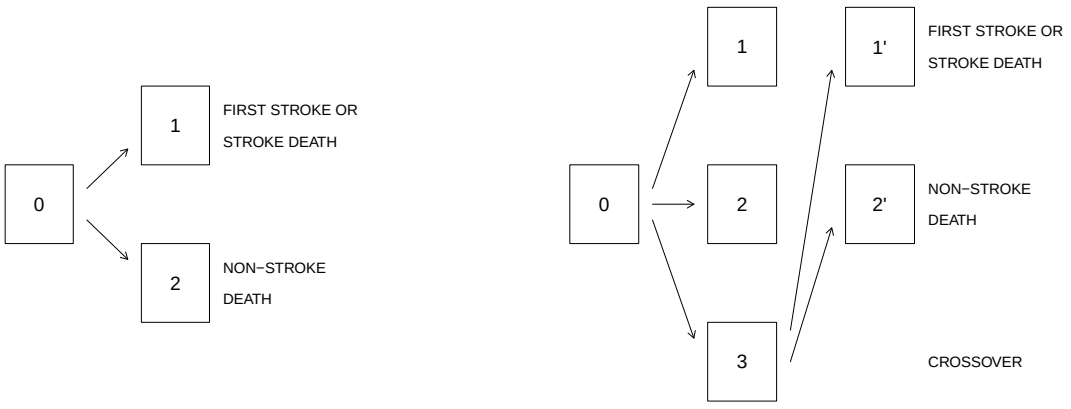
Figure 2.5: Multistate models for a process of interest (panel (a)) and an expanded process incorporating a post-randomization intervention for the control arm (panel (b)).

of principle stratification and estimation of the so-called survivor average causal effect. These methods condition on similar events but use counterfactuals, so do not meet our objective of real world interpretation; Section 2.6 discusses this further.

2.5.2 Surgical prevention of stroke-related events in the NASCET trial

To illustrate the joint modeling approach of Section 2.5.1, we consider data from the NASCET trial (Barnett et al., 1998; North American Symptomatic Carotid Endarterectomy Trial Steering Committee, 1991) in which we focus on individuals randomized to medical care $(n = 1118)$ or carotid endarterectomy $(n = 1087)$ in the stratum with moderate stenosis $(< 70\%)$. We consider for illustration an analysis involving the response of stroke-related events (i.e. stroke or stroke-related death). Figure 2.6(a) shows the multistate diagram for a competing risks process $\{Z(t), 0 < t\}$, where state 0 represents the condition of being stroke-free and alive, state 1 is entered upon the occurrence of a stroke or stroke-death, and state 2 is entered upon a non-stroke death. We note that entry to state 1 due to a non-fatal stroke may be followed by a stroke death or non-stroke death as illustrated in Figure 1.2, but we consider this simplified model, which is relevant for an analysis of the composite event of non-fatal stroke or stroke death; then the only competing event is non-stroke death.

As noted in Section 1.3.2, a number of individuals randomized to receive best available medical care crossed over to undergo carotid endarterectomy. The reasons recorded for this included non-fatal stroke, further narrowing of the carotid artery based on angiography examination, both stroke and angiographic progression, and other reasons. Figure 2.6(b) shows the more complex multistate process relevant for individuals in the medical arm which accommodates the crossover to carotid endarterectomy. Under the intention-to-treat framework the cumulative incidence functions for the model in Figure 2.6(a) can be estimated for both the surgical and medical arms; these are given in Figure 2.8(a) for the composite event of stroke or stroke-related death and Figure 2.8(b) for non-stroke death. Carotid endarterectomy can dislodge plaque into the circulatory system and cause stroke, so there is a perioperative period of high risk of stroke reflected by the steep increase in the cumulative incidence function estimate in Figure 2.8(a) for the surgical arm. Following this perioperative period the cumulative risk of stroke or stroke-death is lower in the surgical arm, so that ultimately there is a 6-8$\%$ absolute reduction in the risk of the composite



(a) A simple competing risk model for an intention-to-treat analysis.

(b) A multistate model accommodating information on medical-surgical crossover.

Figure 2.6: Multistate diagrams for analyses of stroke or stroke-related death and non-stroke related death in the NASCET trial (Barnett et al., 1998).

event in the surgical arm by 8 years. Figure 2.8(b) shows very similar cumulative incidence functions for non-stroke death in the two arms.

The ITT principle is normally justified on the basis that any interventions following randomization are part of routine care. To help in understanding and communicating the relevance of trial findings, it is necessary to clearly describe what constitutes routine care. For the NASCET study involving individuals deemed at moderate risk, the situation is more complicated. We focus here on the moderate risk stratum of NASCET which was treated as a separate, parallel study to one involving individuals designated as high risk due to having greater than 70% stenosis of the carotid artery at the time of randomization. The trial involving high risk patients started at the same time as the trial of moderate risk patients, but was stopped early when strong evidence emerged at an interim analysis of a benefit of carotid endarterectomy; this occurred during the conduct of the trial for the moderate risk patients. A second point is that individuals at moderate risk at the time of randomization to medical care may have experienced a progression of their carotid stenosis over the course of follow-up to the point that it exceeded 70% – this then qualified them as high risk patient. Following the publication of trial results for high risk patients, the standard of care

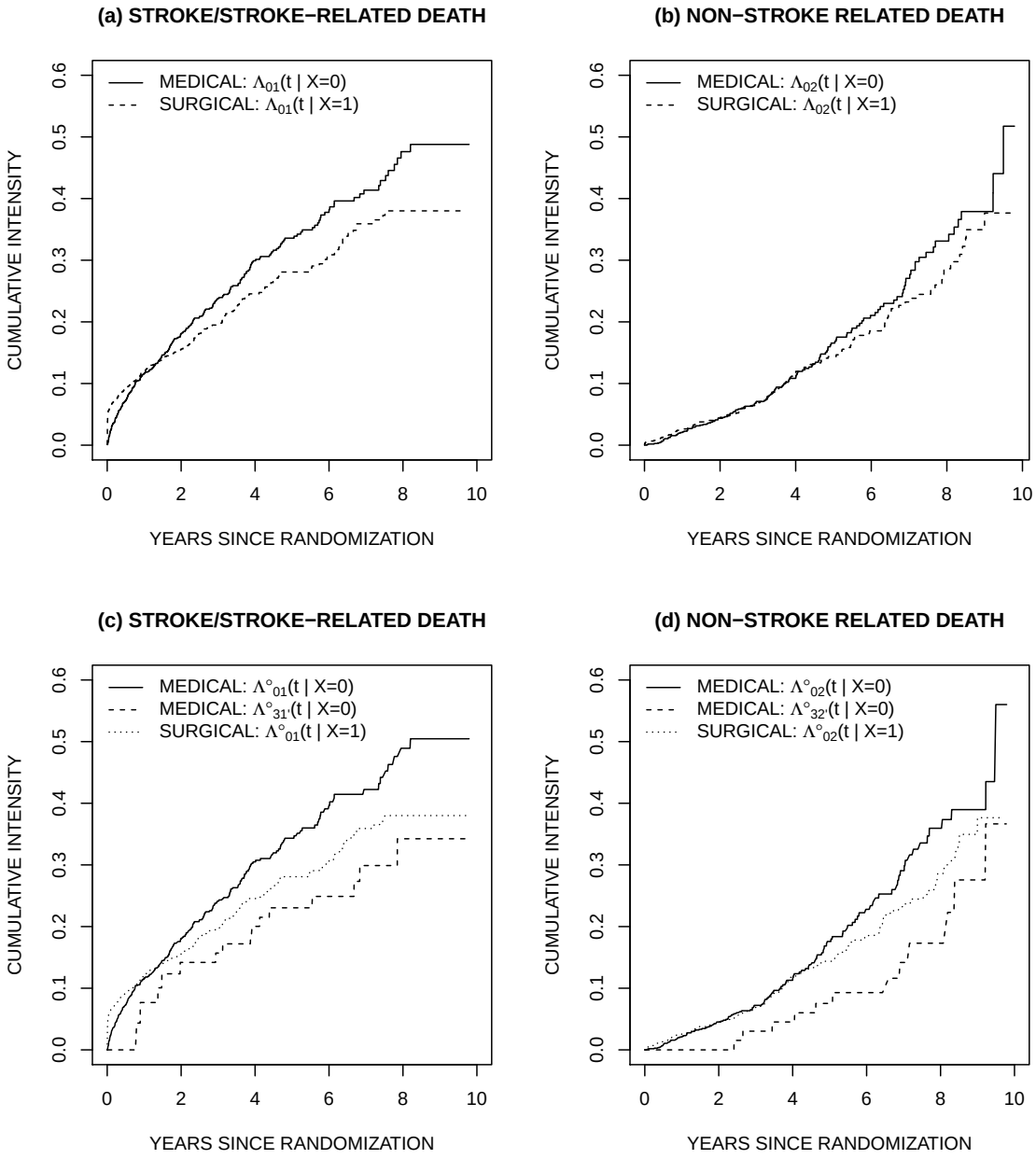


Figure 2.7: Nonparametric Nelson-Aalen estimates of the cumulative intensities for selected transitions based on Figure 2.6(a) in panels (a) and (b), and for selected transitions based on Figure 2.6(b) in panels (c) and (d).

changed to include carotid endarterectomy. As a result some patients randomized to medical care in the trial of moderate risk individuals may have experienced stroke when carotid endarterectomy was not part of standard of care, whereas others may have been randomized, progressed to qualify as high risk, and received carotid endarterectomy while stroke-free. The risk of stroke following surgery in this latter group of would be influenced by the surgical procedure. The results of an intention-to-treat analysis are difficult to interpret when the standard of care changes during the course of a study – a more detailed analysis is warranted to understand risks and associated effects.

Figure 2.6(b) shows the multistate diagram for a more detailed characterization of the course following randomization. States 1 and 2 again correspond to stroke or stroke-death (state 1) and non-stroke death (state 2) but state 3 is entered upon medical-surgical crossover for individuals in the medical arm. By orienting the states in this competing risk arrangement we note that state 1 is now entered only if the stroke or stroke-death occurs prior to surgical crossover, and likewise for the non-stroke death. Medical-surgical crossover can be followed by stroke/stroke-death or non-stroke death so states $1'$ and $2'$ represent these events. Letting $\{Z^{\circ}(t), 0 < t\}$ denote the multistate process in Figure 2.6(b), we note that the Aalen-Johansen (AJ) estimate of the transition probability matrix facilitates a decomposition of the cumulative incidence function estimates in Figures 2.8(a) and (b). If $X = 1$ for an individual randomized to receive surgery and $X = 0$ if they are randomized to medical care, then we note that $P(Z^{\circ}(t) = 1|X = 0) + P(Z^{\circ}(t) = 1'|X = 0)$ is the cumulative incidence function for stroke/stroke-death estimated in Figure 2.8(a). The separate Aalen-Johansen estimates of $P(Z^{\circ}(t) = 1|X = 0)$ and $P(Z^{\circ}(t) = 1'|X = 0)$ in Figure 2.8(c) show that the excess risk of stroke/stroke-death in the medical arm evident in Figure 2.8(a) is made up of risk for those not crossing over, and risk from those crossing over to receive surgery. We cannot attribute this added risk to the surgery itself due to confounding by indication; we may expect an elevation in risk shortly after surgery and a reduction in risk following a perioperative period, but a critical point is that those crossing over to receive surgery are selected in a dynamic way in response to their disease course.

To explore this more fully, we plot Nelson-Aalen estimates of the cumulative transition intensities in Figure 2.7 for the multistate models of Figure 2.6. The Nelson-Aalen estimates in Figures 2.7(a) and (b) correspond to those of Figure 2.6(a) and we again see evidence of the perioperative period of elevated risk following surgery, with similar cumulative intensities for non-stroke death. The Nelson-Aalen estimates of the cumulative transition intensities in Figure 2.7(c) are for transitions

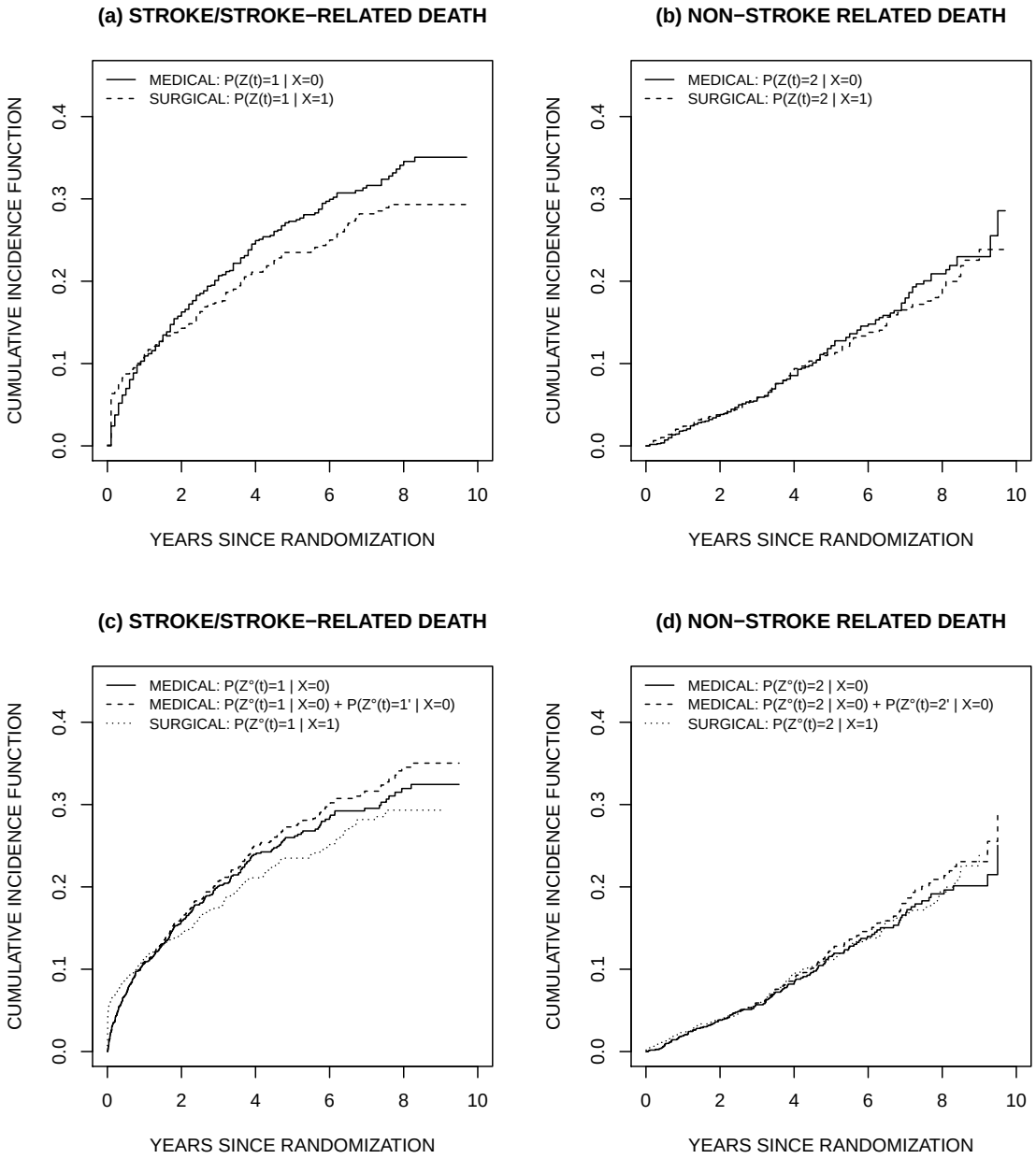


Figure 2.8: Nonparametric Aalen-Johansen estimates for the cumulative incidence functions for stroke/stroke-death and non-stroke death in Figure 2.6(a) (panels (a)-(b)) and Figure 2.6(b) (panels (c)-(d)); Aalen-Johansen estimates are also used for the state occupancy probabilities in panels (c) and (d).

in Figure 2.6(b). Interestingly, the slope of the estimated $0 \longrightarrow 1$ cumulative intensity for the surgical arm and the $3 \longrightarrow 1'$ cumulative intensity for those in the medical arm following surgical-crossover are very similar; the steepest cumulative intensity estimate is for the medical patients who did not crossover. A similar decomposition is given in Figure 2.7(d) for the endpoint of non-stroke death.

Intensity-based treatment comparisons could be based on regression modeling but the crossing cumulative intensities seen in Figure 2.7(a) mean that proportional cause-specific hazards models will not be suitable. A detailed analysis would make use of a model allowing crossover between treatment arms and this would ideally incorporate information on the disease course including blood pressure, cholesterol measurements and angiographic assessment of carotid stenosis. Data on such variables are unfortunately not available so we do not explore this, but note that there is much ongoing work on the assessment of randomized treatments when hazards or cumulative incidence functions cross, and when treatment switching occurs.

2.6 Remarks on potential outcomes

Potential outcomes have played a central role in the development of causal inference theory and methods over the last several decades (Rubin, 2005; Hernán and Robins, 2020), with their origins in cross-sectional or simple longitudinal settings. The potential outcome framework conceptualizes the existence of outcomes for each treatment under consideration for each individual. In a two treatment trial this leads to a pair of potential outcomes which could, conceptually, be compared for each individual in the trial. In trials where each individual receives just one treatment, the pairs of potential outcomes and "causal" measures such as their difference are thus not observable in the real world. The potential outcome corresponding to the treatment received is revealed, while the other is counterfactual. Some find potential outcomes useful for conceptualizing independence conditions, and this is reasonable in settings where one could, both in theory and in practice, randomize individuals to one of two treatments under study (Hernán and Robins, 2020). We note that thinking in terms of potential outcomes is increasingly common in connection with clinical practice, but in this thesis' view this does not line up in most settings with clinicians' views. For example, Bornkamp et al. (2021) say "assume a treating physician is deciding on a treatment to

prescribe. Ideally she would make that decision based on knowledge on what the outcome for the patient would be if given the control treatment, [...], and what the outcome would be under test treatments [...]". The term "ideally" notwithstanding, this thesis's view is that physicians think and counsel patients not in terms of fixed potential outcomes under alternative treatment options but in terms of probabilities that are based on data from trials and observational studies. For example, a physician might inform a breast cancer patient following surgery that the (estimated) probability of a recurrence during the next five years is 0.10 with no adjuvant hormone therapy and 0.08 with adjuvant therapy.

The counterfactual framework can also lead to specification of target estimands of dubious scientific relevance when used in conjunction with the concept of principal strata. Consider the problem of assessing the effect of a new intervention versus standard care on an outcome that can only be measured in individuals who are alive; an example is the measurement of neonatal outcomes among infants born in studies of different techniques for in vitro fertilization (Snowden et al., 2020). Rubin (2006) proposed estimands based on women who would experience a live birth under either of two treatments under study; this is referred to as the survivor average causal effect, with those surviving under either treatment comprising one of the four so-called "principal strata". Aside from concerns about counterfactual outcomes not being observable in the real world, such subgroups are not identifiable (observable) from the available data. Such issues have been discussed by others. As noted by Lipkovich et al. (2022) for example, "a major challenge of using principal stratification [...] is its counterfactual nature that requires strong assumptions to be able to identify and estimate treatment effect within principal strata". Moreover, as noted by Scharfstein (2019), the survivor average causal effect does not convey the effect of treatment on all of those randomized, and the subset of the population in the principle stratum of interest may be small. Issues of competing risks also arise in settings with intercurrent events such as early withdrawal, introduction of rescue medication, crossover to the complementary arm, and death. Hernán and Scharfstein (2018) provide a thoughtful discussion of the ICH E9-R1 addendum and cautionary remarks about potential outcomes and principal strata. Finally, Andersen and Keiding (2012) caution against conditioning on the future in analysis of life history processes if interest lies in maintaining clear real world interpretation of estimands; the survivor average causal effect violates this tenet. Preferable strategies for dealing with intercurrent events include ignoring their occurrence in an intention-to-treat analysis targeting the effect of a "treatment policy" (i.e. the

effect of prescribing one treatment versus the other at the time of study entry). The intercurrent event can also be incorporated into a composite endpoint, which can make sense if the event of interest and the intercurrent event are both undesirable.

Disease processes and patient management are by nature dynamic, and time plays a central role. Aalen (2012) among others have highlighted the importance of time, and of modeling data in ways that recognize the evolutionary nature of disease processes. This way of thinking is aligned with the principle of "staying in the real world" but, as we have discussed, complicates the identification of relevant marginal features which can deliver causal inferences based on randomization to treatment. The dynamic nature of processes also makes the notion of pre-existing counterfactual outcomes artificial; probabilistic models are needed for real world processes.

2.7 Discussion and recommendations

Randomized controlled trials that address disease processes are challenging to plan, conduct and analyze. In the simplest situations an intervention is designed to target a specific adverse event, but in many settings it might also affect other events. In this case there may be alternative choices for the primary event or outcome used for treatment comparisons. In addition, other disease-related events may prevent or interfere with the observation of the primary outcome, or necessitate a change in treatment. A main message of this chapter is the importance of formulating comprehensive models that incorporate features of the disease process as well as intercurrent events related to follow-up and interventions. Such models, in conjunction with medical information, aid in the formulation of protocols and specification of estimands. In addition, they are crucial for secondary intensity-based analysis which provides a deeper understanding of treatment effects and interpretation of estimands based on marginal process features. The challenge of dealing with intercurrent events can sometimes be lessened by good study design and execution; for example, if one can follow individuals to an administrative censoring time then issues of loss to follow-up are avoided. When the standard of care can be defined unambiguously problems in the interpretation of study findings might be avoided, but in international multi-center trials investigators may view having variation in standard of care good for generalizability. In such cases study findings are interpretable in a population with a similar mix of standard of care strategies, which may be

difficult to use in a given setting.

Reference is often made to robustness and sensitivity analysis in conjunction with estimands and their estimation. This thesis' view is that it is the model assumptions and probability limit $\beta^\star$ of the estimator $\widehat{\beta}$ for an estimand β that are crucial. For causal interpretation of a marginal model-based treatment effect the model should be consistent with the observed data. Models naturally just approximate reality and we prefer, as in some of the numerical illustrations, to consider $\beta^\star$ as the *de-facto* estimand; it can be interpreted as the parameter value in the best approximating or least false (Grambauer et al., 2010) model. In this framework, the term *robust* could be applied to a family of marginal models and an estimand whose value $\beta^\star$ does not change much under certain changes in the intensities for the disease or follow-up process. Numerical studies as provided in this chapter can be viewed as sensitivity analyses concerning robustness. In this sense, the cloglog transformation model for the cumulative incidence of a non-fatal event in competition with a fatal event is quite robust across a range of plausible scenarios, whereas the log transformed model is less robust. Another aspect that is sometimes considered as a form of robustness, but is slightly different, is whether a model is collapsible with respect to the effect of treatment when other covariates V affecting treatment are removed (e.g. Aalen et al., 2015). This means that the treatment effect has the same interpretation in models with or without the inclusion of V. Aside from certain types of linear models, most models are not collapsible. Proportional hazards and other relative risk models are in particular generally non-collapsible.
We note in the same vein that causal interpretation of a marginal treatment effect does not rely on the "correctness" of the assumed model. It requires only that there is a well-defined limiting estimand $\beta^\star$ which is marginal in the sense that it is based on a marginal feature that involves no conditioning on events after randomization to treatment. An understanding of the treatment effect does however depend on model adequacy, as well as its source in the effects of treatment on process intensities.

Chapter 3

Cumulative incidence function regression in clinical trials: results on interpretability of estimands and robustness of tests

3.1 Introduction

Marginal models based on transformations of cumulative incidence functions (CIFs) are appealing when estimating and testing for treatment effects in type (II) clinical trial settings, where treatments are primarily designed to target a particular non-fatal event of interest but death is a terminal event. In palliative trials of patients with metastatic cancer (c.f. Section 1.3.3, Hortobagyi et al., 1996), for instance, primary interest lies in the effect of bisphosphonates on the prevention of skeletal complications, but death may occur first.

Generalized linear models (GLMs) for CIFs have seen extensive development over the last few decades (Gray, 1988), with the most popular framework considered by Fine and Gray (1999). The development of methods for valid inference has particularly led to their widespread use. Fine and Gray (1999) pioneered methods for estimation based on a complementary log-log (cloglog) transformation of the CIF by modeling the associated subdistribution hazard function through a weighted Cox-type approach. Inverse probability of censoring weighting is applied to the

partial likelihood contributions from individuals with known at-risk indicators for the non-fatal event. Extensions of the Fine-Gray (FG) estimation procedure have been considered by He et al. (2016) and Scheike et al. (2022). Originally developed under the assumption of covariate-independent censoring, He et al. (2016) adapted the partial pseudo likelihood approach under the FG model to allow the censoring distribution to be dependent on covariates. Scheike et al. (2022) investigated whether an augmented version of the FG estimator may increase statistical efficiency. Eriksson et al. (2015) proposed a subdistribution hazard-based estimation procedure for the logit transformed CIF model, for which interpretation of treatment effects, or more generally covariate effects, may however be less appealing (Gerds et al., 2012). Alternative estimation procedures based on pseudo-values and weighted binomial estimating functions were developed by Klein and Andersen (2005) and Scheike and Zhang (2007), both of which can be used for arbitrary GLMs. Section 11 of Klein et al. (2014) gives a thorough treatment of the latter procedure, typically known as direct binomial (DB) regression; see also Scheike et al. (2008).

Despite the extensive literature on models directed at CIFs, a number of issues concerning the interpretability of treatment effects under model misspecification and robustness of tests have not been adequately addressed. We study the impact of particular kinds of model misspecification and focus on the cloglog transformed CIF model in semi-competing risks settings; see Figure 1.1(b). We consider the relationship between the CIF for the non-fatal event and the process intensity functions in the spirit of Putter et al. (2020), but we do so using large sample theory; see also Latouche et al. (2007) and Grambauer et al. (2010). When the true data-generating mechanism is an illness-death process with proportional intensities for treatment and control groups, we derive the limiting values of two common estimators of treatment effect in the cloglog transformed CIF model: one based on Fine and Gray (1999)'s proposed estimation procedure and the other based on DB regression (Scheike et al., 2008). We investigate the dependence of the limiting values of estimators on features of the underlying illness-death process, and find that the estimand arising from DB regression can be less sensitive than the estimand from the FG approach to variation in the true process. Such investigations give insights into the determinants of estimands which, along with supportive secondary analyses, can help in the interpretation of study results.

Tests for treatment effects are another essential component of randomized clinical trials. A standard approach to testing of no covariate effects on CIFs is the nonparametric log-rank-type test by Gray (1988). Poythress et al. (2020) provide a recent discussion of Gray (1988)'s test

in the context of randomized trials. We argue however that tests in a trial's primary analysis should be related to corresponding estimands; here under the true intensity-based data generation process the implicit estimand is not β but the limiting value $\beta^\star$ of the FG or DB treatment effect estimator as the sample size becomes arbitrarily large; see also Definition 1 of Section 2.3.2. Using asymptotic theory of misspecified models (White, 1982; Struthers and Kalbfleisch, 1986), we examine whether FG- and DB-based Wald tests of no treatment effect are asymptotically valid under different hypotheses concerning the true illness-death process. We show that the tests control the type I error under two key hypotheses, provided they employ robust variance estimates.

The remainder of this chapter is organized as follows. In Section 3.2, we introduce notation for the illness-death process and define generalized linear models for CIFs, following which we give estimating functions for the effect of a binary covariate for the FG approach and DB regression in Section 3.3.1. In Section 3.3.2, we provide a thorough numerical investigation of the limiting values of the estimators arising from FG and DB estimation as a function of the parameters of an intensity-based data-generating process. Section 3.3.3 is concerned with model checking and goodness-of-fit tests. In Section 3.4, we consider tests for treatment effects and carefully outline several explicit hypotheses that may be considered in regulatory settings. Theoretical and numerical results concerning the effects of model misspecification on properties of FG- and DB-based Wald tests are provided in Sections 3.4.1 and 3.4.2. Section 3.4.1 concludes with a robustness property. In Section 3.5, we present an illustrative analysis of a trial of the effect of a bisphosphonate on fracture risk in breast cancer patients with skeletal metastases. Concluding remarks are given in Section 3.6. Some derivations and additional numerical results are provided in Appendix B.

3.2 Model formulation and notation

Suppose the disease process under investigation can be represented by an illness-death model with state space $\mathcal{S} = \{0, 1, 2\}$, as illustrated in Figure 1.1(b). In oncology trials, for example, state 0 may represent the state occupied at the time of randomized treatment assignment, with state 1 entered upon occurrence of a non-fatal event (e.g. progression, relapse), and state 2 entered upon death. Recall that X takes the value 1 for individuals randomized to the experimental treatment

arm and 0 otherwise, and that $Z(t)$ denotes the state occupied at time t. We assume $Z(0) = 0$, and let $\{Z(t), 0 \leq t\}$ be the associated illness-death process. We let $T_k = \inf\{t > 0 : Z(t) = k\}$ be the entry time to state $k = 1, 2$, and $T = \inf\{t > 0 : Z(t) \neq 0\}$ the event-free survival (EFS) time (or waiting time in state 0), as before in Chapter 2.

Consider a type (II) trial setting where the experimental treatment is designed to prevent entry to state 1 but death is a terminal event. In such settings there has been increased interest in marginal models based on cumulative incidence functions (CIFs) for the non-fatal event, given by

$$F_1(t|X) = P(T_1 \leq t|X) = \int_0^t S(u^-|X)\lambda_{01}(u|X)du \ , \tag{3.2.1}$$

where

$$S(u|X) = P(T > u|X) = \exp\bigg(-\int_0^u [\lambda_{01}(s|X) + \lambda_{02}(s|X)]ds\bigg) \tag{3.2.2}$$

is the probability of being alive and event-free (in state 0) at time u. Since a model for $F_1(t|X)$ does not consider death following the non-fatal event of interest, it is sufficient to consider the competing risks process in Figure 1.1(a) with state space $\mathcal{S} = \{0, 1, 2\}$ and intensities $\lambda_{0k}(t|X)$, defined as in (2.3.1) for $k = 1, 2$.

When models are specified for the intensity functions $\lambda_{01}(t|X)$ and $\lambda_{02}(t|X)$ (Prentice et al., 1978), (3.2.1) depends on the process intensities in complex ways and there is no resulting one-dimensional summary of the treatment effect. As discussed in previous sections, generalized linear models (GLMs) of the form (1.2.8), i.e. $g(F_1(t|X)) = \alpha(t) + \beta X$, have been proposed to address this by modeling treatment effects directly on the cumulative incidence function. While the most suitable link function $g(\cdot)$ will depend on the setting, we observed that the cloglog model often provides a reasonable approximation in a broad range of applications; we restrict attention in this chapter to semiparametric GLMs of the form

$$\log\big(-\log(1 - F_1(t|X))\big) = \alpha(t) + \beta X \ , \tag{3.2.3}$$

with the baseline function $\alpha(t) = \log(-\log(1 - F_1(t|X = 0)))$ left unspecified and estimand

$$\beta = \log\bigg(\frac{\log(1 - F_1(t|X = 1))}{\log(1 - F_1(t|X = 0))}\bigg) \ .$$

3.3 Cumulative incidence function regression under misspecification

In Section 3.3.1, we consider estimation for (3.2.3) as described by Fine and Gray (1999) (FG) and based on direct binomial (DB) regression (Scheike et al., 2008). Through large sample theory we derive the limiting values of the associated treatment effect estimators when (3.2.3) is misspecified. We consider in particular the case where the true process is a competing risks model with multiplicative treatment effects on the intensities $\lambda_{0k}(t|X)$, and study the probability limits $\beta^\star$ of the FG and DB estimators of β in (3.2.3) under such a true process. These limits may also be sensitive to misspecification of the censoring model, so we comment on this first.
In what follows we let $Z(T) \in \{1, 2\}$ record the event type, and refer to $k = 1$ as the event of interest and $k = 2$ as the competing event (e.g. event-free death).

3.3.1 Estimation methods and large sample results

We let τ be the planned duration of follow-up for each individual $i = 1, \ldots, n$, i.e. the administrative censoring time. Individuals may be prematurely lost to follow-up at a random censoring time $C_{ri} < \tau$, giving the net censoring time $C_i = \min(C_{ri}, \tau)$. The results here readily extend to settings where individuals' administrative censoring times may differ; for simplicity we consider the case where they are equal, i.e. $\tau_i = \tau$. We assume that a set of covariates W_i is available such that $\{Z_i(u), u > 0\} \perp\!\!\!\perp C_{ri}|X_i, W_i$ and let $G_i(u) = G(u|X_i, W_i) = P(C_{ri} > u|X_i, W_i)$. If the censoring process does not depend on covariates associated with the failure process, the Kaplan-Meier (KM) estimate of $G_i(u) = G(u) = P(C_r > u)$ can be used for estimation of the censoring distribution. In settings where censoring depends on covariates, intensity-based regression models (e.g. Cox proportional hazards model or Aalen (1980)'s additive model) are common for the censoring time C_r. If censoring depends only on a limited number of discrete covariates, an alternative to regression modeling is to define strata and use KM estimation within strata. We review estimating functions and large sample theory for KM estimation and Cox regression in Sections B.1.1 and B.1.2 of Appendix B.1; in particular we review the limiting behavior of the KM and Cox estimators of the censoring survivor function when the censoring model for C_r may be misspecified. In investigations that follow we label the probability limit of the estimator

as $G_i^\star(u)$ (i.e. $\widehat{G} \longrightarrow G^\star$), with $G_i^\star(u)$ representing $G^\star(u)$ in the case of an unstratified KM estimator $\widehat{G}(u)$ or $G^\star(u|X_i, W_i)$ if a stratified KM or Cox estimator $\widehat{G}(u|X_i, W_i)$ is considered. We note that $G^\star = G$ when the "working" censoring model is correct (i.e. $\widehat{G} \longrightarrow G$). A special case is when no additional covariates W_i are needed; we consider this case here and assume that $\{Z_i(u), u > 0\} \perp\!\!\!\perp C_{ri} \mid X_i$ for the remainder of Chapter 3.

3.3.1.1 Fine-Gray method of estimation

The Fine and Gray (1999) method is based on the fact that (3.2.3) may be characterized as a proportional hazards model for T_1 with baseline cumulative subdistribution hazard function $\Gamma(t) = \exp(\alpha(t))$. Then if

$$Y_i^\dagger(t) = \mathbb{1}(T_{1i} \geq t) = \mathbb{1}(T_i \geq t) + \mathbb{1}(T_{2i} \leq t) = \mathbb{1}(Z_i(t^-) \in \{0, 2\})$$

is the so-called Fine-Gray (FG) at-risk indicator, a weighted Cox partial likelihood leads to estimating functions

$$\sum_{i=1}^{n} w_i(t) Y_i^\dagger(t) \big(dN_{1i}(t) - \exp(\beta X_i) d\Gamma(t)\big) = 0 \,, \tag{3.3.1}$$

$$\sum_{i=1}^{n} \int_0^\infty w_i(t) Y_i^\dagger(t) \big(dN_{1i}(t) - \exp(\beta X_i) d\Gamma(t)\big) X_i = 0 \,, \tag{3.3.2}$$

for $d\Gamma(t)$ and β respectively, where $N_{1i}(t) = \mathbb{1}(T_{1i} \leq t)$, $\Delta N_{1i}(t) = N_{1i}(t + \Delta t^-) - N_{1i}(t^-)$ and $dN_{1i}(t) = \lim_{\Delta t \downarrow 0} \Delta N_{1i}(t)$. Note that individuals who have previously experienced the competing event (e.g. event-free death) and are thus in state 2 at time t^- are not truly at-risk of the event of interest, but the FG at-risk indicator treats them as such (Putter et al., 2020). Inverse probability of censoring weights in (3.3.1) and (3.3.2), defined as

$$w_i(t) = \frac{\mathbb{1}(C_i \geq \min(T_i, t))}{G_i(\min(T_i, t))} = \mathbb{1}(t \leq \tau) \cdot \frac{\mathbb{1}(C_{ri} \geq \min(T_i, t))}{G_i(\min(T_i, t))} \,, \tag{3.3.3}$$

are necessary to deal with the fact that for individuals censored before $\min(T_i, t)$ we do not know the value of $N_{1i}(t)$; see Section 4.1.2 of Cook and Lawless (2018). Solving (3.3.1) for $d\Gamma(t)$ with fixed β gives the profile estimate

$$d\widetilde{\Gamma}(t) = \sum_{i=1}^{n} w_i(t) Y_i^\dagger(t) dN_{1i}(t) \Big/ \sum_{i=1}^{n} w_i(t) Y_i^\dagger(t) \exp(\beta X_i) \,, \tag{3.3.4}$$

and substituting (3.3.4) into (3.3.2) gives the partial pseudo-score function for β as

$$U^{FG}(\beta, G) = \sum_{i=1}^{n} \int_0^{\infty} w_i(t) Y_i^{\dagger}(t) \left[X_i - \frac{S^{(1)}(t,\beta)}{S^{(0)}(t,\beta)} \right] dN_{1i}(t) \ , \tag{3.3.5}$$

where $S^{(l)}(t,\beta) = \sum_{i=1}^{n} w_i(t) Y_i^{\dagger}(t) X_i^l \exp(\beta X_i), l = 0, 1$. Setting (3.3.5) equal to zero and solving gives the FG estimator $\widehat{\beta}_{FG}$ following estimation of G. The Breslow-type estimate of $\Gamma(t) = \int_0^t d\Gamma(u)$ is $\widehat{\Gamma}_{FG}(t) = \int_0^t d\widehat{\Gamma}_{FG}(u)$ where $d\widehat{\Gamma}_{FG}(u)$ can be obtained by substituting $\widehat{\beta}_{FG}$ and $\widehat{w}_i(t)$ into (3.3.4):

$$d\widehat{\Gamma}_{FG}(u) = \sum_{i=1}^{n} \widehat{w}_i(u) Y_i^{\dagger}(u) dN_{1i}(u) / \sum_{i=1}^{n} \widehat{w}_i(u) Y_i^{\dagger}(u) \exp(\widehat{\beta}_{FG} X_i) \ . \tag{3.3.6}$$

Three assumptions labelled A1-A3 are used in developing asymptotic results which follow. If

A1 $\{Z(u), u > 0\} \perp\!\!\!\perp C_r \mid X$,

A2 model (3.2.3) for $F_1(t|X)$ is valid , and

A3 the censoring model for C_r is correctly specified such that $G^{\star} = G$ and $\widehat{G} \longrightarrow G$,

then $\widehat{\beta}_{FG}$ is consistent for β and $\widehat{\Gamma}_{FG}(t)$ for $\Gamma(t)$ (Fine and Gray, 1999; He et al., 2016). More generally, if $G_i^{\star}(\cdot)$ is defined as above and

$$w_i^{\star}(t) = \mathbb{1}(t \leq \tau) \cdot \frac{\mathbb{1}(C_{ri} \geq \min(T_i, t))}{G_i^{\star}(\min(T_i, t))} \ ,$$

then we rewrite (3.3.5) as

$$U^{FG}(\beta, G^{\star}) = \sum_{i=1}^{n} \int_0^{\infty} w_i^{\star}(t) Y_i^{\dagger}(t) \left[X_i - \frac{S^{(1,\star)}(t,\beta)}{S^{(0,\star)}(t,\beta)} \right] dN_{1i}(t) \ , \tag{3.3.7}$$

where $S^{(l,\star)}(t,\beta) = \sum_{i=1}^{n} w_i^{\star}(t) Y_i^{\dagger}(t) X_i^l \exp(\beta X_i), l = 0, 1$. The implied estimand is the probability limit $\beta^{\star}$ of the estimator $\widehat{\beta}$ based on (3.3.7), which solves

$$\mathbb{E}\big(U^{FG}(\beta, G^{\star})\big) = \int_0^{\infty} \left\{ s^{(1,\star)}(t) - \frac{s^{(1,\star)}(t,\beta)}{s^{(0,\star)}(t,\beta)} s^{(0,\star)}(t) \right\} dt = 0 \ , \tag{3.3.8}$$

where $s^{(l,\star)}(t,\beta) = \mathbb{E}(S^{(l,\star)}(t,\beta))$ and $s^{(l,\star)}(t) = \mathbb{E}\big(w_i^{\star}(t) Y_i^{\dagger}(t) X_i^l dN_{1i}(t)\big), l = 0, 1$, and where expectations are taken with respect to the true competing risks, censoring and covariate processes

(White, 1982; Struthers and Kalbfleisch, 1986). The conceptual estimand β depends on an assumed model, and if that model is not a correct representation of the true cumulative incidence functions, then its interpretability is unclear. By examining the probability limit $\beta^\star$ of the estimator (and of the baseline function $\alpha^\star(t)$) under different assumptions about the true competing risks process, we are able to assess the adequacy of the CIF model and treatment effect estimates based on it. A recent discussion of this topic in the linear regression context is given by Buja et al. (2019).

If assumptions A1 and A3 hold but A2 does not, the expectations needed to compute (3.3.8) are

$$\begin{aligned}
s^{(0,\star)}(t,\beta) &= s^{(0)}(t,\beta) = \mathbb{E}_X\big((1-F_1(t|X))\exp(\beta X)\big) \\
&= \sum_{x=0,1} P(X=x)\exp(\beta x)(1-F_1(t|X=x)) \ , \\
s^{(1,\star)}(t,\beta) &= s^{(1)}(t,\beta) = \mathbb{E}_X\big((1-F_1(t|X))X\exp(\beta X)\big) \\
&= P(X=1)\exp(\beta)(1-F_1(t|X=1)) \ , \\
s^{(0,\star)}(t) = s^{(0)}(t) &= \mathbb{E}_X\big(\lambda_{01}(t|X)S(t|X)\big) = \sum_{x=0,1} P(X=x)\lambda_{01}(t|X=x)S(t|X=x) \ , \\
s^{(1,\star)}(t) = s^{(1)}(t) &= \mathbb{E}_X\big(\lambda_{01}(t|X)S(t|X)X\big) = P(X=1)\lambda_{01}(t|X=1)S(t|X=1) \ ,
\end{aligned}$$

where $s^{(l)}(t,\beta) = \mathbb{E}(S^{(l)}(t,\beta))$ and $s^{(l)}(t) = \mathbb{E}\big(\sum_{i=1}^n w_i(t)Y_i^\dagger(t)X_i^l dN_{1i}(t)\big), l = 0,1$. We denote the probability limit of the solution to (3.3.8) by $\beta^\star_{FG}$. Since $\mathbb{E}_{C_r|(T,Z(T)),X}(w_i^\star(t)) = 1$ under A1 and A3, $\beta^\star_{FG}$ is not dependent on the censoring process. The probability limit of $d\widetilde{\Gamma}(t)$ is the solution when the expected value of the LHS of (3.3.1) is set equal to zero with fixed $\beta^\star_{FG}$, and can be found to be

$$d\Gamma^\star_{FG}(t) = \frac{\mathbb{E}_X(\lambda_{01}(t|X)S(t|X))}{\mathbb{E}_X(\exp(\beta^\star_{FG}X)(1-F_1(t|X)))} = \frac{\sum_{x=0,1}P(X=x)\lambda_{01}(t|X=x)S(t|X=x)}{\sum_{x=0,1}P(X=x)\exp(\beta^\star_{FG}x)(1-F_1(t|X=x))} \ ,$$

where $F_1(t|X)$ is given by (3.2.1). Consequently, the limit of the FG estimator of $F_1(t|X)$ under (3.2.3) is

$$F^\star_{1,FG}(t|X) = 1-\exp\big(-\exp(\alpha^\star_{FG}(t))\exp(\beta^\star_{FG}X)\big) \ , \tag{3.3.9}$$

with $\exp(\alpha^\star_{FG}(t)) = \Gamma^\star_{FG}(t) = \int_0^t d\Gamma^\star_{FG}(u)$.

3.3.1.2 Estimating function for direct binomial regression

The binomial estimation procedure uses the fact that, at any given time t, $N_{1i}(t)$ follows a binomial distribution with mean $\mathbb{E}(N_{1i}(t)|X_i) = P(T_{1i} \leq t|X_i) = F_1(t|X_i)$ (Scheike et al., 2008). We consider estimation of $F_1(\cdot|X_i)$ based on (3.2.3) at R distinct time points $s_1, s_2, \ldots, s_R$ over $(0, \tau)$. In this case the cumulative incidence function $F_1(t|X_i;\theta)$ under (3.2.3) is specified at times $s_1, \ldots, s_R$ by an $(R+1) \times 1$ parameter vector $\theta = (\alpha', \beta)'$, where $\alpha = (\alpha_1, \ldots, \alpha_R)'$ with $\alpha_r := \alpha(s_r)$, $r = 1, 2, \ldots, R$. Estimation of θ is based on the indicators $N_{1i}(s_r)$ which are ideally available for all $r = 1, 2, \ldots, R$ and all $i = 1, 2, \ldots, n$. Note that the binary indicators $N_{11}(s_1), \ldots, N_{11}(s_R)$ from a given individual are not independent, but in the absence of random censoring one could use a composite binomial log-likelihood under a working independence assumption of the form

$$\sum_{i=1}^{n}\sum_{r=1}^{R}\big(N_{1i}(s_r)\log(h(\alpha_r + \beta X_i)) + (1 - N_{1i}(s_r))\log(1 - h(\alpha_r + \beta X_i))\big)\ , \tag{3.3.10}$$

where $h(u) = g^{-1}(u) = 1 - \exp(-\exp(u))$ is the inverse of the cloglog link function. Differentiating (3.3.10) with respect to θ gives unbiased estimating functions for α and β. In trials with random loss to follow-up $N_{1i}(s_r)$ is known if $\min(T_i, C_i) \geq s_r$ or $T_i < \min(s_r, C_i)$, but unknown if $C_i < \min(T_i, s_r)$. The weighted response

$$\widetilde{N}_{1i}(s_r) = \frac{\mathbb{1}(C_i \geq \min(T_i, s_r))N_{1i}(s_r)}{G_i(\min(T_i, s_r))} \underset{s_r \leq \tau}{=} \frac{\mathbb{1}(C_{ri} \geq \min(T_i, s_r))N_{1i}(s_r)}{G_i(\min(T_i, s_r))} \tag{3.3.11}$$

is however observable when $G_i(\cdot)$ is known or estimated, and has the same expectation as $N_{1i}(s_r)$ if assumptions A1 and A3 are satisfied. Replacing $N_{1i}(s_r)$ by $\widetilde{N}_{1i}(s_r)$ in (3.3.10) and taking the derivatives with respect to θ gives the DB estimating equations as

$$U^{DB}(\theta, G) = \sum_{i=1}^{n}\begin{pmatrix} U_i(\alpha, G) \\ U_i(\beta, G)\end{pmatrix} = \sum_{i=1}^{n}\begin{pmatrix} U_i(\alpha_1, G) \\ \ldots \\ U_i(\alpha_R, G) \\ U_i(\beta, G)\end{pmatrix} = \sum_{i=1}^{n} D_i A_i^{-1}(\widetilde{N}_{1i} - \mu_i)\ , \tag{3.3.12}$$

where $A_i = \text{diag}\big(\{F_1(s_r|X_i;\theta)(1 - F_1(s_r|X_i;\theta))\}, r = 1, \ldots, R\big)$ is an $R \times R$ diagonal (working independence) covariance matrix, $\widetilde{N}_{1i} = (\widetilde{N}_{1i}(s_1), \ldots, \widetilde{N}_{1i}(s_R))'$ is an $R \times 1$ vector of weighted

responses, $\mu_i = (\mu_i(s_1), \ldots, \mu_i(s_R))'$ is an $R \times 1$ vector with entries $\mu_i(s_r) = \mathbb{E}(N_{1i}(s_r)|X_i;\theta) = h(\alpha_r + \beta X_i)$, and $D_i = (\partial\mu_i/\partial\alpha', \partial\mu_i/\partial\beta)'$ is a $(R+1) \times R$ matrix of derivatives. Following specification of the grid $s_1, \ldots, s_R$ and estimation of G, the DB estimator $\widehat{\theta} = (\widehat{\alpha}'_{DB}, \widehat{\beta}_{DB})'$ solves $U^{DB}(\theta, G) = (0, 0, \ldots, 0)'$. It is possible to base estimation on all observed type 1 event times as in the FG approach, but we follow Klein and Andersen (2005) who recommend 5-10 time points equally-spaced over $(0, \tau)$. If assumptions A1-A3 are valid, then $\widehat{\theta}$ is consistent for θ (Scheike et al., 2008). More generally $\widehat{\theta}$ is consistent for $\theta^\star = ((\alpha^\star_{DB})', \beta^\star_{DB})$, the solution to $\mathbb{E}(U^{DB}(\theta, G^\star)) = (0, 0, \ldots, 0)'$ where

$$U^{DB}(\theta, G^\star) = \sum_{i=1}^{n} \begin{pmatrix} U_i(\alpha, G^\star) \\ U_i(\beta, G^\star) \end{pmatrix} = \sum_{i=1}^{n} \begin{pmatrix} U_i(\alpha_1, G^\star) \\ \ldots \\ U_i(\alpha_R, G^\star) \\ U_i(\beta, G^\star) \end{pmatrix} = \sum_{i=1}^{n} D_i A_i^{-1} (\widetilde{N}^\star_{1i} - \mu_i) \tag{3.3.13}$$

for $\widetilde{N}^\star_{1i} = (\widetilde{N}^\star_{1i}(s_1), \ldots, \widetilde{N}^\star_{1i}(s_R))'$ and

$$\widetilde{N}^\star_{1i}(s_r) = \frac{\mathbb{1}(C_{ri} \geq \min(T_i, s_r)) N_{1i}(s_r)}{G^\star_i(\min(T_i, s_r))} ,$$

and where expectations are taken with respect to the true competing risks, censoring and covariate processes. The expectations needed to derive $\theta^\star$ under assumptions A1 and A3 are

$$\begin{aligned} \mathbb{E}(U_1(\alpha_r, G^\star)) &= \mathbb{E}(U_1(\alpha_r, G)) \qquad (3.3.14) \\ &= \sum_{x=0,1} \frac{P(X=x) h'(\alpha_r + \beta x)}{h(\alpha_r + \beta x)(1 - h(\alpha_r + \beta x))} (F_1(s_r|X=x) - h(\alpha_r + \beta x)) \\ \mathbb{E}(U_1(\beta, G^\star)) &= \mathbb{E}(U_1(\beta, G)) \qquad (3.3.15) \\ &= \sum_{r=1}^{R} \frac{P(X=1) h'(\alpha_r + \beta)}{h(\alpha_r + \beta)(1 - h(\alpha_r + \beta))} (F_1(s_r|X=1) - h(\alpha_r + \beta)) , \end{aligned}$$

where $F_1(s_r|X)$ is given as in (3.2.1). The probability limit of the DB estimator of $F_1(s_r|X)$ at time s_r is

$$F^\star_{1,DB}(s_r|X) = h(\alpha^\star_{rDB} + \beta^\star_{DB} X) , \qquad r = 1, 2, \ldots, R. \tag{3.3.16}$$

An alternative to weighting the state 1 occupancy indicators is to weight the binomial estimating equations (EEs) as follows:

$$\sum_{i=1}^{n} W_i D_i A_i^{-1}(N_{1i} - \mu_i) = 0 \ ,$$

where W_i is a $R \times R$ diagonal matrix with entries $w_i(s_r)$ given by (3.3.3). Note that the limiting values are the same for both strategies when the censoring model is correctly specified. Unlike the weighted outcome approach, which uses information from all individuals at time s_r, individuals who were censored prior to $\min(T, s_r)$ do not contribute to the pseudo-score function at s_r in the weighted EE approach. For the case of the logit model (1.2.8) and a single time point s_1 ($R = 1$), Blanche et al. (2023) showed that the relative efficiency of these two strategies depends on the distribution of the censoring times. In investigations that follow we focus on the original DB method as described by Scheike et al. (2008).

3.3.2 Limiting values of FG and DB estimators under intensity-based processes

We examined the limiting values of the FG and DB estimators using the large sample results of Section 3.3.1 in a setting where the true underlying process has $0 - k$ intensities of the form $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X), k = 1, 2$. We present corresponding results for time-inhomogeneous intensities in Section B.2.2. Without loss of generality we set the administrative censoring time to $\tau = 1$ and determined λ_k so that $P(T \leq 1|X = 0) = 0.6$ where $T = \min(T_1, T_2)$ and $P(T_1 < T_2|T \leq 1, X = 0) \in (0, 1)$. We set $\exp(\gamma_1)$ to $1, 0.75$ or 0.6 to correspond to a $0\%, 25\%$ or 40% reduction in the $0 - 1$ intensity with treatment; values for $\exp(\gamma_2)$ varied from 0.6 to 1.3, giving either a decrease, no change or an increase in the $0 - 2$ intensity. For each parameter setting the limiting values $\beta^\star_{FG}$ and $\beta^\star_{DB}$ were computed using (3.3.8) and (3.3.14)-(3.3.15), respectively.

Figure 3.1 contains contour plots depicting the dependence of $\exp(\beta^\star_{FG})$ (left panels) and $\exp(\beta^\star_{DB})$ (right panels) on $\exp(\gamma_2), P(T_1 < T_2|T \leq 1, X = 0)$ and $\exp(\gamma_1)$ for the case where the adopted censoring model is correct (i.e. $G^\star = G, \widehat{G} \longrightarrow G$) and $R = 6$ for the DB approach. The limiting values $\beta^\star_{FG}$ and $\beta^\star_{DB}$ are not expected to be identical as they solve

different estimating functions. If treatment does not affect the $0-1$ intensity or the $0-2$ intensity ($\exp(\gamma_1)=\exp(\gamma_2)=1$), then $\exp(\beta^\star_{FG})=\exp(\beta^\star_{DB})=1$; see vertical contour in upper panels. In the middle and lower panels treatment affects the $0-1$ intensity (i.e. $\exp(\gamma_1)\neq 1$); if $\exp(\gamma_2)=1$, then $\exp(\beta^\star_{FG})$ and $\exp(\beta^\star_{DB})$ are very close to $\exp(\gamma_1)$. It can be seen from all panels of Figure 3.1 that $\exp(\beta^\star_{FG})>\exp(\gamma_1)$ and $\exp(\beta^\star_{DB})>\exp(\gamma_1)$ if $\exp(\gamma_2)<1$, and $\exp(\beta^\star_{FG})<\exp(\gamma_1)$ and $\exp(\beta^\star_{DB})<\exp(\gamma_1)$ if $\exp(\gamma_2)>1$. The magnitude of the limiting value further depends on the probability observed events are of the cause of interest in the control arm. Naturally the greater the probability of the competing event (e.g. event-free death) the greater the impact of $\exp(\gamma_2)$ on the estimand of interest.

We conclude that model (3.2.3) can be useful for the evaluation of experimental interventions in type (II) trials aiming to prevent type 1 events when the competing (type 2) event rate is low and treatment has a fairly small effect on the corresponding $0-2$ intensity; in this case $\beta^\star$ represents the effect of treatment on $F_1(t|X)$ well, and is close to the treatment effect γ_1 on the main event intensity. More generally, $\beta^\star$ is a function of the true intensity-based process and the degree of model misspecification. The estimand $\beta^\star_{DB}$ appears less sensitive than $\beta^\star_{FG}$ to misspecification of (3.2.3), as $\beta^\star_{DB}$ is closer to γ_1 than $\beta^\star_{FG}$ in all scenarios considered; see also Figures B.1 and B.2 in Appendix B.2.1 for additional illustration. The difference between $\exp(\beta^\star_{FG})$ and $\exp(\beta^\star_{DB})$ decreases with increasing probability observed events are of type 1 in the control group.
In Appendix B.2.1 we further discuss the impact of model misspecification on the estimation of $F_1(t|X)$ by comparing the true $F_1(t|X)$ (as given in (3.2.1)) to $F^\star_{1,\cdot}(t|X)$ (as given in (3.3.9) and (3.3.16)) for each $X=0,1$. It is readily apparent from Figures B.3 and B.4 that treatment effects on the $0-2$ intensity (i.e. $\exp(\gamma_2)\neq 1$) impact the CIF-based estimands through a reduction ($\exp(\gamma_2)>1$) or increase ($\exp(\gamma_2)<1$) in the cumulative probability of the event of interest for the experimental treatment.
We next provide guidance on methods for checking the assumption of a constant treatment effect in (3.2.3). We note that the cloglog form in (3.2.3) may also be inadequate in some cases. This can be examined by plotting nonparametric estimates of $F_1(t|X)$ for $X=0,1$ or by fitting a parametric family of transformation models. We illustrate the former approach in Section 3.5.

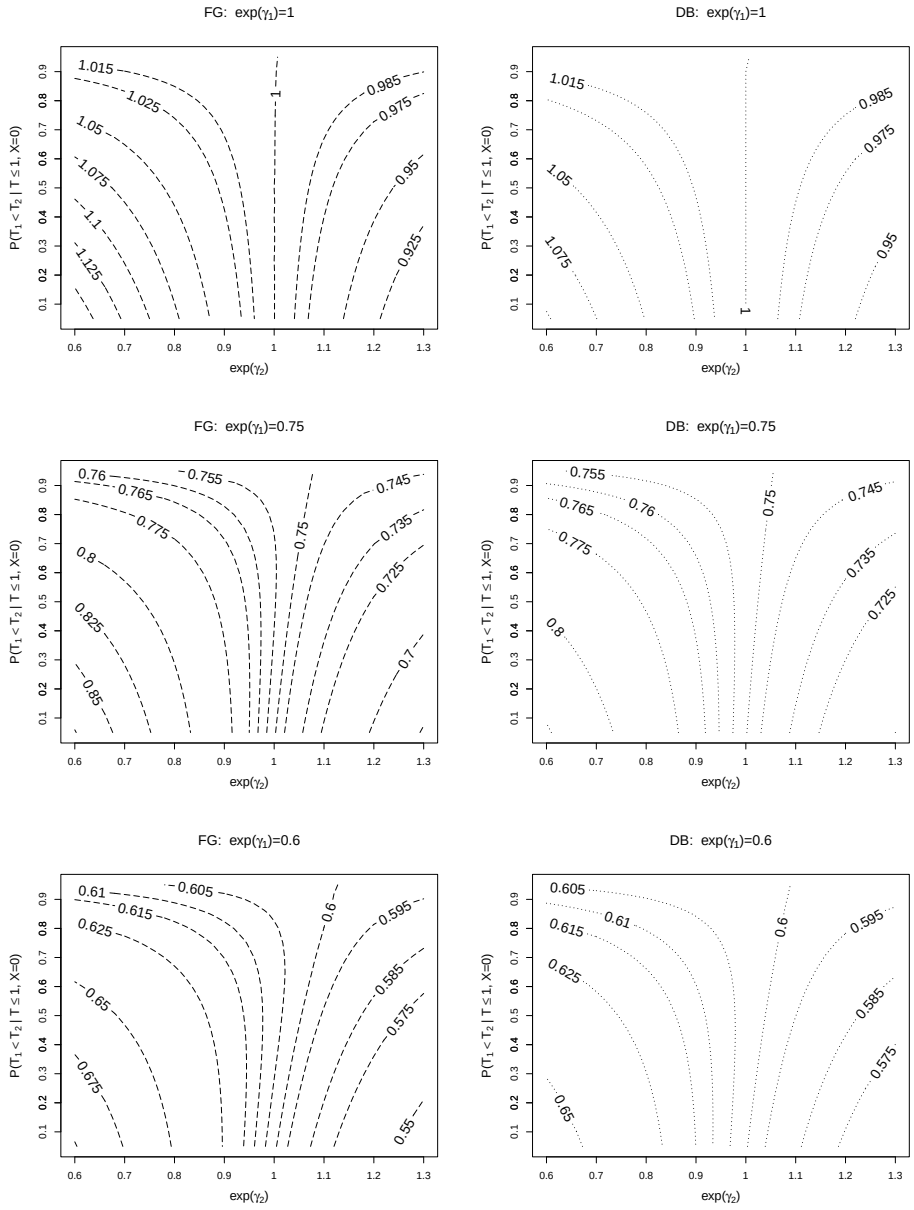


Figure 3.1: Contours showing the limiting values $\exp(\beta^{\star}_{FG})$ (left column) and $\exp(\beta^{\star}_{DB})$ (right column) as a function of $\exp(\gamma_2)$ and $P(T_1 < T_2 | T \leq 1, X = 0)$ when the true process has intensities $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X)$ for $\exp(\gamma_1) = 1$ (top row), 0.75 (middle row) and 0.6 (bottom row); $\tau = 1, P(T \leq 1|X = 0) = 0.6$, $R = 6, G^{\star} = G$.

3.3.3 Goodness of fit

A range of methods have been suggested in the literature to assess the plausibility of a time-constant treatment effect within the FG model (Zhou et al., 2013; Li et al., 2015; Boher et al., 2017). A simple approach is to consider extended cloglog models for $F_1(t|X)$ of the form

$$\log\big(-\log(1-F_1(t|X))\big) = \alpha(t) + \beta(t)X = \alpha(t) + \big[\beta + \nu b(t)\big]X \ , \tag{3.3.17}$$

with the time-varying treatment effect defined as $\beta(t) = \beta + \nu b(t)$ for a pre-specified function $b(t)$ of time and an unknown regression coefficient ν. A Wald test of $H_0 : \nu = 0$ is thus a test of a time-constant treatment effect and can be easily executed using standard FG model software (e.g. R function crr()). Different choices of $b(t)$ are sensitive to different departures from the null hypothesis; common choices include $b(t) = t$ and $b(t) = \log(t)$. Goodness-of-fit tests based on scaled Schoenfeld residuals are common practice to assess the proportional hazards assumption in Cox regression (Grambsch and Therneau, 1994). Zhou et al. (2013) proposed a similar procedure based on modified weighted Schoenfeld residuals for FG-based CIF regression; in particular, they also assume $\beta(t)$ to be of the specific form $\beta + \nu b(t)$ and consider a score test of $H_0 : \nu = 0$ based on constrained optimization of the pseudo-score function under model (3.3.17). Simulation studies in Zhou et al. (2013) have shown that the score test tends to be more powerful than the Wald test when the assumed form of $\beta(t)$ is correctly specified, except for the case where $\beta = 0$. Alternatively, Li et al. (2015) proposed a procedure based on cumulative sums of martingale residuals to test against a general alternative of $\beta(t)$.

Less work has been done for the DB approach. Scheike and Zhang (2008) proposed nonparametric goodness-of-fit tests for $H_0 : \beta_r = \beta$ based on Kolmogorov-Smirnov-type and Cramer-von-Mises-type statistics. Analogous to (3.3.17), a Wald test of $H_0 : \nu = 0$ in the extended model

$$\log\big(-\log(1-F_1(s_r|X))\big) = \alpha_r + \beta_r X = \alpha_r + \big[\beta + \nu b(s_r)\big]X \tag{3.3.18}$$

can be considered. An alternative procedure is to carry out a Wald test of $H_0 : \beta_r = \beta$ for all $r = 1, 2, \ldots, R$, or equivalently $H_0 : C\beta^\dagger = 0$ for $\beta^\dagger = (\beta_1, \ldots, \beta_R)'$, using the test statistic

$$\big(C\widehat{\beta}^\dagger\big)'\big(C\widehat{\Sigma}^\dagger_\beta C'\big)^{-1}C\widehat{\beta}^\dagger \ \sim \ \chi^2_{R-1} \ , \tag{3.3.19}$$

and with $\widehat{\beta}^\dagger = (\widehat{\beta}_1, \ldots, \widehat{\beta}_R)'$ assumed to be approximately normal with mean $\beta^\dagger$ and covariance matrix $\Sigma^\dagger_\beta$. The so-called $(R-1) \times R$ contrast matrix C has non-zero entries $c_{jl} = 1$ if $j = l$

and $c_{jl} = -1$ if $l = j + 1, j = 1, \ldots, R-1, l = 1, \ldots, R$. We refer to Section 4.1.3 of Cook and Lawless (2018) for details on estimation of $\theta^\dagger = (\alpha_1, \ldots, \alpha_R, \beta^{\dagger\prime})'$ in model (3.3.18). Bootstrap methods as outlined in Section 11.4.2 of Klein et al. (2014) can be used to obtain $\widehat{\Sigma}^\dagger_\beta$.

3.4 Robustness of tests

Hypothesis testing is central to the evaluation of experimental treatments in randomized clinical trials, so we consider this topic with two aims. We first discuss a variety of null hypotheses that may be tested with competing risks processes, examine their relationships, and provide a short overview of existing tests. Second, using large sample theory and empirical studies we investigate the behavior of Wald tests of no treatment effect on $F_1(t|X)$ in settings where the cloglog model (3.2.3) is misspecified.

3.4.1 Formulation of testing framework and robustness property for Wald tests

Several different null hypotheses can be formulated for testing treatment effects within the context of Figure 1.1(a), including

$$H_0^{\lambda_1\lambda_2} : \lambda_{0k}(t|X=1) = \lambda_{0k}(t|X=0) \text{ for } t \geq 0 \text{ and all } k = 1, 2 \tag{3.4.1}$$

$$H_0^{\lambda_k} : \lambda_{0k}(t|X=1) = \lambda_{0k}(t|X=0) \text{ for } t \geq 0 \tag{3.4.2}$$

$$H_0 : F_1(t|X=1) = F_1(t|X=0) \text{ for } t \geq 0 \tag{3.4.3}$$

$$H_0^S : S(t|X=1) = S(t|X=0) \text{ for } t \geq 0 \tag{3.4.4}$$

$$H_0^{F_1F_2} : F_k(t|X=1) = F_k(t|X=0) \text{ for } t \geq 0 \text{ and all } k = 1, 2\ , \tag{3.4.5}$$

and with the cumulative incidence function for competing events in (3.4.5) defined as

$$F_2(t|X) = P(T_2 \leq t|X) = \int_0^t S(u^-|X)\lambda_{02}(u|X)du\ .$$

A schematic showing the relation between the null hypotheses (3.4.1)-(3.4.5) is given in Figure 3.2.

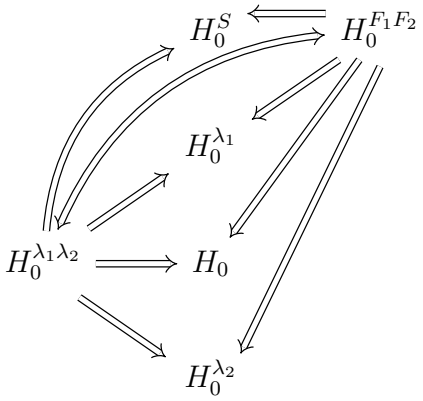


Figure 3.2: Relationships between different null hypotheses in a competing risks process.

The global null $H_0^{\lambda_1\lambda_2} = H_0^{\lambda_1} \cap H_0^{\lambda_2}$ corresponds to no treatment effect on the full disease process and, thus, no effect on any process feature. The null hypothesis $H_0^{\lambda_k}$ is weaker than $H_0^{\lambda_1\lambda_2}$ as it reflects no treatment effect only on the $0-k$ intensity, $k = 1, 2$. Neither $H_0^{\lambda_1}$ nor $H_0^{\lambda_2}$ alone imply H_0^S or H_0 due to (3.2.2) and (3.2.1). For example, if $H_0^{\lambda_1}$ is true but $H_0^{\lambda_2}$ is not, then $F_1(t|X)$ depends on X. Clearly, $H_0^{\lambda_k}$ alone does not imply $H_0^{F_1F_2}$ either. The null hypothesis H_0 is aligned with the estimator of β such that if $\beta = 0$ in (3.2.3) then $F_1(t|X) = F_1(t)$. Mathematically, however, H_0 does not imply $H_0^{\lambda_1}, H_0^{\lambda_2}$ or $H_0^{\lambda_1\lambda_2}$, but it does impose complicated constraints on the two intensities. For example, $f_1(t|X = 1) = f_1(t|X = 0)$ implies that

$$\lambda_{01}(t|X=0)e^{-(\Lambda_{01}(t|X=0)+\Lambda_{02}(t|X=0))} = \lambda_{01}(t|X=1)e^{-(\Lambda_{01}(t|X=1)+\Lambda_{02}(t|X=1))} .$$

Such a constraint would not be satisfied with most intensity-based models. Since $S(t|X) = 1 - F_1(t|X) - F_2(t|X)$ and $\lambda_{0k}(t|X) = f_k(t|X)/S(t|X)$, $H_0^{F_1F_2}$ implies $H_0^S, H_0^{\lambda_k}$ and $H_0^{\lambda_1\lambda_2}$. If H_0 is true but $H_0^{F_2} : F_2(t|X = 1) = F_2(t|X = 0)$ is not, then $S(t|X)$ depends on X. Finally, H_0^S does not imply any of the other null hypotheses in (3.4.1)-(3.4.3) and (3.4.5).

A key question concerning tests of hypotheses (3.4.1)-(3.4.5) is: if a test has nominal size ω, what is the actual (empirical) size $\omega^\star$ under a different null hypothesis? For instance, if a test of H_0 is based on a marginal model such as (3.2.3), what is the rejection rate if the true process satisfies $H_0^{\lambda_1}$? We discuss specific tests in detail shortly, but first state the following theorem pertaining to Figure 3.2.

Theorem 1 *Consider tests $T^{\lambda_1\lambda_2}, T^{\lambda_k}, T^{F_1}, T^S$ and $T^{F_1F_2}$ of nominal size ω for the null hypotheses (3.4.1)-(3.4.5), respectively; in what follows, $\neq \omega$ means not necessarily equal to ω.*

1. *Under $H_0^{\lambda_1\lambda_2}$, $P(T^{\lambda_k}$ rejects $H_0^{\lambda_k}) = \omega, P(T^{F_1}$ rejects $H_0) = \omega, P(T^S$ rejects $H_0^S) = \omega$ and $P(T^{F_1F_2}$ rejects $H_0^{F_1F_2}) = \omega$.*

2. *Under $H_0^{\lambda_k}$, $P(T^{\lambda_1\lambda_2}$ rejects $H_0^{\lambda_1\lambda_2}) \neq \omega, P(T^{F_1}$ rejects $H_0) \neq \omega, P(T^S$ rejects $H_0^S) \neq \omega$ and $P(T^{F_1F_2}$ rejects $H_0^{F_1F_2}) \neq \omega$, with the actual rejection rate in all cases dependent on $\lambda_{02}(t|X)$ if $k = 1$ and $\lambda_{01}(t|X)$ if $k = 2$.*

3. *Under H_0, $P(T^{\lambda_1\lambda_2}$ rejects $H_0^{\lambda_1\lambda_2}) \neq \omega, P(T^{\lambda_k}$ rejects $H_0^{\lambda_k}) \neq \omega, P(T^S$ rejects $H_0^S) \neq \omega$ and $P(T^{F_1F_2}$ rejects $H_0^{F_1F_2}) \neq \omega$, with the actual rejection rate in all cases dependent on $\lambda_{01}(t|X)$ and $\lambda_{02}(t|X)$.*

4. *Under H_0^S, $P(T^{\lambda_1\lambda_2}$ rejects $H_0^{\lambda_1\lambda_2}) \neq \omega, P(T^{\lambda_k}$ rejects $H_0^{\lambda_k}) \neq \omega, P(T^{F_1}$ rejects $H_0) \neq \omega$ and $P(T^{F_1F_2}$ rejects $H_0^{F_1F_2}) \neq \omega$, with the actual rejection rate in all cases dependent on both $\lambda_{01}(t|X)$ and $\lambda_{02}(t|X)$.*

5. *Under $H_0^{F_1F_2}$, $P(T^{\lambda_1\lambda_2}$ rejects $H_0^{\lambda_1\lambda_2}) = \omega, P(T^{\lambda_k}$ rejects $H_0^{\lambda_k}) = \omega, P(T^{F_1}$ rejects $H_0) = \omega$ and $P(T^S$ rejects $H_0^S) = \omega$.*

The local null $H_0^{\lambda_1}$ is typically tested using the nonparametric log-rank test for the event of interest (Mantel, 1966; Peto and Peto, 1972; Cox, 1972), which is asymptotically equivalent to a Wald (or score) test for $\gamma_1 = 0$ based on the Cox model $\lambda_{01}(t|X) = \lambda_1(t)\exp(\gamma_1 X)$. In the simulations that follow in Section 3.4.2 we refer to these two test statistics as $T^{\lambda_1}_{LR}$ and $T^{\lambda_1}_{Cox}$. Similar tests can be considered for $H_0^{\lambda_2}$. A common test for the global null $H_0^{\lambda_1\lambda_2}$ is based on a Wald statistic for the hypothesis $\gamma_1 = \gamma_2 = 0$ in the joint Cox model $\lambda_{01}(t|X) = \lambda_1(t)\exp(\gamma_1 X), \lambda_{02}(t|X) = \lambda_2(t)\exp(\gamma_2 X)$. Since $\left[\widehat{\gamma}_1/se(\widehat{\gamma}_1)\right]^2 \sim \chi^2_1$ and $\left[\widehat{\gamma}_2/se(\widehat{\gamma}_2)\right]^2 \sim \chi^2_1$ are asymptotically chi-squared and independent, this is a two degrees of freedom test and we label the test statistic as $T^{\lambda_1\lambda_2}_{Cox}$. In a similar way, one could construct a χ^2 joint test based on two independent log-rank test statistics (Li and Yang, 2016). The null H_0^S can be tested based on a log-rank test statistic for the event-free survival time T, or based on a Wald statistic for $\zeta = 0$ in a cloglog model of the form (2.2.6); we disregard tests of H_0^S in simulations. Two common tests for H_0 are a generalized log-rank test due

to Gray (1988) and a Wald test based on (3.2.3). We label the Gray test statistic as $T_{Gray}^{F_1}$ and the Wald tests based on the FG and DB estimation procedures for β as $T_{FG}^{F_1}$ and $T_{DB}^{F_1}$, respectively. Tests for $H_0^{F_1F_2}$ have not received much attention. Shi et al. (2013) proposed a Wald test for the hypothesis $\theta = 0$ based on parametric models of the form $F_1(t|X) = 1 - (1 - F_1(\infty|X = 0)v(t; b_1, c_1))^{\exp(\theta X)}$ and $F_2(t|X) = (1 - F_1(\infty|X = 0))^{\exp(\theta X)}v(t; b_2, c_2)$, where $F_1(t|X = 0) = F_1(\infty|X = 0)v(t; b_1, c_1)$ satisfies a modified three-parameter logistic function with $v(t; b_k, c_k) \longrightarrow 1$ as $t \longrightarrow \infty$, $b_k > 0, c_k \in \mathbb{R}$. We discuss limitations of this class of models in Section 3.4.2.1. Alternatively, a Wald statistic for $\theta_1 = \theta_2 = 0$ in a multinomial logistic model of the form $\log(F_k(t|X)/S(t|X)) = A_k(t) + \theta_k X, k = 1, 2$ could be considered (Gerds et al., 2012). Various other less commonly used tests have been discussed in the literature (Li and Yang, 2016; Dobler and Pauly, 2017; Tiwari et al., 2006); we do not consider them here.

For the remainder of this section we focus on the FG- and DB-based Wald tests $T_{FG}^{F_1}$ and $T_{DB}^{F_1}$ for $\beta = 0$ and assess their properties under model misspecification. We let $\widehat{\beta}$ be the solution to the FG pseudo-score function (3.3.5) or the DB estimating equations (3.3.12), respectively. Statistics of the form $(\widehat{\beta} - 0)/se(\widehat{\beta})$ have in general an asymptotic $N(0, 1)$ distribution under $\beta = 0$, provided the underlying modeling assumptions are all satisfied. With a misspecified model (3.2.3) the null hypothesis actually being tested is $\beta_{FG}^\star = 0$ or $\beta_{DB}^\star = 0$, respectively, based on

$$\sqrt{n}(\widehat{\beta}_{FG} - \beta_{FG}^\star) \xrightarrow{\mathcal{D}} N(0, V_{FG}) \quad \text{or} \quad \sqrt{n}(\widehat{\beta}_{DB} - \beta_{DB}^\star) \xrightarrow{\mathcal{D}} N(0, V_{DB}) \,, \tag{3.4.6}$$

where the robust sandwich-type variance under the FG procedure takes the form

$$V_{FG} = \frac{\Omega(\beta_{FG}^\star)}{I^2(\beta_{FG}^\star)}$$

for $I(\beta) = \mathbb{E}\big(-\partial U^{FG}(\beta, G^\star)/\partial\beta\big)$ and $\Omega(\beta) = \mathbb{E}\big(\big[U^{FG}(\beta, G^\star)\big]^2\big) + \Psi$ (Fine and Gray, 1999; He et al., 2016). The second term of $\Omega(\beta)$ reflects the uncertainty in estimating the censoring distribution G; an explicit expression for Ψ can be found in Section 4 in Fine and Gray (1999) for the case of covariate-independent censoring and in Appendix A of He et al. (2016) for covariate-dependent censoring. The asymptotic variance $V_{DB} = \widetilde{\Omega}(\beta_{DB}^\star)/\tilde{I}^2(\beta_{DB}^\star)$ under the DB procedure is likewise of robust sandwich-type form; see Eq. (7) in Scheike et al. (2008) for further details. We note that robust variance estimation for $\widehat{\beta}$ is available for FG and DB regression in the `R` functions `crr()`, `comp.risk()` and `cifreg()` (Scheike and Zhang, 2011; Holst et al., 2016).

It is important to understand under which circumstances a Wald test based on $(\widehat{\beta} - 0)/se(\widehat{\beta})$ remains asymptotically valid even if (3.2.3) is misspecified. We conclude this section by stating the main result in the following theorem; a proof of Theorem 2 is provided in Appendix B.3.

Theorem 2 *When the cloglog model (3.2.3) for $F_1(t|X)$ is misspecified, a hypothesis test based on the Wald statistic*

$$(\widehat{\beta} - 0)/se(\widehat{\beta})$$

with a robust standard error estimate is asymptotically valid if the probability limit $\beta^\star$ of $\widehat{\beta}$ is 0 under the true process. As before, let $\widehat{\Gamma}_{FG}(t)$ denote the Breslow-type estimator for $\Gamma(t) = \exp(\alpha(t))$ under the FG procedure, and let $\widehat{\alpha}_r$ be the estimator for $\alpha_r = g(F_1(s_r|X = 0))$ under DB estimation, $r = 1, 2, \ldots, R$. Assume $P(X = 1) = 0.5$ and the regularity conditions in Appendix A.1 of He et al. (2016) and Appendix A of Scheike et al. (2008) such that $d\widehat{\Gamma}_{FG}(t) \longrightarrow d\Gamma^\star_{FG}(t), \widehat{\beta}_{FG} \longrightarrow \beta^\star_{FG}$ and $\widehat{\alpha}_{rDB} \longrightarrow \alpha^\star_{rDB}, \widehat{\beta}_{DB} \longrightarrow \beta^\star_{DB}$.

1. *If the FG procedure in Section 3.3.1.2 is used for estimation with a correctly specified censoring model ($G^\star = G$), then $\beta^\star_{FG} = 0$ under $H_0^{\lambda_1\lambda_2}$ and H_0, but $\beta^\star_{FG} \neq 0$ under $H_0^{\lambda_1}$ alone.*

2. *If we consider DB estimation at R distinct time points $s_1, s_2, \ldots, s_R$ (as in Section 3.3.1.2) with a correctly specified censoring model, and treat the baseline parameters α_r as unconstrained, then $\beta^\star_{DB} = 0$ under $H_0^{\lambda_1\lambda_2}$ and H_0, but $\beta^\star_{DB} \neq 0$ under $H_0^{\lambda_1}$.*

To summarize, the FG- and DB-based Wald tests of $\beta = 0$ are robust when the true process satisfies $H_0^{\lambda_1\lambda_2}$ or H_0, in the sense that they control the nominal level ω in large samples. This robustness property does not hold if robust variance estimates are not used.

3.4.2 Simulation studies

We report on simulation studies designed to assess the type I error rates of the FG- and DB-based Wald tests and their power for various alternatives. Although the main focus here is on tests based on cumulative incidence function regression models, we also examined the empirical rejection rates for a selection of intensity-based tests including $T^{\lambda_1\lambda_2}_{Cox}$, $T^{\lambda_1}_{LR}$ and $T^{\lambda_1}_{Cox}$ in order to confirm

the findings of Theorem 1 in empirical studies. Freidlin and Korn (2005) and Williamson et al. (2007) consider testing with competing risks processes based on the framework of latent failure times. The assumptions of this framework are impossible to assess in any given setting, unlike the intensity-based models we consider here.

3.4.2.1 Simulation setup

The binary treatment indicator X was simulated from a Bernoulli distribution with $P(X = 1) = P(X = 0) = 0.5$. To simulate competing risks data under $H_0^{\lambda_1\lambda_2}$ and $H_0^{\lambda_1}$, we generated processes with intensity functions $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X)$ following the procedure in Section 2.5 of Cook and Lawless (2018); for each of $X = 0, 1$ the event time $T|X$ was first generated from an exponential distribution with rate $\lambda_{01}(t|X) + \lambda_{02}(t|X)$. Given $T = t$, the event type was then drawn from a Bernoulli trial with probability $\lambda_{01}(t|X)/(\lambda_{01}(t|X) + \lambda_{02}(t|X))$ that the event was of the cause of interest. As in Section 3.3.2, with $\tau = 1$ we set the values for the time-homogeneous baseline intensities λ_1 and λ_2 to satisfy the constraints: $P(T \leq 1|X = 0) = 0.6$ and $P(T_1 < T_2|T \leq 1, X = 0) = 0.4, 0.6$ and 0.8. This gave type 1 event rates by $\tau = 1$ for individuals in the control group of $F_1(1|X = 0) = 0.24, 0.36$ and 0.48, and type 2 event rates of $F_2(1|X = 0) = 0.36, 0.24$ and 0.12, respectively. We further set $\exp(\gamma_1)$ to $0.6, 0.75, 0.9$ and 1, and $\exp(\gamma_2)$ to $0.5, 0.8, 0.9, 1, 1.1$ and 1.5. Note that $H_0^{\lambda_1\lambda_2}$ is true when $\exp(\gamma_1) = \exp(\gamma_2) = 1$ and $H_0^{\lambda_1}$ is true when $\exp(\gamma_1) = 1$. The random withdrawal time C_r was simulated from an exponential distribution with rate $\rho > 0$ set so that $\pi_r = P(C_r < T_1|T_1 \leq \min(T_2, 1)) = \mathbb{E}_X(P(C_r < T_1|T_1 \leq \min(T_2, 1), X))$ is 0.2, so 20% of the type 1 events occurring before the administrative censoring time are right-censored due to random loss to follow-up.

Section 5.3.6 of Beyersmann et al. (2012) discusses methods for simulating competing risks processes where $F_1(t|X)$ has the cloglog form. With their indirect simulation method, models for $F_1(t|X)$ and $F_2(t|X)$ take the form

$$\begin{aligned} g(F_1(t|X)) = g\big(F_1(\infty|X = 0)(1 - \exp[-t])\big) + \beta X \\ \iff \quad F_1(t|X) = 1 - \big(1 - F_1(\infty|X = 0)(1 - \exp[-t])\big)^{\exp(\beta X)} \end{aligned} \tag{3.4.7}$$

$$\log(F_2(t|X = 1)) = \exp(\beta)\log(F_2(\infty|X = 0)) + B(t; \beta)\ , \tag{3.4.8}$$

where $g(u) = \text{cloglog}(u), q := F_1(\infty|X = 0)$ is the proportion of type 1 events at $t = \infty$ in the

control group and $F_2(\infty|X) = 1 - F_1(\infty|X)$, $X = 0, 1$. Limitations to this approach include the fact that (3.4.8) does not have a standard GLM form since $B(t;\beta) = \log(1 - \exp[-t\exp(\beta)])$ is a function of β, and more importantly, $F_1(t|X)$ and $F_2(t|X)$ depend on the same β. If $\beta = 0$ in (3.4.7), then $F_1(t|X) = F_1(t)$ and $F_2(t|X) = F_2(t)$, and also $\lambda_{0k}(t|X) = \lambda_{0k}(t)$ for $k = 1, 2$; thus within the class of models given by (3.4.7) and (3.4.8) H_0 implies $H_0^{F_2}, H_0^{F_1F_2}, H_0^{\lambda_k}$ and $H_0^{\lambda_1\lambda_2}$. We therefore also considered models for $F_1(t|X)$ and $F_2(t|X)$ as follows:

$$g(F_1(t|X;\beta)) = g\big(F_1(\infty|X=0)(1-\exp[-\psi_1(t)])\big) + \beta X$$
$$\iff \quad F_1(t|X) = 1 - \big(1 - F_1(\infty|X=0)(1-\exp[-\psi_1(t)])\big)^{\exp(\beta X)} , \quad (3.4.9)$$
$$F_2(t|X;\beta,\beta_2) = (1 - F_1(\infty|X))g^{-1}(\alpha_2(t) + \beta_2 X) , \quad (3.4.10)$$

where $F_1(t|X=0) = q(1-\exp[-\psi_1(t)])$, $F_1(\infty|X) = 1-(1-q)^{\exp(\beta X)}$ and $\alpha_2(t) = \log(\psi_2(t))$ for $\psi_k(t) = t$, in which case (3.4.9) reduces to (3.4.7). If $\beta = 0$ in (3.4.9), then $F_2(\infty|X = 1) = F_2(\infty|X = 0)$. If $\beta_2 = 0$ and $\beta \neq 0$, then $F_2(t|X = 1) \neq F_2(t|X = 0)$. These models are both too restrictive for general use, but they provide a framework to check on the properties of tests under H_0 and alternatives based on (3.2.3). More generally we note that if $f_k(t|X) = \partial F_k(t|X)/\partial t$ is the subdensity function of T_k and $B(t;\beta) = \log\big(1-\exp(-t\exp(\beta))\big)$ as in Beyersmann et al. (2012), we obtain the $0-k$ intensities under models (3.4.7)-(3.4.8) and (3.4.9)-(3.4.10) as

$$\lambda_{0k}^{\dagger}(t|X) = \lim_{\Delta t \downarrow 0} \frac{P(T \in [t, t+\Delta t^-), Z(T) = k \mid T \geq t, X)}{\Delta t} = \frac{f_k(t|X)}{1 - F_1(t|X) - F_2(t|X)} . \quad (3.4.11)$$

We further note that both approaches satisfy the constraint $0 \leq F_1(t|X) + F_2(t|X) \leq 1$. Given X and $\lambda_{0k}^{\dagger}(t|X)$, the event time $T|X$ can then be generated using the inversion method: we first simulated a standard uniform random variable U, set the realized value u equal to $1 - S(t|X) = F_1(t|X) + F_2(t|X)$ and solved for t. Given $T = t$, the event type k can be determined by a Bernoulli experiment which assigns it to be a type 1 event with probability $\lambda_{01}^{\dagger}(t|X)/(\lambda_{01}^{\dagger}(t|X) + \lambda_{02}^{\dagger}(t|X)), k = 1, 2$. With $\tau = 1$, we specified q so that $F_1(1|X = 0) = 0.24, 0.36$ and 0.48, and set $\exp(\beta)$ to $0.8, 0.9, 1$ and 1.1, and $\exp(\beta_2)$ to 0.8 and 1. This gave type 2 event rates by $\tau = 1$ for individuals in the control group of $F_2(\tau|X = 0) = 0.39, 0.27$ and 0.15 under both approaches (3.4.7)-(3.4.8) and (3.4.9)-(3.4.10). As before, the random withdrawal time C_r follows an exponential distribution with rate $\rho > 0$ set to satisfy $\pi_r = 0.2$.

For each null model and parameter setting we consider the following tests: a two-d.f. Wald test for $H_0^{\lambda_1\lambda_2}$ based on Cox models for $\lambda_{01}(t|X), \lambda_{02}(t|X)$ or $\lambda_{01}^{\dagger}(t|X), \lambda_{02}^{\dagger}(t|X)$ ($T_{Cox}^{\lambda_1\lambda_2}$), the log-rank test for $H_0^{\lambda_1}$ ($T_{LR}^{\lambda_1}$), a one-d.f. Wald test for $H_0^{\lambda_1}$ based on a Cox cause-specific hazards model for $\lambda_{01}(t|X)$ or $\lambda_{01}^{\dagger}(t|X)$ ($T_{Cox}^{\lambda_1}$), Gray's test for H_0 ($T_{Gray}^{F_1}$), a one-d.f. Wald test for H_0 based on the FG procedure with a robust variance estimate ($T_{FG}^{F_1}$), and a one-d.f. Wald test for H_0 based on DB regression with $R = 6$ or $R = 3$ equi-spaced time points on $(0, 1)$ and a robust variance estimate ($T_{DB_R}^{F_1}$). We simulated datasets with $n = 1000$ individuals and conducted the tests using the `coxph()`, `survdiff()`, `cuminc()`, `crr()` and `comp.risk()` functions in the respective `R` packages `survival`, `cmprsk` and `timereg` (Therneau, 2024; Scheike and Zhang, 2011); as mentioned before, the `crr()` and `comp.risk()` functions for the FG and DB procedures automatically provide robust variance estimates. Moreover, unstratified KM estimation was used in these functions to get an estimate of the common censoring distribution $G(u) = P(C_r > u)$. We note that if $w(t)$ in the FG procedure is replaced by stabilized weights of the form $G(t)\mathbb{1}(C_r \geq \min(T, t))/G(\min(T, t))$, Theorem 2 continues to hold. We then calculated the empirical rejection rates as the proportion of $n_{sim} = 10000$ simulations in which the respective null hypothesis was rejected at the nominal 5% level ($\omega = 0.05$): with this many simulation if the test rejects the null at the nominal 5% level, we expect the empirical rejection rate to fall within the interval $[0.0457, 0.0543]$ 95% of the time.

3.4.2.2 Simulation results

Table 3.1 reports the empirical rejection rates of all tests considered for the setting where the data were generated according to the multiplicative intensity-based models and where $F_1(1|X = 0) = 0.36$ and $F_2(1|X = 0) = 0.24$ (i.e. $P(T_1 < T_2|T \leq 1, X = 0) = 0.6$), along with the corresponding limiting values $\beta_{FG}^{\star}, \beta_{DB_6}^{\star}$ and $\beta_{DB_3}^{\star}$ under FG and DB estimation. The $13th$ row (when $\exp(\gamma_1) = \exp(\gamma_2) = 1$) shows the results under the true process implied by $H_0^{\lambda_1\lambda_2}$; the $1st$ row of each block (when $\exp(\gamma_1) = 1$) represents the empirical type I error rates when $H_0^{\lambda_1}$ is true. In addition to Table 3.1, the empirical type I error rates for tests $T_{Gray}^{F_1}, T_{FG}^{F_1}, T_{DB_6}^{F_1}$ and $T_{DB_3}^{F_1}$ under $H_0^{\lambda_1\lambda_2}$ and $H_0^{\lambda_1}$ are also graphically depicted in the middle panel of Figure 3.3.

Consistent with Theorem 2, the FG- and DB-based Wald tests $T_{FG}^{F_1}, T_{DB_6}^{F_1}$ and $T_{DB_3}^{F_1}$ are valid under $H_0^{\lambda_1\lambda_2}$ in the sense that their type I error rate is compatible with the nominal 5% level. Gray's test for H_0, $T_{Gray}^{F_1}$, also shows good type I error control under $H_0^{\lambda_1\lambda_2}$. Under $H_0^{\lambda_1}$, all tests

for H_0 have inflated type I error rates except when $\exp(\gamma_2) = 1$. The inflation is very small when $\exp(\gamma_2)$ is close to 1 (see e.g. the $1st$ row of the $3rd$ block), but increases as the magnitude of $\exp(\gamma_2)$ increases and as the proportion of type 1 events in the control group decreases; see Figure 3.3. This agrees with the theoretical results presented in Theorem 1. The FG statistics $T^{F_1}_{Gray}$ and $T^{F_1}_{FG}$ have a larger type I error inflation than $T^{F_1}_{DB_6}$ and $T^{F_1}_{DB_3}$; note that $\beta^{\star}_{DB_6}$ and $\beta^{\star}_{DB_3}$ are closer to 0 than $\beta^{\star}_{FG}$. The tests $T^{\lambda_1}_{LR}$ and $T^{\lambda_1}_{Cox}$ for $H^{\lambda_1}_0$ yield very similar type I error rates consistent with the nominal 5% level under $H^{\lambda_1\lambda_2}_0$ and $H^{\lambda_1}_0$, as expected. The test $T^{\lambda_1\lambda_2}_{Cox}$ maintains control over the type I error rate under $H^{\lambda_1\lambda_2}_0$, but has inflated type I error rates under $H^{\lambda_1}_0$; as per Theorem 1, the inflation varies with the magnitude of $\exp(\gamma_2)$ and $P(T_1 < T_2|T \leq 1, X = 0)$; see Tables B.1 and B.2 in Section B.4 of Appendix B where $P(T_1 < T_2|T \leq 1, X = 0) = 0.4$ and 0.8.

Simulation results for scenarios where the data were generated according to models (3.4.7)-(3.4.8) or (3.4.9)-(3.4.10) and where $F_1(1|X = 0) = 0.36$ are given in Table 3.2. The $3rd$ row of the first block and the $5th$ and $6th$ rows of the second block (when $\exp(\beta) = 1$) represent the empirical rejection rates under H_0. For the former case, since H_0 implies $H^{\lambda_1}_0$ and $H^{\lambda_1\lambda_2}_0$ within this class of models, all tests have good type I error control. For the latter case, the empirical type I error rates of tests $T^{F_1}_{FG}, T^{F_1}_{DB_6}, T^{F_1}_{DB_3}$ and $T^{F_1}_{Gray}$ are close to the nominal level, confirming the findings in Theorem 2. The intensity-based tests $T^{\lambda_1}_{LR}, T^{\lambda_1}_{Cox}$ and $T^{\lambda_1\lambda_2}_{Cox}$ control the nominal level if $\exp(\beta_2) = 1$, in which case (3.4.9)-(3.4.10) reduces to (3.4.7)-(3.4.8). When there is a separate effect of treatment on $F_2(t|X)$ (i.e. $\exp(\beta_2) \neq 1$), these tests are based on non-proportional intensities and hence give inflated type I error rates; see also Table B.3 in Appendix B.4.

We next consider the power of the tests under a variety of different alternatives and begin with the results presented in Tables 3.1, B.1 and B.2 where $P(T_1 < T_2|T \leq 1, X = 0) = 0.6, 0.4$ and 0.8, respectively. As one would expect, the power of all tests increases with increasing magnitude of γ_1, the treatment effect on the $0 - 1$ intensity, and increasing proportion of type 1 events, here through $P(T_1 < T_2|T \leq 1, X = 0)$; see also Figure 3.4. The power of tests for $H^{\lambda_1}_0$ is unaffected by γ_2, with $T^{\lambda_1}_{LR}$ and $T^{\lambda_1}_{Cox}$ being equally powerful. These results are widely accepted, and we now compare T^{λ_1} and T^{F_1} in terms of power. First we note that the FG statistics $T^{F_1}_{Gray}$ and $T^{F_1}_{FG}$ are equally powerful in all scenarios. The power of $T^{F_1}_{FG}, T^{F_1}_{DB_6}$ and $T^{F_1}_{DB_3}$ is in general affected by the corresponding probability limits $\beta^{\star}$ and the asymptotic variance estimators of $\widehat{\beta}$. Although $\beta^{\star}_{DB_R}$ is quite robust to the number of time points R, i.e. $\beta^{\star}_{DB_6} \approx \beta^{\star}_{DB_3}$, the asymptotic variances become larger with decreasing R; the DB_6 analysis is thus expected to be more powerful than the

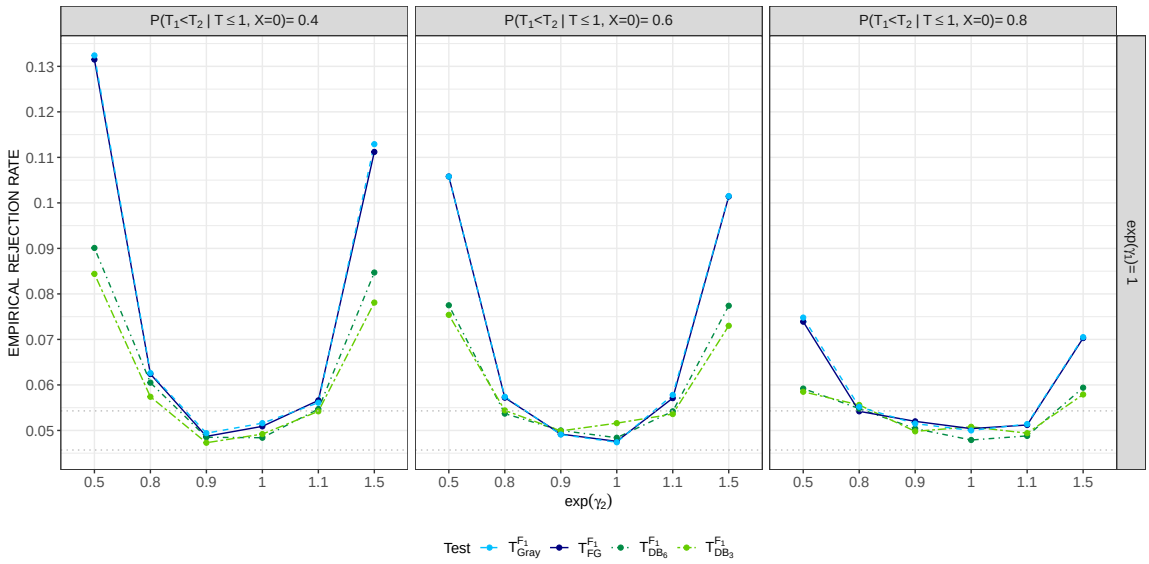


Figure 3.3: Empirical type I error rates of tests T^{F_1} under $H_0^{\lambda_1\lambda_2}$ ($\exp(\gamma_1) = \exp(\gamma_2) = 1$) and $H_0^{\lambda_1}$ ($\exp(\gamma_1) = 1$) for different values of $\exp(\gamma_2)$ and $P(T_1 < T_2|T \leq 1, X = 0) = 0.4$ (left panel), 0.6 (middle panel) and 0.8 (right panel), nominal test size is $\omega = 0.05; n = 1000$ individuals, $n_{sim} = 10000$, $\tau = 1, P(T \leq 1|X = 0) = 0.6, \pi_r = 0.2$.

DB_3 analysis, and this is what the simulation study reveals. Compared to $T^{F_1}_{FG}$, the DB_6 statistic leads to a less powerful analysis in most settings.

Figure 3.4 depicts the power of $T^{\lambda_1}_{LR}, T^{F_1}_{Gray}, T^{F_1}_{FG}$ and $T^{F_1}_{DB_6}$ as a function of $\exp(\gamma_1)$ and $\exp(\gamma_2)$ for the different values of $P(T_1 < T_2|T \leq 1, X = 0)$. When $\exp(\gamma_1) < 1$ and $\exp(\gamma_2) = 1$, the tests $T^{\lambda_1}_{LR}, T^{F_1}_{Gray}$ and $T^{F_1}_{FG}$ achieve comparable power, while $T^{F_1}_{DB_6}$ is slightly less powerful; see the third set of panels from the right. If instead $\exp(\gamma_2) < 1$ (or $\exp(\gamma_2) > 1$), the tests $T^{F_1}_{Gray}, T^{F_1}_{FG}$ and $T^{F_1}_{DB_6}$ are less (or more) powerful than $T^{\lambda_1}_{LR}$; see the first three sets of panels from the left and the first two sets of panels from the right, respectively. However, when $\exp(\gamma_2) > 1$, the tests for H_0 have inflated type I error rates. Interestingly, when there is only a relatively small treatment effect on the $0 - 2$ intensity (i.e. $\exp(\gamma_2) = 1.1$), $T^{F_1}_{DB_6}$ is less powerful than $T^{\lambda_1}_{LR}$.
Naturally as the treatment effect β on $F_1(t|X)$ in Table 3.2 decreases, the power of the FG- and DB-based Wald tests $T^{F_1}_{FG}, T^{F_1}_{DB_6}$ and $T^{F_1}_{DB_3}$ decreases as well. If data were generated by models (3.4.9)-(3.4.10), the power of $T^{F_1}_{FG}, T^{F_1}_{DB_6}$ and $T^{F_1}_{DB_3}$ is unaffected by β_2.

		REJECTION RATES							PROBABILITY LIMITS		
		$H_0^{\lambda_1}$		$H_0^{\lambda_1\lambda_2}$	H_0						
$\exp(\gamma_2)$	$\exp(\gamma_1)$	$T_{LR}^{\lambda_1}$	$T_{Cox}^{\lambda_1}$	$T_{Cox}^{\lambda_1\lambda_2}$	$T_{Gray}^{F_1}$	$T_{FG}^{F_1}$	$T_{DB_6}^{F_1}$	$T_{DB_3}^{F_1}$	$\beta_{FG}^{\star}$	$\beta_{DB_6}^{\star}$	$\beta_{DB_3}^{\star}$
0.5	1	0.0479	0.0476	0.9691	0.1058	0.1058	0.0775	0.0754	0.0825	0.0593	0.0560
	0.9	0.1466	0.1459	0.9771	0.0557	0.0550	0.0617	0.0649	-0.0210	-0.0453	-0.0487
	0.75	0.6549	0.6535	0.9938	0.3702	0.3697	0.3849	0.3753	-0.2003	-0.2264	-0.2299
	0.6	0.9791	0.9786	0.9997	0.9055	0.9052	0.8905	0.8803	-0.4205	-0.4483	-0.4520
0.8	1	0.0477	0.0470	0.2303	0.0574	0.0572	0.0537	0.0544	0.0324	0.0234	0.0221
	0.9	0.1350	0.1337	0.3093	0.0840	0.0832	0.0886	0.0864	-0.0706	-0.0808	-0.0822
	0.75	0.6481	0.6464	0.6923	0.5238	0.5229	0.4879	0.4755	-0.2495	-0.2614	-0.2630
	0.6	0.9763	0.9762	0.9734	0.9509	0.9511	0.9295	0.9166	-0.4691	-0.4828	-0.4846
0.9	1	0.0478	0.0473	0.0862	0.0491	0.0492	0.0500	0.0499	0.0161	0.0117	0.0110
	0.9	0.1403	0.1398	0.1600	0.1106	0.1113	0.1057	0.1046	-0.0868	-0.0924	-0.0932
	0.75	0.6477	0.6464	0.5771	0.5824	0.5804	0.5289	0.5120	-0.2655	-0.2728	-0.2738
	0.6	0.9740	0.9737	0.9548	0.9592	0.9593	0.9398	0.9306	-0.4850	-0.4941	-0.4953
1	1	0.0493	0.0484	0.0534	0.0474	0.0476	0.0484	0.0516	0.0000	0.0000	0.0000
	0.9	0.1446	0.1430	0.1171	0.1390	0.1401	0.1296	0.1228	-0.1028	-0.1040	-0.1041
	0.75	0.6406	0.6388	0.5304	0.6211	0.6219	0.5610	0.5402	-0.2814	-0.2842	-0.2846
	0.6	0.9729	0.9727	0.9469	0.9697	0.9688	0.9470	0.9373	-0.5007	-0.5054	-0.5060
1.1	1	0.0542	0.0537	0.0872	0.0578	0.0571	0.0542	0.0536	-0.0159	-0.0116	-0.0109
	0.9	0.1426	0.1419	0.1540	0.1699	0.1687	0.1479	0.1420	-0.1187	-0.1154	-0.1150
	0.75	0.6322	0.6297	0.5614	0.6597	0.6585	0.5942	0.5706	-0.2971	-0.2956	-0.2954
	0.6	0.9696	0.9692	0.9484	0.9741	0.9740	0.9516	0.9435	-0.5163	-0.5166	-0.5166
1.5	1	0.0519	0.0509	0.7824	0.1015	0.1014	0.0774	0.0730	-0.0780	-0.0569	-0.0540
	0.9	0.1385	0.1375	0.8145	0.3136	0.3123	0.2422	0.2246	-0.1805	-0.1604	-0.1577
	0.75	0.6225	0.6203	0.9387	0.7994	0.7989	0.7107	0.6787	-0.3585	-0.3400	-0.3376
	0.6	0.9701	0.9699	0.9955	0.9923	0.9920	0.9769	0.9705	-0.5774	-0.5605	-0.5583

Table 3.1: Empirical rejection rates for different statistical tests of nominal size $\omega = 0.05$ when the true process has intensities $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X)$, along with the limiting values $\beta_{FG}^{\star}$ and $\beta_{DB}^{\star}$; $\gamma_1 : \log(0.9) = -0.1054, \log(0.75) = -0.2877, \log(0.6) = -0.5108$; $n = 1000$ individuals, $n_{sim} = 10000$, $\tau = 1$, $P(T \leq 1|X = 0) = 0.6$, $P(T_1 < T_2|T \leq 1, X = 0) = 0.6$, $\pi_r = 0.2$.

			REJECTION RATES							PROBABILITY LIMITS
			$H_0^{\lambda_1}$		$H_0^{\lambda_1\lambda_2}$	$H_0^{F_1}$				
Models	$\exp(\beta)$	$\exp(\beta_2)$	$T_{LR}^{\lambda_1}$	$T_{Cox}^{\lambda_1}$	$T_{Cox}^{\lambda_1\lambda_2}$	$T_{Gray}^{F_1}$	$T_{FG}^{F_1}$	$T_{DB_6}^{F_1}$	$T_{DB_3}^{F_1}$	$\beta_{FG}^{\star}, \beta_{DB_6}^{\star}, \beta_{DB_3}^{\star}$
(3.4.7) - (3.4.8)	0.8	-	0.4921	0.4906	0.3975	0.4530	0.4514	0.4022	0.3851	-0.2231
	0.9		0.1569	0.1554	0.1208	0.1441	0.1447	0.1327	0.1279	-0.1054
	1		0.0493	0.0489	0.0482	0.0473	0.0472	0.0475	0.0464	0.0000
	1.1		0.1338	0.1326	0.1045	0.1264	0.1267	0.1144	0.1123	0.0953
(3.4.9) - (3.4.10)	0.8	0.8	0.4820	0.4791	0.3865	0.4441	0.4433	0.3926	0.3737	-0.2231
		1	0.3673	0.3654	0.4230	0.4496	0.4491	0.3982	0.3810	
	0.9	0.8	0.1866	0.1850	0.1969	0.1432	0.1423	0.1279	0.1229	-0.1054
		1	0.1207	0.1199	0.1248	0.1437	0.1427	0.1276	0.1267	
	1	0.8	0.0750	0.0738	0.4432	0.0501	0.0500	0.0488	0.0507	0.0000
		1	0.0492	0.0488	0.0480	0.0472	0.0471	0.0474	0.0463	
	1.1	0.8	0.0709	0.0703	0.3573	0.1322	0.1319	0.1186	0.1167	0.0953
		1	0.1130	0.1122	0.1149	0.1311	0.1307	0.1211	0.1164	

Table 3.2: Empirical rejection rates for different statistical tests of nominal size $\omega = 0.05$ when the true process is implied by models (3.4.7)-(3.4.8) and (3.4.9)-(3.4.10), along with the limiting values $\beta_{FG}^{\star}$ and $\beta_{DB}^{\star}$; $\beta : \log(0.8) = -0.2231, \log(0.9) = -0.1054, \log(1.1) = 0.0953$; $n = 1000$ individuals, $n_{sim} = 10000$, $\tau = 1, F_1(1|X = 0) = 0.36, F_2(1|X = 0) = 0.27, \pi_r = 0.2$.

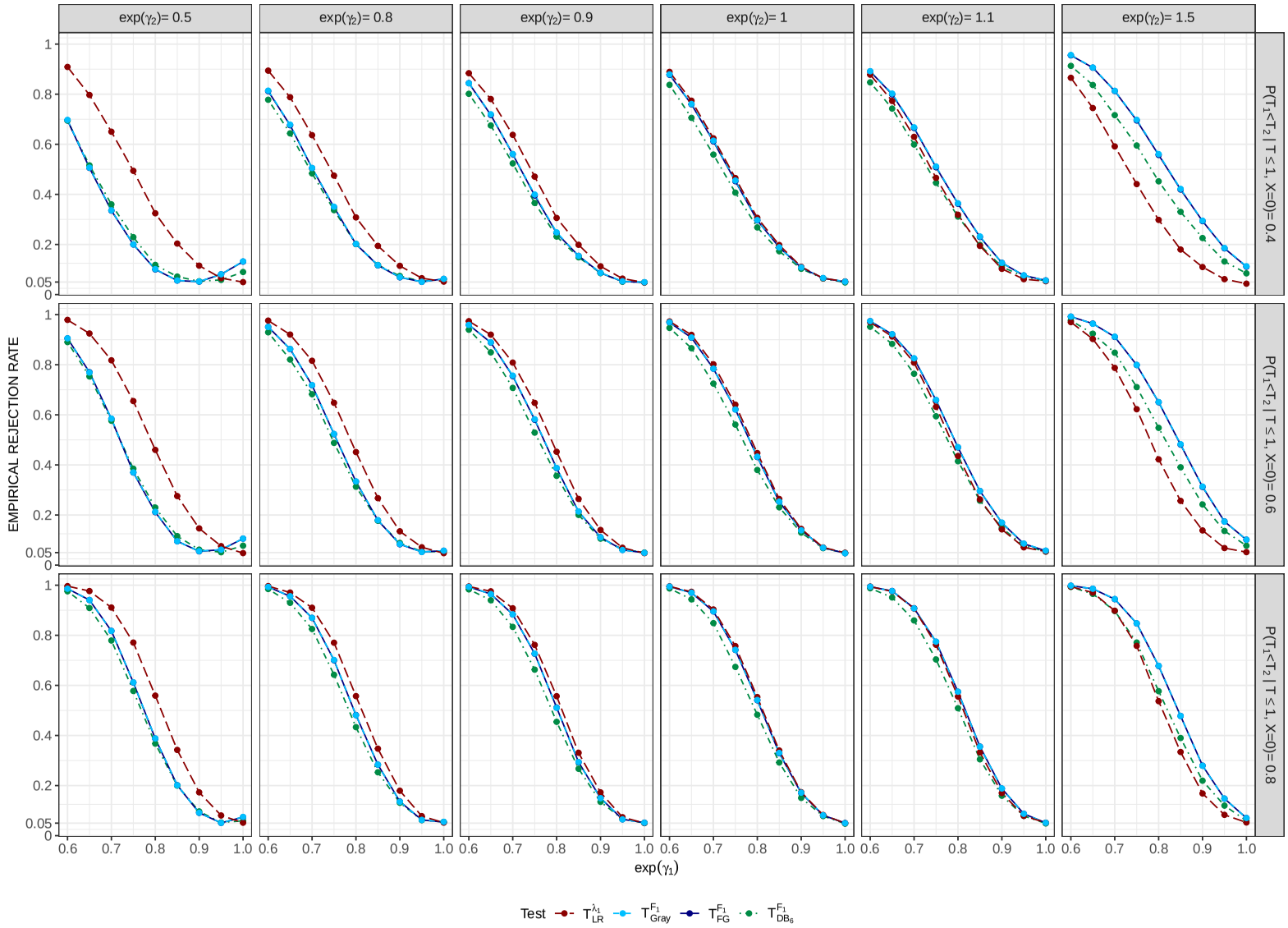


Figure 3.4: Empirical rejection rates for different statistical tests under various alternatives; $n = 1000$ individuals, $n_{sim} = 10000$, $\tau = 1$, $P(T \leq 1|X = 0) = 0.6$, $P(T_1 < T_2|T \leq 1, X = 0) = 0.4$ (top row), 0.6 (middle row) and 0.8 (bottom row), $\pi_r = 0.2$.

3.4.2.3 Implications of simulation studies and recommendations

Some recommendations based on the simulation results are as follows:

1. If the primary objective of a trial is in testing H_0, we recommend use of robust Wald statistics such as $T_{FG}^{F_1}$ and $T_{DB_R}^{F_1}$, e.g. $R = 6$. We have shown that these tests are asymptotically valid under $H_0^{\lambda_1\lambda_2}$ and H_0 in the sense that they control the type I error. With true processes satisfying $H_0^{\lambda_1}$ alone however, the type I error is close to the nominal level ω only if $\beta^\star \approx 0$; in settings where both $\exp(\gamma_2)$ and $P(T_1 < T_2 | T \geq 1, X = 0)$ are not too far from 1 the limiting values $\beta_{FG}^\star$ and $\beta_{DB}^\star$ are found to be roughly 0. For the case where $\exp(\gamma_2) = 0.8$ and $P(T_1 < T_2 | T \leq 1, X = 0) = 0.6$, for instance, the empirical rejection rates for $T_{FG}^{F_1}$ and $T_{DB_6}^{F_1}$ were at most 0.057 when $n = 1000$, 0.055 when $n = 500$ and 0.050 when $n = 250$; see the $5th$ rows in Tables B.4 and B.7 in Appendix B.4. We characterize rejection of H_0 under $H_0^{\lambda_1}$ as a type I error. In fact since the CIFs are determined by both $\lambda_{01}(t|X)$ and $\lambda_{02}(t|X)$, rejection of H_0 is a true positive finding in settings where $\beta^\star \not\approx 0$, even if it is not for the anticipated reason (i.e. a reduction in the $0 - 1$ intensity). The FG statistic tends to yield more powerful tests than the DB_6 statistic in settings most closely aligned with real-world applications.

2. If in secondary intensity-based analyses interest is in $H_0^{\lambda_1}$, tests based on the log-rank statistic $T_{LR}^{\lambda_1}$ or the Cox Wald statistic $T_{Cox}^{\lambda_1}$ are recommended and commonly used in practice. The performance of tests for $H_0^{\lambda_1}$ when H_0 is true is of less concern; the same applies to tests for $H_0^{\lambda_1\lambda_2}$. A discussion of multiplicity adjustments is beyond the scope of this thesis.

3.5 Prevention of skeletal complications in metastatic breast cancer

The analysis presented in this section considers data from the palliative therapy trial in stage IV breast cancer metastatic to bone (see Section 1.3.3, Hortobagyi et al., 1996), and examines the effect of pamidronate versus placebo on the cumulative incidence of skeletal complications, termed skeletal-related events (SREs). The Aalen-Johansen (AJ) estimates of $F_1(t|X)$ in the left panel of Figure 3.5 (black curves) show a steadily increasing cumulative probability of experiencing SREs in both treatment groups, with 2-year estimates of roughly 56% in the pamidronate group (X=1) and 78% in the placebo group (X=0); the two cumulative incidence curves begin to separate at 3 months. Gray (1988)'s test of no treatment effect on the CIF for SREs gives a p-value of 0.000226, and was carried out using the `R` function `cuminc()` in the competing risks package `cmprsk`. We note that this function also gives the nonparametric AJ estimate, but we obtained the CIF curves in Figure 3.5 using the `etm()` function (Allignol et al., 2011).

Plots of $g\big(\widehat{F}_1^{AJ}(t|X=1)\big) - g\big(\widehat{F}_1^{AJ}(t|X=0)\big)$ along with pointwise 95% CIs (constructed by the Δ-method) can help to find an appropriate link function $g(\cdot)$. According to the middle and right panels of Figure 3.5, both the log and cloglog transformation offer reasonable approximations; the logit link approximates the differences in $\widehat{F}_1^{AJ}(t|X=1)$ and $\widehat{F}_1^{AJ}(t|X=0)$ less well. In this analysis we restrict attention to $g(u) = \text{cloglog}(u)$ and compare the FG and DB estimation procedures. We used the `crr()` function to fit model (3.2.3) under the FG procedure and `comp.risk()` for DB regression (Scheike and Zhang, 2011), as before in Section 3.4.2.1.
There was no evidence that censoring was dependent on covariates, so Kaplan-Meier (KM) estimates of a common censoring distribution were used in (3.3.1)-(3.3.2) for the FG estimator and (3.3.12) for the DB estimator. The FG procedure yielded an estimated treatment effect of $\widehat{\beta}_{FG} = -0.516$ (robust SE=0.141, 95% CI: $-0.793, -0.239$, p=0.00026), while DB estimation with $R = 10$ equi-spaced time points over $(0, 21)$ yielded $\widehat{\beta}_{DB} = -0.526$ (robust SE=0.153, 95% CI: $-0.827, -0.226$, p=0.000601). There is not much difference in the estimated treatment effects, but a slight increase in the associated standard errors with the DB approach. The analysis suggests that pamidronate has a significant effect on the cumulative probability of SREs. The left panel of Figure 3.5 also gives the estimated cumulative incidence curves obtained from fitting the semiparametric cloglog model (3.2.3) under each estimation procedure.

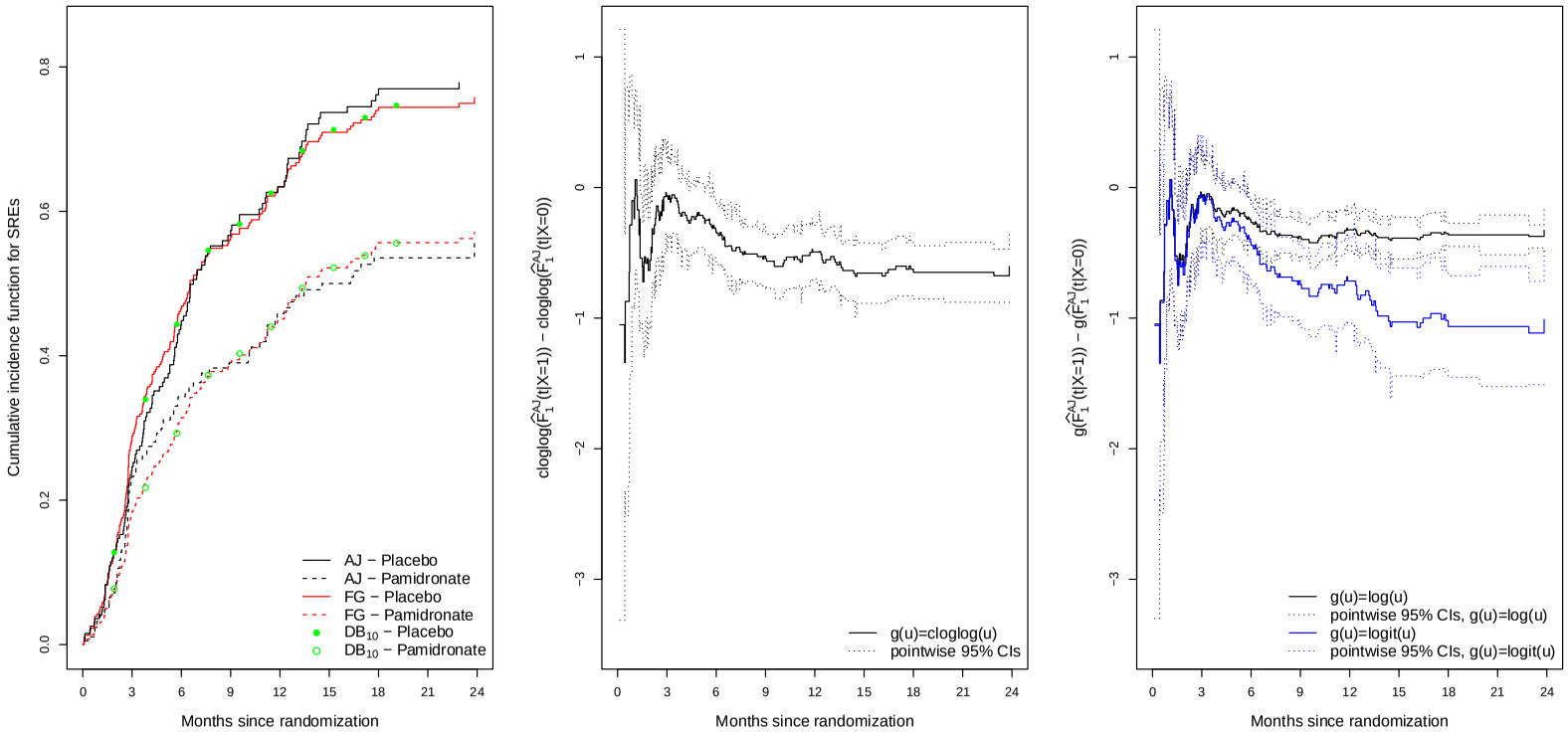


Figure 3.5: Left panel: comparisons of nonparametric Aalen-Johansen (AJ) estimates of the cumulative incidence function for skeletal-related events (SREs) and semiparametric estimates implied by the Fine-Gray (FG) method and direct binomial (DB) estimation (with $R = 10$ equi-spaced time points in $(0, 21)$) under the cloglog transformation model (3.2.3), middle and right panels: plots of $g\big(\widehat{F}_1^{AJ}(t|X = 1)\big) - g\big(\widehat{F}_1^{AJ}(t|X = 0)\big)$ for $g(\cdot) = \text{cloglog}(\cdot), \log(\cdot)$ and $\text{logit}(\cdot)$ along with pointwise 95% CIs; $X = 1$: pamidronate group, $X = 0$: placebo group.

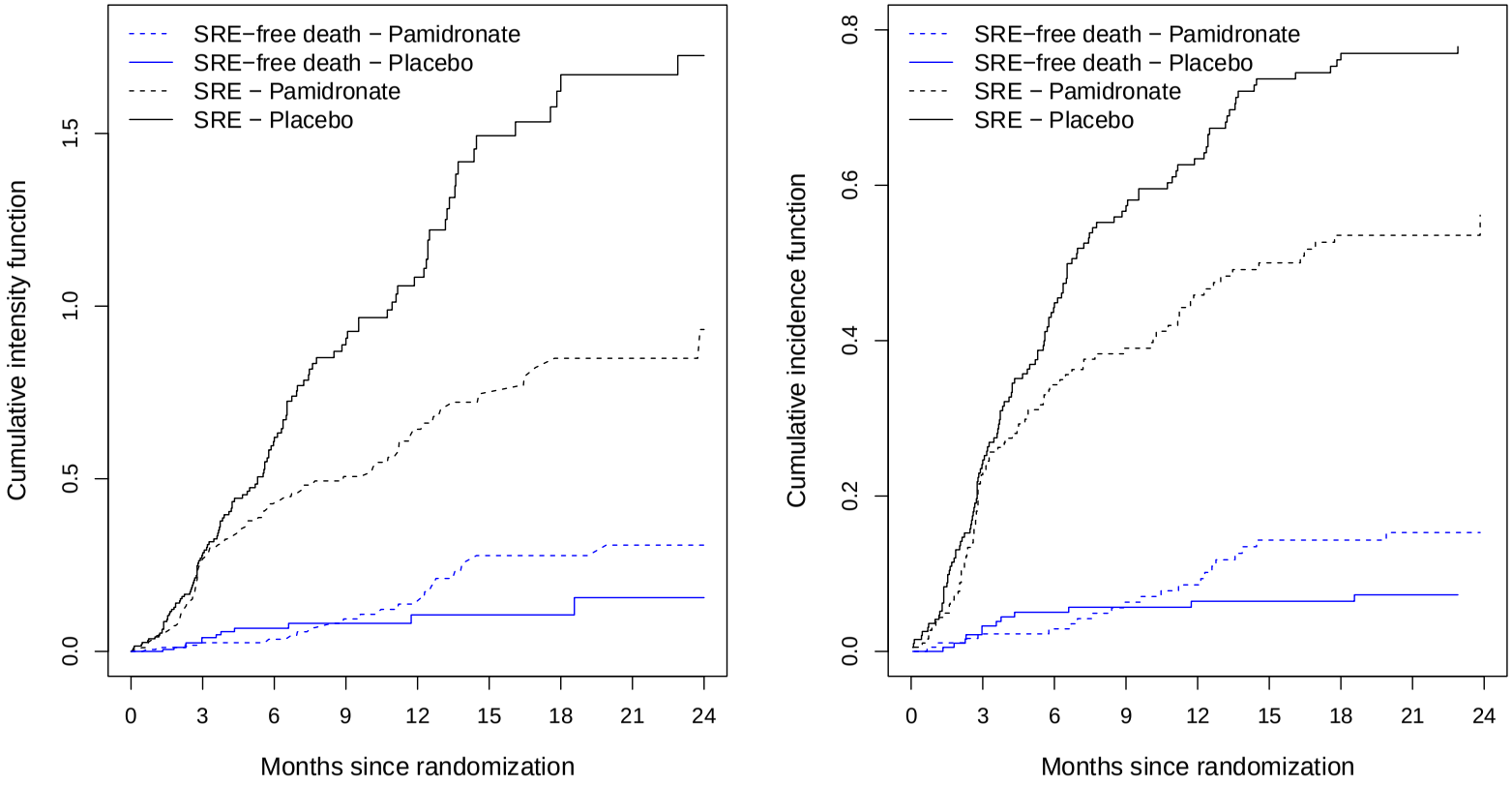


Figure 3.6: Nelson-Aalen estimates of the cumulative intensity functions (left panel) and Aalen-Johansen estimates of the cumulative incidence functions (right panel) for skeletal complications and complication-free death for breast cancer patients in Hortobagyi et al. (1996).

			Test statistic	p-value
FG	Wald test of $H_0 : \nu = 0$	$b(t) = t$	2.444	0.1180
		$b(t) = \log(t)$	1.967	0.1608
	Score test of $H_0 : \nu = 0$	$b(t) = t$	2.690	0.1010
		$b(t) = \log(t)$	2.148	0.1427
DB	Wald test of $H_0 : \nu = 0$	$b(t) = t$	2.262	0.1326
	Wald test of $H_0 : \beta_r = \beta$ based on (3.3.19) and 10000 bootstrap samples		14.622	0.1018

Table 3.3: Goodness-of-fit-tests for checking the assumption of a time-constant treatment effect.

Any difference between the semiparametric and the nonparametric estimates is an indirect reflection of model inadequacy. Both methods and the cloglog model offer a reasonable representation of the treatment effect, with some lack of fit early and later on in the trial. We complement the graphical model assessment with formal goodness-of-fit tests described in Section 3.3.3. Results can be found in Table 3.3, where the `R` package `crrSC` was used to perform the FG-based score test (Zhou et al., 2013). All other test statistics were calculated based on the results provided by `crr()` and `comp.risk()`. There is insufficient evidence to reject the assumption of a constant treatment effect over time, providing some justification for model (3.2.3).

Further secondary analyses were carried out to enhance understanding of the marginal treatment effects $\widehat{\beta}_{FG}$ and $\widehat{\beta}_{DB}$, including nonparametric and semiparametric analyses of the cause-specific hazard (CSH) functions for the time to SRE and death. Plots of the nonparametric Nelson-Aalen (NA) estimates of the CSH functions are displayed in the left panel of Figure 3.6 for each event type, and were obtained using the `mvna` package (Allignol et al., 2008). Within the first 3 months of follow-up pamidronate does not show any effect on the risk of SRE and death, which explains why $\widehat{F}_1^{AJ}(t|X=0) \approx \widehat{F}_1^{AJ}(t|X=1)$ and $\widehat{F}_2^{AJ}(t|X=0) \approx \widehat{F}_2^{AJ}(t|X=1)$ for $t \leq 3$ months; see right panel of Figure 3.6 and also (3.2.1). Overall, patients treated with pamidronate have a significantly reduced hazard of skeletal complications (HR=$e^{\widehat{\gamma}_1} = 0.590$, 95% CI: $0.447, 0.778$, p=0.000185; log-rank test: p=0.0002) as compared to placebo patients; the results shown in the round brackets were obtained using the `coxph()` and `survdiff()` functions in the `survival` package (Therneau, 2024). The risk of death, however, is slightly higher in the pamidronate arm; the crossing cause-specific hazards for complication-free death suggest that a Cox analysis would not be suitable in this case. The cumulative hazard in the placebo arm is relatively flat for

complication-free death beyond 6 months, indicating that most complication-free deaths occurred in placebo patients within the first 6 months of follow-up. The secondary analyses demonstrate that the difference in the cumulative incidence curves for SREs is mainly due to the reduction in the intensity for SRE by the use of pamidronate.

3.6 Discussion and recommendations

Cumulative incidence function regression, which includes the Fine-Gray model as a special case, has seen increasing use for the analysis of competing and semi-competing risks data in randomized clinical trials. Based on large sample results we have studied treatment effects in FG- and DB-based CIF regression when the true process has specified intensities, and established a robustness property for FG- and DB-based Wald tests of no treatment effect on $F_1(t|X)$ under key hypotheses concerning the intensity-based process. We focused on the cloglog transformed CIF model, but other link functions can be likewise studied; see for instance Blanche et al. (2023) who recently investigated logistic CIF regression under binomial estimation. While DB estimation (Scheike et al., 2008) can be used with any link function, the original FG procedure (Fine and Gray, 1999) is defined in terms of the cloglog link. A key message in this thesis chapter is the need for secondary intensity-based analysis when assessing the effect of a randomized treatment on the CIF for the event of interest. We provide a brief summary of this chapter's main points and findings in the next paragraph.

CIF regression targets a descriptive causal estimand based on an observable process feature, but its interpretation is only meaningful if the underlying model assumptions are plausible. Both estimands $\beta^\star_{FG}$ and $\beta^\star_{DB}$ are functions of the degree of misspecification and the intensities of the underlying process including the treatment effects. We found that $\beta^\star$ has a simple interpretation if the incidence rate of the competing event is low and there is a relatively modest effect on the corresponding $0-2$ intensity, since in this case $\beta^\star$ is in close agreement with the treatment effect on the $0-1$ intensity. In numerical investigations the censoring distribution was assumed to be "correctly" specified; in the case of a misspecified censoring model $\beta^\star$ is affected by the censoring process. It is straightforward to adapt the derivations of $\beta^\star$ given in Section 3.3.1 to investigate how sensitive the limiting value is to such a misspecification. We refer to He et al.

(2016) and Klein and Andersen (2005) who investigated in empirical studies the consequences of a misspecified censoring model on the estimation of β when (3.2.3) is valid.
The robustness property in Theorem 2 supports the use of FG- and DB-based Wald tests using robust standard errors in primary analyses of clinical trials. Similar robustness properties have been established for hazard-based Cox analyses in the survival setting (DiRienzo and Lagakos, 2001) or logistic regression in binary response settings (Rosenblum and van der Laan, 2009).

The planning stage of a trial, with primary estimation and testing based on model (3.2.3), should involve careful consideration of the process intensities, and how treatment may affect them. Control data collected in similar studies can offer insights into the nature of baseline intensities. Graphs such as those in Figure 3.1 can help to anticipate for $\widehat{\beta}$ under different assumed treatment effects in the true process. This could in turn be used for power and sample size calculations.
According to the sample size formula for FG regression by Latouche et al. (2004), the number of individuals n required to ensure $100(1-\omega')\%$ power for a two-sided Wald test of $\beta = 0$ vs $(\beta = \beta_1)$ at the $100\omega\%$ significance level is given by

$$n = \frac{1}{P(T_1 \leq C)} \cdot \frac{(z_{1-\omega/2} + z_{1-\omega'})^2}{\beta_1^2 P(X=1)P(X=0)}\,, \tag{3.6.1}$$

where $P(T_1 \leq C)$ is the proportion of observed type 1 events by the end of follow-up and z_q is the $100q\%$ percentile of the standard normal distribution. Their Schoenfeld-type formula was derived under the simplistic assumption of "complete data" (c.f. Section 3.1 of Fine and Gray, 1999), in which case the FG at-risk indicator is known for all individuals and no inverse probability of censoring weighting in (3.3.1) and (3.3.2) is needed (i.e. $w_i(t) = 1$ for all t and $i = 1, \ldots, n$). As a result, (3.6.1) was obtained by using a simple "naive" model-based variance estimator for $\widehat{\beta}$ of the form $(P(X=1)P(X=0)N_{1\cdot}(\infty))^{-1}$ under both the null and alternative hypotheses, and so does theoretically not apply to trials where random right-censoring occurs. Simulation studies in Latouche et al. (2004) suggest that (3.6.1) provides a reasonable approximation in finite samples for a limited range of trial settings, provided the assumed model (3.2.3) is correct. Further study of the utility of the Latouche et al. (2004) formula for more complex scenarios, e.g. when there is moderate to heavy right-censoring and the censoring intensity may depend on covariates, is however warranted. A closed-form sample size formula for DB regression is more challenging to derive and beyond the scope of this chapter. For now, we suggest a simulation-based approach to estimate power for DB-based Wald tests.

The impact of non-proportional hazards in power calculations for the log-rank test has been extensively studied (Zhang and Quan, 2009; Royston and Parmar, 2016; Yung and Liu, 2020; Dormuth et al., 2023). Similar investigations with FG and DB regression have not yet received much attention. Large sample results such as those in (3.4.6) could facilitate sample size calculations for FG and DB regression under model misspecification.

Chapter 4

Mean function regression in clinical trials involving recurrent and terminal event processes: interpretability of estimands

4.1 Introduction

Clinical trials have traditionally been designed based on the effect of treatment on the time to the first non-fatal event, or on the time to the first non-fatal or fatal event. In disease processes with frequent recurrent non-fatal events this ignores much information (Anker and McMurray, 2012; Anker et al., 2016; Claggett et al., 2018). In the palliative trial (II) of Section 1.3.3, for example, advanced breast cancer patients with bone metastases may experience more than one skeletal-related event prior to death or loss to follow-up (Hortobagyi et al., 1996). Other examples include recurrent relapses in progressive forms of multiple sclerosis (Montalban et al., 2017) or recurrent myocardial infarctions (MIs) and strokes in cardiovascular (CV) diseases (Marso et al., 2016). This has lead to the popularity of marginal mean function regression models, such as those of Ghosh and Lin (2002) and Mao and Lin (2016), when randomized clinical trials involve recurrent and terminal events (Solomon et al., 2017, 2019; Vaduganathan et al., 2024).

The multistate process in Figure 1.1(e) with state space $\mathcal{S} = \{(k, D_k) : k = 0, 1, \ldots, K\}$ portrays a setting where up to K recurrent non-fatal events may occur, as well as death. Ghosh and Lin (2002) proposed a mean function model based on the expected number of non-fatal events $\mu(t) = \mathbb{E}(N(t))$ occurring up to time t, taking into account that death precludes the occurrence of further events. They proposed that in case of a binary treatment, where $X = 1$ indicates assignment to the experimental arm and $X = 0$ otherwise, the mean function takes a multiplicative form $\mu(t|X) = \mu_0(t)\exp(\beta X)$. If we let E represent the set of non-fatal events and $g(\cdot) = \log(\cdot)$, the Ghosh-Lin (GL) model can be formulated within the count framework (1.2.6) for $\mu_0(t) = \exp(\alpha(t))$.

Mao and Lin (2016) suggested a multiplicative model $\mu^\dagger(t|X) = \mu_0^\dagger(t)\exp(\beta^\dagger X)$ for the mean total number of events (fatal or non-fatal) up to time t. By incorporating fatal events of any type into the composite counting process, the traditional Mao-Lin (ML) model is motivated by contexts where treatment is designed to avoid both serious non-fatal and fatal events. In trials (II) aiming to prevent non-fatal events only (see e.g. palliative trial of Section 1.3.3), the ML model approach has less appeal. Furberg et al. (2022) adapted the ML model for the setting displayed in Figure 1.1(f), where fatal events are classified as disease-related (D_k) or non-disease related (D'_k). In this case E may address only recurrent non-fatal events or disease-related fatal events, and $\mu^\dagger(t|X)$ can be likewise formulated within the count framework (1.2.6) of Section 1.2.3.1 for $g(\cdot) = \log(\cdot)$ and $\mu_0^\dagger(t) = \exp(\alpha(t))$; non-disease related death is handled as a competing risk. This composite endpoint approach may be suited to cardiovascular (CV) trial settings (I), where treatments are typically intended to prevent occurrences of strokes, myocardial infarctions, hospitalizations and CV-related death (Marso et al., 2016).

Nonparametric estimation of the mean functions $\mu(t)$ or $\mu^\dagger(t)$ does not provide scalar treatment effects (Cook et al., 2009; Ghosh and Lin, 2000; Cortese and Scheike, 2022; Erdmann et al., 2024), unless the estimates $\widehat{\mu}(t|X)$ or $\widehat{\mu}^\dagger(t|X)$ for a given treatment group are compared at a specific time τ. The marginal GL and ML models for the expected numbers of events are thus appealing in that they provide scalar measures of treatment effects (i.e. β or $\beta^\dagger$), which can be used as a basis for trial design. However, they come with similar challenges as in cumulative incidence function regression discussed in Chapter 3: the actual mean functions are complicated functions of the true underlying intensity-based processes and do in general not satisfy the multiplicative GL or ML form, so that the GL and ML models are in general misspecified and should be considered

as "approximations" (Buja et al., 2019). Key challenges with marginal GL and ML analyses are thus (a) model misspecification and (b) interpretation of the estimands β and $\beta^{\dagger}$. Specifically, the estimands depend on the underlying dynamic event processes shown in Figures 1.1(e) and 1.1(f), the effect of treatment on the various event intensities (Cook and Lawless, 2007), and on whether the rather simple parametric relationship between $\mu(t|X=1)$ and $\mu(t|X=0)$ (or $\mu^{\dagger}(t|X=1)$ and $\mu^{\dagger}(t|X=0)$) is plausible.

The ML approach has seen increasing use in phase III trials of cardiovascular diseases. In the PARAGON-HF trial of patients with chronic heart failure (HF) and preserved ejection fraction (HFpEF), for example, the effect of sacubitril/valsartan versus valsartan alone was measured on the mean function for the composite event process of recurrent HF hospitalizations and CV death (Solomon et al., 2017, 2019). The ongoing FINEARTS-HF trial will be analyzed based on a similar primary outcome (Vaduganathan et al., 2024). Despite the widespread use of marginal mean models in randomized trials, relatively little attention has been given to the properties of treatment effect estimators $\widehat{\beta}$ and $\widehat{\beta}^{\dagger}$ in settings where the GL and ML models may be misspecified (Toenges et al., 2021).

The objectives of this chapter are to discuss the fit of mean function models and interpretability of estimands, and to make recommendations for randomized trials with disease processes like those in Figures 1.1(e) and 1.1(f). Using large sample theory we derive the limiting values of the GL and ML treatment effect estimators under some specific true processes, which we refer to as the "data-generating frameworks". We consider in particular two such frameworks: one where the true event processes are based on a broad class of multiplicative intensity-based models including Poisson and more general Markov models (Cook and Lawless, 2007, Section 6.6.1), and one where the true processes are based on random effects models (Cook and Lawless, 2007, Section 6.6.2). Within the former framework we assume the non-fatal and fatal event processes to be either independent or associated through dependencies of the fatal event intensities on the number of previous non-fatal events. We explore the determinants of the limiting values $\beta^{\star}$ and $\beta^{\dagger\star}$, and point out that they are sensitive to several features of the true processes.

The remainder of this chapter is organized as follows. In Section 4.2, we introduce notation for recurrent and terminal event processes, define rate functions, and describe the marginal Ghosh-Lin (GL) and Mao-Lin (ML) models in the context of Figures 1.1(e) and 1.1(f) in more detail. In Section 4.3.1, we review inverse probability of censoring weighted estimating functions and large

sample theory for the semiparametric GL and ML models. Alternative estimation methods based on pseudo-Poisson analyses using weighted counts or weighted Poisson estimating functions are proposed in Section 4.3.2. Section 4.3.3 compares the different estimation procedures in extensive simulation studies. In Section 4.4, we consider various intensity-based and random effects models for generating recurrent and terminal event processes, and study the factors determining the limiting values of the GL and ML treatment effect estimators under such processes. An illustrative application is given in Section 4.5; we discuss in particular a palliative trial of individuals with breast cancer metastatic to bone and demonstrate how event-related loss to follow-up can be handled in GL analyses. Concluding remarks are given in Section 4.6.

4.2 Notation and marginal models for mean functions

We first introduce notation for an arbitrary recurrent and terminal event process $\{Z(t), 0 < t\}$ with a finite state space $\mathcal{S}$ (Cook and Lawless, 2007). As before, we let $Z(t)$ represent the state occupied at time $t \geq 0$, and assume that $Z(0) = 0$ for all individuals where $t = 0$ typically corresponds to the time of randomization and where $0 \in \mathcal{S}$ denotes the initial state. We let X be a binary treatment indicator equal to 1 if an individual was randomized to the experimental treatment group and 0 otherwise. We also let U be an auxiliary scalar covariate with $\mathbb{E}(U) = 1$, $\text{Var}(U) = \phi$, and $X \perp\!\!\!\perp U$ due to randomization. If $\mathcal{H}(t) = \{Z(s) : 0 < s < t; X, U\}$ denotes the process history at time t^- including X and U, the multistate process can be fully expressed by the intensity functions

$$\lambda_{jl}(t \mid \mathcal{H}(t)) = \lim_{\Delta t \downarrow 0} \frac{P(Z(t + \Delta t^-) = l \mid Z(t^-) = j, \mathcal{H}(t))}{\Delta t} \qquad j, l \in \mathcal{S}\ , \tag{4.2.1}$$

for $(j, l) \in \{(k-1, k), (k, D_k)\}$ in the setting of Figure 1.1(e) and $(j, l) \in \{(k-1, k), (k, D_k), (k, D'_k)\}$ in the setting of Figure 1.1(f), respectively, $k = 1, \ldots, K$.

4.2.1 Marginal rate functions in recurrent event processes terminated by death

We introduce the marginal Ghosh-Lin model in the context of Figure 1.1(e), where the transient states $0, 1, \ldots, K$ represent the number of non-fatal events experienced over time and the absorbing state D_k is entered upon death by individuals who have experienced k non-fatal events. As we discuss in Section 4.4.1, the intensities $\lambda_{(k-1)k}(t|\mathcal{H}(t))$ and $\lambda_{kD_k}(t|\mathcal{H}(t))$ in (4.2.1) may depend on how many recurrent events have occurred by t^-. We let $T = \inf\{t > 0 : Z(t) \in \{D_0, \ldots, D_K\}\}$ denote the time of death, and let $\{N(t), t > 0\}$ be a right-continuous counting process so that $dN(t) = 1$ if a non-fatal event occurs at time t and 0 otherwise. Then, $N(t) = N(\min(t, T))$ denotes the number of recurrent non-fatal events up to t, allowing for the fact that death terminates the event process. That is, for all $t \geq T$, $dN(t) = 0$ and thus $N(t) = N(T)$. We let $N_D(t) = \mathbb{1}(T \leq t)$ indicate that an individual has died by time t, with associated counting process $\{N_D(t), t > 0\}$ and increment $dN_D(t) = \mathbb{1}(T = t)$.

The mean function for non-fatal events is defined as

$$\mu(t|X) = \mathbb{E}(N(t)|X) = \sum_{k=0}^{\infty} k\big(P(Z(t) = k|X) + P(Z(t) = D_k|X)\big) \ . \tag{4.2.2}$$

The mean function for the semiparametric Ghosh and Lin (2002) model takes the form

$$\mu(t|X) = \mu_0(t) \exp(\beta X) \ , \tag{4.2.3}$$

where $\mu_0(t) = \mathbb{E}(N(t)|X = 0)$ is an unspecified baseline mean function and

$$\beta = \log\left(\frac{\mu(t|X = 1)}{\mu(t|X = 0)}\right)$$

is a scalar estimand that permits causal interpretation according to the principles of Section 2.2.2. In what follows we refer to (4.2.3) as the GL model, and let $d\mu_0(t) = \mathbb{E}(dN(t)|X = 0)$.

An alternative expression for $\mu(t|X)$ in (4.2.2) is based on the conditional rate function $r(t|X)dt = \mathbb{E}(dN(t)|T \geq t, X)$ and $S(t) = P(T > t|X)$ (Cook and Lawless, 1997, 2007):

$$\mu(t|X) = \int_0^t r(u|X)S(u^-|X)du \ . \tag{4.2.4}$$

It should be noted that $r(t|X)$ is not the event intensity but rather is a partially conditional rate function, conditioning only on survival up to time t. Since it conditions on a post-randomization outcome (i.e. survival to time t), it is not suitable for causal comparisons of treatment groups, but is a component of $\mu(t|X)$, which is suitable. Assuming censoring due to loss to follow-up is independent of the recurrent event and death processes, (4.2.4) can be estimated nonparametrically for each treatment group using an Kaplan-Meier (KM) estimate for $S(u|X)$ and a Nelson-Aalen (NA) estimate for $r(u|X)du = dR(u|X)$. In applications that follow in Section 4.5, we refer to this estimator as the KM-based Cook-Lawless (CL) estimator. A consistent estimate for (4.2.2) can be obtained using Aalen-Johansen (AJ) estimates of the state occupancy probabilities $P(Z(t) = k|X)$ and $P(Z(t) = D_k|X)$ (Aalen et al., 2001). We note that AJ estimation offers an alternative to KM estimation of $S(u|X)$ in (4.2.4) through the relationship

$$P(T \leq t|X) = 1 - S(t|X) = \sum_{k=0}^{\infty} P(Z(t) = D_k|X) \; . \tag{4.2.5}$$

We refer to the resulting nonparametric estimator for $\mu(t|X)$ based on (4.2.5) as the AJ-based CL estimator. Either approach to estimation of $\mu(t|X)$ can be made robust to process-dependent loss to follow-up through the incorporation of inverse probability of censoring weights (Cook et al., 2009), as we illustrate in Section 4.5.

Cook and Lawless (2007, Section 6.6.3) note that an alternative semiparametric approach for $\mu(t|X)$ is to model the survival function using the Cox model $S(t|X) = S_0(t)^{\exp(\zeta)}$ and the rate function using the Andersen-Gill (AG) model $r(t|X) = r_0(t)\exp(\xi X)$, and then to estimate $\mu(t|X)$ using (4.2.4). Whereas the GL model requires a weighted estimation procedure, this approach can be implemented with standard software for the Cox and AG models. However, the treatment effect ξ does not have a descriptive causal interpretation, and the resulting estimates of $\mu(t|X)$ are functions of ξ, ζ and time, so we do not obtain a one-dimensional estimand. When treatment has little effect on mortality, as represented in $S(t|X)$, estimates of ξ in the AG model are often close to estimates of β from the GL model; Furberg et al. (2022) also observed this when analyzing recurrent MIs in the LEADER trial.

We next formulate the Mao-Lin model for disease processes as depicted in Figure 1.1(f).

4.2.2 Marginal rate functions with recurrent-terminal composite events and competing risks

If $T_D = \inf\{t > 0 : Z(t) \in \{D_0, \ldots, D_K\}\}$ and $T_{D'} = \inf\{t > 0 : Z(t) \in \{D'_0, \ldots, D'_K\}\}$ represent the time to disease-related and disease-unrelated death (e.g. CV death and non-CV death), the overall survival time is $T = \min(T_D, T_{D'})$ with survivor function $S(t) = P(T > t)$. Here we let $N_D(t) = N_D(\min(T, t)) = \mathbb{1}(T_D \leq \min(T, t))$ denote the counting process for disease-related death; the number of recurrent non-fatal events up to time t is given by $N(t) = N(\min(T, t))$. As discussed in Section 1.1.1, a marginal feature of possible interest in the setting of Figure 1.1(f) is the mean function for the composite event process $\{N^\dagger(t) = N(t) + N_D(t), t > 0\}$ counting the number of recurrent non-fatal events and disease-related death, which can be expressed as

$$\mu^\dagger(t|X) = \mathbb{E}(N^\dagger(t)|X) = \sum_{k=0}^{\infty} \big[k\big(P(Z(t) = k|X) + P(Z(t) = D'_k|X)\big) + (k+1)P(Z(t) = D_k|X)\big] . \tag{4.2.6}$$

If $r^\dagger(t|X)dt = \mathbb{E}(dN^\dagger(t)|T \geq t, X)$ is the partially conditional composite event rate, the CL-type expression for $\mu^\dagger(t|X)$ is

$$\mu^\dagger(t|X) = \int_0^t r^\dagger(u|X)S(u^-|X)du . \tag{4.2.7}$$

A multiplicative model that provides a one-dimensional marginal estimand $\beta^\dagger$ is

$$\mu^\dagger(t|X) = \mu_0^\dagger(t) \exp(\beta^\dagger X) , \tag{4.2.8}$$

where $\mu_0^\dagger(t) = \mathbb{E}(N^\dagger(t)|X = 0)$ is an unspecified mean function for individuals in the control group (Mao and Lin, 2016); we may refer to $\mu_0^\dagger(t)$ as a baseline mean function, and let $d\mu_0^\dagger(t) = \mathbb{E}(dN^\dagger(t)|X = 0)$. We note that the Mao-Lin (ML) model was originally introduced for composite event processes consisting of recurrent non-fatal events and all-cause death, i.e. $\{N(t) + \mathbb{1}(T \leq t), t > 0\}$. Furberg et al. (2022) extended the ML approach to deal with competing causes of death, when one cause of death is part of the composite event process (e.g. CV death) and the other is not (e.g. non-CV death). Although we focus on this in investigations below, we refer to (4.2.8) as the ML model.

Toenges et al. (2021) observed that the estimand $\xi^\dagger$ in a multiplicative AG-type model $r^\dagger(t|X) = r_0^\dagger(t)\exp(\xi^\dagger X)$, where $r_0^\dagger(t) = r^\dagger(t|X=0)$, is often close to $\beta^\dagger$ in a ML model when the mortality rate is low over the course of the study. This is also the case when the $\lambda_{kD_k}(t|\mathcal{H}(t))$ intensities do not depend on the recurrent event process. As in the case of recurrent events, the estimand $\xi^\dagger$ does however not have a marginal causal interpretation and should not, despite Toenges et al. (2021)'s recommendation, be used for primary treatment comparisons.

4.3 Mean function regression under misspecification

Here we consider estimation of $(\beta, d\mu_0(\cdot))'$ in the GL model (4.2.3) and $(\beta^\dagger, d\mu_0^\dagger(\cdot))'$ in the ML model (4.2.8) using three different methods. We first review the ordinary inverse probability of censoring weighted estimating functions and large sample theory for the semiparametric GL and ML models in Section 4.3.1 (Ghosh and Lin, 2002; Mao and Lin, 2016; Furberg et al., 2022). In Section 4.3.2 we propose two alternative procedures based on pseudo-Poisson analyses using weighted event counts or weighted Poisson estimating functions. The former is similar in spirit to the use of direct binomial estimation for cumulative incidence function regression in semi-competing risks processes (Scheike et al., 2008); see Section 3.3.1.2. Using large sample theory we derive the probability limits $\beta^\star$ (or $\beta^{\dagger\star}$) of the associated treatment effect estimators of β (or $\beta^\dagger$) when model (4.2.3) (or (4.2.8)) is misspecified.

For notational convenience and to harmonize the presentation we let $\beta^m \in \{\beta, \beta^\dagger\}$ denote the target estimand under the marginal model $m \in \{GL, ML\}$, and let $N_i^m(t) \in \{N_i(t), N_i^\dagger(t)\}$ represent the corresponding counting process with $\mu^m(t|X_i) = \mathbb{E}(N_i^m(t)|X_i) \in \{\mu(t|X_i), \mu^\dagger(t|X_i)\}$ and $r^m(t|X_i) = \mathbb{E}(dN_i^m(t)|T_i \geq t, X_i) \in \{r(t|X_i), r^\dagger(t|X_i)\}$. Hence, if $m = GL$, then $\beta^m = \beta, N_i^m(t) = N_i(t), dN_i^m(t) = dN_i(t), \mu^m(t|X_i) = \mu(t|X_i)$ and $r^m(t|X_i) = r(t|X_i)$. Otherwise, if $m = ML$, $\beta^m = \beta^\dagger, N_i^m(t) = N_i^\dagger(t), dN_i^m(t) = dN_i^\dagger(t), \mu^m(t|X_i) = \mu^\dagger(t|X_i)$ and $r^m(t|X_i) = r^\dagger(t|X_i)$, respectively.

Suppose individual i in a sample of n independent processes is observed over the time period $(0, \min(T_i, C_{ri}, \tau)]$, where τ denotes the administrative censoring time and C_{ri} is a random right-censoring time at which premature loss to follow-up may occur. Individuals may have in general different administrative censoring times τ_i, but for simplicity we consider the case where $\tau_i = \tau$.

We assume $\{Z_i(u), u > 0\} \perp\!\!\!\perp C_{ri}|X_i$, and let $C_i = \min(C_{ri}, \tau)$ be the net censoring time for i, $i = 1, \ldots, n$.

4.3.1 Ghosh-Lin and Mao-Lin methods of estimation

The inverse probability of censoring weighted estimating equations are

$$\sum_{i=1}^{n} \widehat{w}_i(t)\big(\mathbb{1}(T_i \geq t)dN_i^m(t) - d\mu_0^m(t)\exp(\beta^m X_i)\big) = 0 \tag{4.3.1}$$

for the baseline rate function $d\mu_0^m(\cdot)$, and

$$\sum_{i=1}^{n} \int_0^{\infty} \widehat{w}_i(t)\big(\mathbb{1}(T_i \geq t)dN_i^m(t) - d\mu_0^m(t)\exp(\beta^m X_i)\big)X_i = 0 \tag{4.3.2}$$

for the regression coefficient β^m (Ghosh and Lin, 2002; Mao and Lin, 2016), with

$$\widehat{w}_i(t) = \frac{\mathbb{1}(C_i \geq \min(T_i, t))}{\widehat{G}_i(\min(T_i, t))} = \mathbb{1}(t \leq \tau) \cdot \frac{\mathbb{1}(C_{ri} \geq \min(T_i, t))}{\widehat{G}_i(\min(T_i, t))} \tag{4.3.3}$$

an estimated weight function depending on an estimate of the random censoring distribution $G_i(u) = G(u|X_i) = P(C_{ri} > u|X_i)$. A nonparametric Kaplan-Meier (KM) estimate $\widehat{G}_i(u) = \widehat{G}(u) = \widehat{P}(C_r > u)$ of $G_i(u) = G(u) = P(C_r > u)$ can be used if the censoring process does not depend on X. In settings where censoring depends on X, stratified KM estimation or intensity-based regression models for C_r (e.g. the Cox (1972) proportional hazards model or the Aalen (1980) additive model) are commonly used to obtain $\widehat{G}_i(u) = \widehat{G}(u|X_i) = \widehat{P}(C_{ri} > u|X_i)$. In what follows we label the estimator of $G_i(u)$ as $\widehat{G}_i(u)$, with $\widehat{G}_i(u)$ either representing $\widehat{G}(u)$ in the former case of unstratified KM estimation or $\widehat{G}(u|X_i)$ in the latter case of stratified KM estimation or Cox regression.

Solving (4.3.1) for $d\mu_0^m(t)$ with fixed β^m gives the profile Breslow-type estimator

$$d\widehat{\mu}_0^m(t) = \sum_{i=1}^{n} \widehat{w}_i(t)\mathbb{1}(T_i \geq t)dN_i^m(t) \Big/ \sum_{i=1}^{n} \widehat{w}_i(t)\exp(\beta^m X_i), \quad t > 0\ , \tag{4.3.4}$$

which in turn can be substituted into (4.3.2) to obtain the partial pseudo-score function for β^m

$$U(\beta^m, \widehat{G}) = \sum_{i=1}^{n} \int_0^{\infty} \widehat{w}_i(t)\left[X_i - \frac{S^{(1)}(t, \beta^m)}{S^{(0)}(t, \beta^m)}\right]\mathbb{1}(T_i \geq t)dN_i^m(t), \tag{4.3.5}$$

where $S^{(l)}(t,\beta^m) = \sum_{i=1}^n \widehat{w}_i(t) X_i^l \exp(\beta^m X_i)$ for $l = 0, 1$. Expression (4.3.5) can be set to zero and solved to get $\widehat{\beta}^m$. Ghosh and Lin (2002) and Mao and Lin (2016) showed for their respective models that $\widehat{\beta}^m$ is consistent for β^m if

A1 $\{Z(u), u > 0\} \perp\!\!\!\perp C_r \mid X$,

A2 model (4.2.3) or (4.2.8) for $\mu^m(t|X) = \mathbb{E}(N^m(t)|X)$ is valid ,

A3 the censoring model for C_r is correctly specified such that $\widehat{G} \longrightarrow G$,

and if certain regularity conditions are satisfied. Under the same assumptions the Breslow-type estimator for the baseline mean function

$$\widehat{\mu}_0^m(t|X) = \int_0^t d\widehat{\mu}_0^m(u) = \int_0^t \frac{\sum_{i=1}^n \widehat{w}_i(u) \mathbb{1}(T_i \geq t) dN_i^m(u)}{\sum_{i=1}^n \widehat{w}_i(u) \exp(\widehat{\beta}^m X_i)}$$

is consistent for $\mu_0^m(t|X)$; $d\widehat{\mu}_0^m(u)$ can be obtained by substituting $\widehat{\beta}^m$ into (4.3.1).

The censoring model for C_r may however be subject to some form of misspecification, in which case assumption A3 no longer holds. Appendix B.1 reviews the limiting behavior of the KM and Cox estimators of the censoring survivor function under potential misspecification. Following the notation of Appendix B.1, we let $G_i^\star(u)$ be the probability limit of the "working" estimator $\widehat{G}_i(u)$, i.e. $\widehat{G}_i(u) \longrightarrow G_i^\star(u)$ (or in short $\widehat{G} \longrightarrow G^\star$); note that $G^\star = G$ when the "working" model is correctly specified (c.f. A3). Recall from Section 3.3.1 that $G_i^\star(u)$ represents $G^\star(u)$ if unstratified KM methods are used for estimation and $G^\star(u|X_i)$ if a stratified KM or Cox estimation is used, respectively. Consequently, under a misspecified censoring model,

$$\widehat{w}_i(t) \longrightarrow w_i^*(t) = \mathbb{1}(t \leq \tau) \cdot \frac{\mathbb{1}(C_{ri} \geq \min(T_i, t))}{G_i^\star(\min(T_i, t))} ,$$

and we rewrite (4.3.5) as

$$U(\beta^m, G^\star) = \sum_{i=1}^n \int_0^\infty w_i^\star(t) \left[X_i - \frac{S^{(1,\star)}(t,\beta^m)}{S^{(0,\star)}(t,\beta^m)} \right] \mathbb{1}(T_i \geq t) dN_i^m(t) , \tag{4.3.6}$$

with $S^{(l)}(t,\beta^m)$ updated to $S^{(l,\star)}(t,\beta^m) = \sum_{i=1}^n w_i^\star(t) X_i^l \exp(\beta^m X_i)$ for $l = 0, 1$. If A1 holds but A2 and A3 may not, $\widehat{\beta}^m$ is consistent for the limiting value $\beta^{m\star}$, the solution to

$$\mathbb{E}\big(U(\beta^m, G^\star)\big) = \int_0^\infty \left\{ s^{(1,m,\star)}(t) - \frac{s^{(1,\star)}(t,\beta^m)}{s^{(0,\star)}(t,\beta^m)} s^{(0,m,\star)}(t) \right\} dt = 0 , \tag{4.3.7}$$

where $s^{(l,\star)}(t,\beta^m) = \mathbb{E}(S^{(l,\star)}(t,\beta^m))$ and $s^{(l,m,\star)}(t) = \mathbb{E}\big(w_i^\star(t)X_i^l \mathbb{1}(T_i \geq t) dN_i^m(t)\big)$, and where expectations are taken with respect to the true recurrent and terminal event, censoring and covariate processes (Struthers and Kalbfleisch, 1986). In numerical investigations that follow in Sections 4.4.1 and 4.4.2 we denote the limiting values of $\widehat{\beta}$ and $\widehat{\beta}^\dagger$ under the GL and ML estimation procedures as $\beta_{GL}^\star$ and $\beta_{ML}^{\dagger\star}$, respectively.

4.3.2 Estimation based on weighted pseudo-Poisson estimating functions

In the spirit of the DB estimation procedure used for cumulative incidence function regression models (Scheike and Zhang, 2007; Scheike et al., 2008), here we consider direct use of weighted Poisson estimating equations for regression models in the recurrent and terminal event settings of Figures 1.1(e) and 1.1(f). Primary interest lies in estimating $\mu^m(t|X) = \mu_0^m(t)\exp(\beta^m X)$, and let $\alpha^m(t) = \log(\mu_0^m(t))$. In generalized linear model (GLM) form,

$$\log(\mu^m(t|X)) = \alpha^m(t) + \beta^m X \ . \tag{4.3.8}$$

At any fixed time t the event count $N_i^m(t)$ follows a "working" Poisson distribution with mean $\mathbb{E}(N_i^m(t)|X_i) = \mu^m(t|X_i)$. We consider R time points $0 < s_1 < s_2 < \ldots < s_R \leq \tau$ at which estimation is to be conducted, and let $\theta^m = ((\alpha^m)', \beta^m)'$ be an $(R+1)\times 1$ vector of unknown parameter values with entries $\alpha^m = (\alpha_1^m, \alpha_2^m, \ldots, \alpha_R^m)$ with $\alpha_r^m = \alpha^m(s_r), r = 1,2,\ldots,R$. In the case of uncensored data $N_i(s_r)$ equals k if $Z_i(s_r) \in \{k, D_k\}$, and $N_i^\dagger(s_r)$ is equal to k if $Z_i(s_r) \in \{k, D_k'\}$ and $k+1$ if $Z_i(s_r) = D_k$. A pseudo-Poisson log-likelihood function for a random sample of n individuals under model m is then given by

$$\ell(\theta^m) = \sum_{i=1}^{n}\sum_{r=1}^{R}\big\{N_i^m(s_r)\log\big(h(\alpha_r^m + \beta^m X_i)\big) - h(\alpha_r^m + \beta^m X_i)\big\} \ , \tag{4.3.9}$$

where for convenience we denote $h(u) = \exp(u)$. The corresponding score functions for α_r^m and β^m are

$$\frac{\partial \ell(\theta^m)}{\partial \alpha_r^m} = \sum_{i=1}^{n}\frac{h'(\alpha_r^m + \beta^m X_i)}{h(\alpha_r^m + \beta^m X_i)}\big(N_i^m(s_r) - h(\alpha_r^m + \beta^m X_i)\big) \ , \qquad r = 1,2,\ldots,R \ , \tag{4.3.10}$$

$$\frac{\partial \ell(\theta^m)}{\partial \beta^m} = \sum_{i=1}^{n}\sum_{r=1}^{R} X_i\frac{h'(\alpha_r^m + \beta^m X_i)}{h(\alpha_r^m + \beta^m X_i)}\big(N_i^m(s_r) - h(\alpha_r^m + \beta^m X_i)\big) \ . \tag{4.3.11}$$

Setting (4.3.10)-(4.3.11) equal to zero gives an estimator of θ^m.

When data is subject to right-censoring the counting process $\{N_i^m(t), 0 > t\}$ is incompletely observed for individuals who withdraw from study at a random time $C_{ri} < \min(T_i, \tau)$ prior to death or administrative censoring. For those individuals $N_i^m(s_r)$ is known if $s_r \leq C_{ri}$ but is unknown for all time points $s_r \in (C_{ri}, \tau]$, making the estimating functions (4.3.10) and (4.3.11) unusable. Thus, we consider weighted estimation designed to address this.

4.3.2.1 Weighted estimating equations

We propose to base estimation and inference for θ^m in models (4.3.8) on sums of weighted Poisson score functions of the form

$$U(\alpha_r^m, \widehat{G}) = \sum_{i=1}^{n} \widehat{w}_i(s_r) \frac{h'(\alpha_r^m + \beta^m X_i)}{h(\alpha_r^m + \beta^m X_i)} \big(N_i^m(s_r) - h(\alpha_r^m + \beta^m X_i)\big) \ , \tag{4.3.12}$$

$$U(\beta^m, \widehat{G}) = \sum_{i=1}^{n} \sum_{r=1}^{R} X_i \widehat{w}_i(s_r) \frac{h'(\alpha_r^m + \beta^m X_i)}{h(\alpha_r^m + \beta^m X_i)} \big(N_i^m(s_r) - h(\alpha_r^m + \beta^m X_i)\big) \ , \tag{4.3.13}$$

with the estimated weight function $\widehat{w}_i(s_r)$ defined as in (4.3.3). This corresponds to reweighting the contributions from individuals for whom $N_i^m(s_r)$ is known. In matrix notation, (4.3.12) and (4.3.13) can be written as

$$U^P(\theta^m, \widehat{G}) = \sum_{i=1}^{n} \begin{pmatrix} U_i(\alpha^m, \widehat{G}) \\ U_i(\beta^m, \widehat{G}) \end{pmatrix} = \sum_{i=1}^{n} \widehat{W}_i D_i A_i^{-1} (N_i^m - \mu_i^m) \ , \tag{4.3.14}$$

where $N_i^m = (N_i^m(s_1), N_i^m(s_2), \ldots, N_i^m(s_R))'$ is a $R \times 1$ vector of event counts for individual i, $\widehat{W}_i = \text{diag}(\widehat{w}_i(s_r), r = 1, \ldots, R)$ is a $R \times R$ diagonal weight matrix, $A_i = \text{diag}(\mu^m(s_r|X_i), r = 1, \ldots, R)$ is a $R \times R$ working covariance matrix, $\mu_i^m = (\mu_i^m(s_1), \mu_i^m(s_2), \ldots, \mu_i^m(s_R))'$ is an $R \times 1$ vector with entries

$$\mu_i^m(s_r) = \mathbb{E}(N_i^m(s_r)|X_i; \theta^m) = \mu^m(s_r|X_i; \theta^m) = h(\alpha_r^m + \beta^m X_i) \ ,$$

and $D_i = (\partial \mu_i^m / \partial \alpha^{m\prime}, \partial \mu_i^m / \partial \beta)'$ is an $((R+1) \times R)$ matrix of derivatives. Setting (4.3.14) equal to zero and solving gives $\widehat{\theta}^m = ((\widehat{\alpha}^m)', \widehat{\beta}^m)'$. If assumptions A1-A3 of Section 4.3.1 are satisfied, i.e. when $\{Z(t), 0 < t\} \perp\!\!\!\perp C_r|X$, model (4.3.8) for $\mathbb{E}(N^m(t)|X)$ is valid and $\widehat{G} \longrightarrow G$,

then $\widehat{\theta}^m$ can be shown to be consistent for θ^m. More generally if A1 holds but A2 and A3 may not the solution is consistent for the probability limit

$$\theta^{m\star} = (\alpha_1^{m\star}, \ldots, \alpha_R^{m\star}, \beta_P^{m\star}) \ ,$$

defined as the solution to the equation system (White, 1982)

$$\mathbb{E}(U^P(\theta^m, G^\star)) = (0, 0 \ldots, 0)' \ ,$$

where

$$U^P(\theta^m, G^\star) = \sum_{i=1}^{n} \begin{pmatrix} U_i(\alpha^m, G^\star) \\ U_i(\beta^m, G^\star) \end{pmatrix} = \sum_{i=1}^{n} W_i^\star D_i A_i^{-1} (N_i^m - \mu_i^m) \ , \tag{4.3.15}$$

and $W_i^\star = \text{diag}(w_i^\star(s_r), r = 1, \ldots, R)$ with

$$w_i^\star(s_r) = \mathbb{1}(C_{ri} > \min(T_i, s_r))/G_i^\star(\min(T_i, s_r)) \ .$$

In (4.3.15) the expectations are taken with respect to the true recurrent and terminal event, censoring and covariate processes. In investigations that follow we refer to the procedure based on (4.3.14) as pseudo-Poisson estimation and label the resulting limiting values of $\widehat{\beta}$ and $\widehat{\beta}^\dagger$ as $\beta_P^\star$ and $\beta_P^{\dagger\star}$, respectively.

4.3.2.2 Weighted event counts

An alternative approach is to base estimation and inference for α_r^m and β^m in models (4.3.8) on a composite pseudo-Poisson log-likelihood of the form

$$\ell^{\widetilde{P}}(\theta^m) \propto \sum_{i=1}^{n} \sum_{r=1}^{R} \left[\widetilde{N}_i^m(s_r) \log(h(\alpha_r^m + \beta^m X_i)) - h(\alpha_r^m + \beta^m X_i) \right] \ ,$$

where, following estimation of the random censoring distribution $G_i(\cdot)$, the weighted event count

$$\widetilde{N}_i^m(s_r) = \frac{\mathbb{1}(C_i \geq \min(T_i, s_r)) N_i^m(s_r)}{\widehat{G}_i(\min(T_i, s_r))} \overset{s_r \leq \tau}{=} \frac{\mathbb{1}(C_{ri} \geq \min(T_i, s_r)) N_i^m(s_r)}{\widehat{G}_i(\min(T_i, s_r))} = \widehat{w}_i(s_r) N_i^m(s_r)$$

can be computed for all individuals at any time s_r; $\widetilde{N}_i^m(s_r)$ is in particular 0 for those individuals who were lost to follow-up prior to time s_r. It can be shown that $\mathbb{E}(\widetilde{N}_i^m(s_r)) = \mu^m(s_r|X_i)$ if assumptions A1-A3 hold. The resulting estimating functions for α_r^m and β^m are given by (4.3.10) and (4.3.11) with $N_i^m(s_r)$ replaced by $\widetilde{N}_i^m(s_r)$, and can be written in matrix notation as

$$U^{\widetilde{P}}(\theta^m, \widehat{G}) = \sum_{i=1}^{n} D_i A_i^{-1}(\widetilde{N}_i^m - \mu_i^m) \ , \tag{4.3.16}$$

where $\widetilde{N}_i^m = (\widetilde{N}_i^m(s_1), \widetilde{N}_i^m(s_2), \ldots, \widetilde{N}_i^m(s_R))'$ is a $R \times 1$ vector of weighted event counts for individual i.

We note that (4.3.16) differs from (4.3.14) of Section 4.3.2.1 in how individuals who were lost to follow-up prior to time s_r (i.e. $C_{ri} < \min(s_r, T_i)$) contribute to the score functions for α_r^m and β^m; see also Blanche et al. (2023). In the weighted outcome (wO) or event count approach the contributions from such an individual i at s_r are $(0 - h(\alpha_r^m + \beta^m X_i))$ for the baseline function α_r^m, and $X_i(0 - h(\alpha_r^m + \beta^m X_i))$ for the regression coefficient β^m. In the weighted estimating equation (wEE) approach, i does not contribute to (4.3.12) and (4.3.13) since $\widehat{w}_i(s_r) = 0$ for all $s_r > C_{ri}$. This is similar to the approach used with the GL and ML estimation methods of Section 4.3.1. The limiting values are the same for both pseudo-Poisson procedures when the censoring model is correctly specified.

4.3.3 Empirical studies of finite sample properties

Simulation studies were conducted to assess the finite sample properties of estimators $\widehat{\beta}$ obtained by solving the pseudo-Poisson score functions (4.3.14) and (4.3.16) for the Ghosh-Lin (GL) model in the case in which it is a correctly specified mean function. For the remainder of this section we denote the corresponding estimators by $\widehat{\beta}_{wEE}$ and $\widehat{\beta}_{wO}$, respectively. We first describe the simulation setup in Section 4.3.3.1, following which we summarize the simulation results in Section 4.3.3.2.

4.3.3.1 Simulation setup

Recurrent and terminal event data under the GL model (4.2.3) were simulated as follows (Ghosh and Lin, 2002): we first generate X from a binomial distribution with probability 0.5 an individual is in the experimental treatment group. We let $\tau = 1$, and assume that the non-fatal and fatal events occur over the interval $(0, \tau]$ according to the process in Figure 1.1(e) with proportional intensities $\lambda_{(k-1)k}(t|\mathcal{H}(t)) = \lambda(t|X) = \lambda_0(t)\exp(\gamma X)$ and $\lambda_{kD_k}(t|\mathcal{H}(t)) = \rho(t|X) = \rho_0(t)\exp(\delta X)$. We note that $r(t|X) = \lambda(t|X)$ and $S(t|X) = S(t|X=0)^{\exp(\delta X)} = \left(\exp(-\int_0^t \rho_0(s)ds)\right)^{\exp(\delta X)}$ under this true data-generating process, and consider two different cases: $\delta = 0$ and $\delta \neq 0$. If $\delta = 0$ (i.e. no treatment effect on death intensities), the cumulative mean function for non-fatal events

$$\mu(t|X) \overset{(4.2.4)}{=} e^{\gamma X} \cdot \Bigg[\underbrace{\int_0^t \lambda_0(u)\exp\left(-\int_0^u \rho_0(s)ds\right)du}_{=\mu_0(t)} \Bigg]$$

is of Ghosh-Lin form for $\beta = \gamma$. If $\delta \neq 0$ (i.e. treatment affects death intensities), we let $r(t|X=1) = r(t|X=0)S(t|X=0)^{1-\exp(\delta)}\exp(\xi)$ in which case

$$\mu(t|X=1) \overset{(4.2.4)}{=} \int_0^t r(u|X=1)S(u|X=1)du = e^{\xi} \cdot \Bigg[\underbrace{\int_0^t r(u|X=0)S(u|X=0)du}_{=\mu_0(t)} \Bigg],$$

and so the proportional means assumption inherent to the GL model is satisfied for $\beta = \xi = \gamma$. This is obviously not a plausible model for actual settings but it provides a framework to check on the properties of estimators arising from the pseudo-Poisson procedures.

We consider time-homogeneous baseline intensities and set λ_0, ρ_0 so that (1) $\pi_D = P(T \leq \tau|X=0) = 0.15$ or 0.3 and (2) $\mathbb{E}(N(\tau)|X=0) = 0.8$. The effects of treatment on the non-fatal and fatal intensities were set to $\exp(\gamma) = 0.75$ and $\exp(\delta) = 1$ or 0.8, respectively. An early withdrawal time $C_r|X$ was generated from an exponential distribution with rate parameter $\lambda_0^r e^{\gamma_r X}$; $\lambda_0^r > 0$ was set to satisfy $\pi_r = P(C_r \leq \tau|X=0) = 0.2, 0.4$ or 0.6, i.e. $\lambda_0^r = -\log[1-\pi_r]/\tau$. We further set $\exp(\gamma_r) = 1$ or 0.85. We simulated data sets with $n = 500$ individuals, and for each data set obtained estimates of β in model (4.2.3) using the ordinary GL estimation method of Section 4.3.1 and by the pseudo-Poisson estimation methods based on $R = 6$ or $R = 12$

equi-spaced time points in $(0, \tau)$. The GL procedure can be implemented using the `R` function `coxph()` from the `survival` package (Therneau, 2024), with a robust variance estimate and $\widehat{w}_i(t)$ accommodated by the `weights` argument. The estimates $\widehat{\beta}_{wEE}$ and $\widehat{\beta}_{wO}$ can be obtained using the `geese()` function in the `R` package `geepack` (Halekoh et al., 2006), once a suitable dataframe has been constructed. In Table C.2 of Appendix C.1 a dataframe is provided involving hypothetical data from four individuals; it contains the variables necessary for estimation based on (4.3.14) and (4.3.16) using the `geese()` function (Halekoh et al., 2006). Estimates of the censoring distribution $G_i(t)$ for calculation of the weights (4.3.3) were obtained by different nonparametric and semiparametric approaches, including unstratified Kaplan-Meier (KM) estimation (when $\exp(\gamma_r) = 1$), stratified KM estimation and Cox regression (when $\exp(\gamma_r) = 1$ or 0.85). Standard errors provided from such analyses treat the weights as known rather than estimated. Despite this, we find in the simulation studies that follow that they can perform well in many settings. Resampling methods such as the nonparametric bootstrap can be used in settings where there may be concerns about the effect of sampling variation of the estimated weights. Derivation of robust variance estimates for the Poisson-based approaches that recognize the sampling variation of the weights are a topic of future research. For each of the three estimation procedures we report on the empirical bias $\widehat{\beta}$ (BIAS), the empirical standard error (ESE), the average standard error (ASE) provided by `coxph()` or `geese()`, the mean squared error (MSE), and the empirical coverage probability (CP) for β of the nominal 95% Wald confidence intervals.

4.3.3.2 Simulation results

Simulation results are summarized in Tables 4.1-4.4 for $n_{sim} = 2000$ simulation runs and for the case where the adopted censoring model is correctly specified (i.e. $\widehat{G} \longrightarrow G, G^{\star} = G$). Results for settings where the censoring process is independent of X (i.e. $\exp(\gamma_r) = 1$) are provided in Tables 4.1 and 4.2; Tables 4.3 and 4.4 represent simulation results when censoring depends on X (i.e. $\exp(\gamma_r) = 0.85$).

The pseudo-Poisson procedure based on weighted estimating equations, denoted by P_{wEE}, yields estimates with negligible empirical biases in all scenarios considered. There is good agreement between the empirical and average standard errors, and the empirical coverage probabilities are close to the desired nominal level. Variability in $\widehat{\beta}_{wEE}$ increases as the censoring rate π_r increases

and as R decreases from 12 to 6. Note that for the pseudo-Poisson procedure based on weighted event counts, denoted by P_{wO}, the average standard error is larger than the empirical standard error when the censoring rates are high. The empirical coverage probability is slightly inflated in such settings (e.g. when $\pi_r = 0.6$ in Tables 4.1-4.4).

Blanche et al. (2023) compared the relative efficiency of these two weighting procedures in logistic cumulative incidence function regression when estimation is based only on a single time point s_1 (i.e. $R = 1$). Similar investigations for the pseudo-Poisson procedures in recurrent and terminal event settings are a topic of future research.

Overall the GL procedure of Section 4.3.1 is most efficient among all procedures. Similar findings were found by Scheike et al. (2008) when comparing the estimation procedure of Fine and Gray (1999) with binomial estimation in cumulative incidence function regression. Note that for the GL procedure there is good agreement between the average standard error and the empirical standard error even when censoring rates are high. This indicates that the robust variance estimates provided by `coxph()` are valid and can be used in practice; we report such an estimate in the application of Section 4.5.

4.4 Limiting values of estimator of treatment effect under particular data-generating mechanisms

The marginal Ghosh and Lin (2002) and Mao and Lin (2016) models make a strong assumption about the form of any difference in $\mu^m(t|X = 1)$ and $\mu^m(t|X = 0)$. The actual mean functions in (4.2.2) and (4.2.6) are complicated functions of the underlying true process intensities, and may in general not satisfy the GL or ML form. In this section we consider some specific "data-generating frameworks" and evaluate the effects of the true process intensities on the adequacy of the GL and ML models. We also study the limiting values $\beta^{m\star}$ of estimators of treatment effect arising from the GL/ML estimation procedure (Section 4.3.1) and the Poisson procedure based on weighted estimating equations (Section 4.3.2.1) under such true processes. The numerical investigations concerning the GL model are patterned on the multistate process in Figure 1.1(e); the process displayed in Figure 1.1(f) and patterned on the LEADER trial (Marso et al., 2016) serves as a basis for discussions of the ML model in Section 4.4.2.

			$\pi_r = 0.2$					$\pi_r = 0.4$					$\pi_r = 0.6$				
π_D	Method	R	BIAS	ESE	ASE	MSE	CP	BIAS	ESE	ASE	MSE	CP	BIAS	ESE	ASE	MSE	CP
unstratified KM estimation: $\exp(\gamma) = 0.75,\ \exp(\delta) = 1$ and $\exp(\gamma_r) = 1$																	
0.15	GL		-0.0001	0.1149	0.1163	0.0132	0.9560	-0.0047	0.1272	0.1251	0.0162	0.9540	-0.0053	0.1425	0.1395	0.0203	0.9530
	P_{wEE}	6	-0.0012	0.1342	0.1366	0.0180	0.9490	-0.0060	0.1493	0.1466	0.0223	0.9495	-0.0068	0.1605	0.1617	0.0258	0.9555
		12	-0.0006	0.1317	0.1343	0.0173	0.9545	-0.0057	0.1473	0.1443	0.0217	0.9515	-0.0070	0.1579	0.1596	0.0250	0.9560
	P_{wO}	6	-0.0011	0.1372	0.1402	0.0188	0.9510	-0.0079	0.1569	0.1546	0.0247	0.9485	-0.0068	0.1761	0.1763	0.0310	0.9520
		12	-0.0005	0.1350	0.1380	0.0182	0.9515	-0.0076	0.1552	0.1527	0.0241	0.9470	-0.0069	0.1741	0.1749	0.0303	0.9520
0.3	GL		-0.0017	0.1226	0.1183	0.0150	0.9395	0.0013	0.1310	0.1271	0.0172	0.9470	0.0017	0.1468	0.1409	0.0216	0.9420
	P_{wEE}	6	-0.0045	0.1411	0.1360	0.0199	0.9430	0.0012	0.1486	0.1456	0.0221	0.9470	0.0024	0.1618	0.1599	0.0262	0.9515
		12	-0.0043	0.1388	0.1340	0.0193	0.9405	0.0015	0.1464	0.1436	0.0214	0.9455	0.0024	0.1605	0.1581	0.0257	0.9515
	P_{wO}	6	-0.0040	0.1441	0.1396	0.0208	0.9390	0.0009	0.1571	0.1537	0.0247	0.9415	0.0033	0.1780	0.1747	0.0317	0.9470
		12	-0.0037	0.1418	0.1377	0.0201	0.9395	0.0011	0.1555	0.1521	0.0242	0.9420	0.0034	0.1776	0.1736	0.0315	0.9475
stratified KM estimation: $\exp(\gamma) = 0.75,\ \exp(\delta) = 1$ and $\exp(\gamma_r) = 1$																	
0.15	GL		0.0031	0.1192	0.1163	0.0142	0.9420	-0.0004	0.1270	0.1248	0.0161	0.9475	-0.0003	0.1372	0.1397	0.0188	0.9520
	P_{wEE}	6	-0.0002	0.1388	0.1364	0.0193	0.9560	-0.0030	0.1494	0.1457	0.0223	0.9450	-0.0026	0.1646	0.1619	0.0271	0.9435
		12	0.0004	0.1365	0.1341	0.0186	0.9520	-0.0025	0.1477	0.1435	0.0218	0.9465	-0.0021	0.1618	0.1598	0.0262	0.9445
	P_{wO}	6	-0.0002	0.1388	0.1399	0.0193	0.9610	-0.0030	0.1494	0.1538	0.0223	0.9575	-0.0026	0.1646	0.1766	0.0271	0.9575
		12	0.0004	0.1365	0.1378	0.0186	0.9605	-0.0025	0.1477	0.1520	0.0218	0.9555	-0.0021	0.1618	0.1752	0.0262	0.9635
0.3	GL		0.0016	0.1195	0.1184	0.0143	0.9495	-0.0031	0.1288	0.1269	0.0166	0.9485	-0.0039	0.1395	0.1410	0.0195	0.9500
	P_{wEE}	6	-0.0011	0.1362	0.1360	0.0185	0.9510	-0.0011	0.1473	0.1452	0.0217	0.9470	0.0008	0.1583	0.1602	0.0250	0.9510
		12	-0.0009	0.1341	0.1340	0.0180	0.9530	-0.0011	0.1453	0.1432	0.0211	0.9490	0.0003	0.1566	0.1584	0.0245	0.9485
	P_{wO}	6	-0.0011	0.1362	0.1395	0.0185	0.9555	-0.0011	0.1473	0.1533	0.0217	0.9605	0.0008	0.1583	0.1750	0.0250	0.9690
		12	-0.0009	0.1341	0.1377	0.0180	0.9585	-0.0011	0.1453	0.1518	0.0211	0.9610	0.0003	0.1567	0.1739	0.0245	0.9685
Cox regression: $\exp(\gamma) = 0.75,\ \exp(\delta) = 1$ and $\exp(\gamma_r) = 1$																	
0.15	GL		-0.0001	0.1149	0.1163	0.0132	0.9555	-0.0047	0.1272	0.1252	0.0162	0.9540	-0.0053	0.1425	0.1395	0.0203	0.9545
	P_{wEE}	6	-0.0011	0.1342	0.1366	0.0180	0.9495	-0.0060	0.1493	0.1466	0.0223	0.9495	-0.0069	0.1604	0.1617	0.0258	0.9550
		12	-0.0006	0.1318	0.1343	0.0174	0.9545	-0.0057	0.1473	0.1443	0.0217	0.9510	-0.0070	0.1578	0.1596	0.0249	0.9555
	P_{wO}	6	-0.0013	0.1349	0.1402	0.0182	0.9560	-0.0065	0.1499	0.1546	0.0225	0.9590	-0.0072	0.1619	0.1763	0.0262	0.9705
		12	-0.0008	0.1324	0.1380	0.0175	0.9610	-0.0062	0.1478	0.1527	0.0219	0.9600	-0.0073	0.1588	0.1749	0.0253	0.9715
0.3	GL		-0.0018	0.1226	0.1183	0.0150	0.9385	0.0013	0.1310	0.1271	0.0172	0.9475	0.0016	0.1468	0.1410	0.0215	0.9425
	P_{wEE}	6	-0.0046	0.1411	0.1360	0.0199	0.9430	0.0013	0.1485	0.1456	0.0220	0.9465	0.0022	0.1617	0.1599	0.0261	0.9525
		12	-0.0044	0.1388	0.1340	0.0193	0.9385	0.0015	0.1464	0.1436	0.0214	0.9470	0.0022	0.1603	0.1581	0.0257	0.9530
	P_{wO}	6	-0.0047	0.1417	0.1396	0.0201	0.9450	0.0010	0.1498	0.1537	0.0224	0.9565	0.0019	0.1633	0.1747	0.0267	0.9640
		12	-0.0044	0.1393	0.1377	0.0194	0.9480	0.0012	0.1474	0.1521	0.0217	0.9560	0.0020	0.1617	0.1736	0.0261	0.9670

Table 4.1: Empirical properties of treatment effect estimators $\widehat{\beta}$ from a GL model fitted under different estimation procedures; $\tau = 1, \mathbb{E}(N(\tau)|X = 0) = 0.8, \pi_D = P(T \leq \tau|X = 0) = 0.15$ and $0.3, \pi_r = P(C_r \leq \tau|X = 0) = 0.2, 0.4$ and $0.6, \exp(\gamma) = 0.75, \exp(\delta) = 1, \exp(\gamma_r) = 1, \widehat{G} \longrightarrow G, n = 500$ individuals, $n_{sim} = 2000$ simulation runs.

			$\pi_r = 0.2$					$\pi_r = 0.4$					$\pi_r = 0.6$				
π_D	Method	R	BIAS	ESE	ASE	MSE	CP	BIAS	ESE	ASE	MSE	CP	BIAS	ESE	ASE	MSE	CP
unstratified KM estimation: $\exp(\gamma) = 0.75,\ \exp(\delta) = 0.8$ and $\exp(\gamma_r) = 1$																	
0.15	GL		0.0069	0.1182	0.1161	0.0140	0.9490	-0.0018	0.1285	0.1252	0.0165	0.9520	0.0061	0.1370	0.1394	0.0188	0.9525
	P_{wEE}	6	0.0046	0.1361	0.1365	0.0185	0.9530	-0.0033	0.1513	0.1464	0.0229	0.9455	0.0006	0.1620	0.1615	0.0262	0.9485
		12	0.0050	0.1343	0.1342	0.0181	0.9500	-0.0038	0.1492	0.1442	0.0223	0.9425	0.0009	0.1599	0.1594	0.0255	0.9490
	P_{wO}	6	0.0050	0.1400	0.1400	0.0196	0.9500	-0.0022	0.1588	0.1545	0.0252	0.9470	-0.0013	0.1742	0.1760	0.0303	0.9535
		12	0.0055	0.1383	0.1379	0.0192	0.9500	-0.0028	0.1570	0.1526	0.0246	0.9440	-0.0011	0.1725	0.1746	0.0297	0.9540
0.3	GL		0.0058	0.1159	0.1181	0.0135	0.9550	0.0032	0.1240	0.1267	0.0154	0.9580	0.0001	0.1409	0.1409	0.0199	0.9505
	P_{wEE}	6	0.0030	0.1332	0.1358	0.0177	0.9520	-0.0027	0.1457	0.1454	0.0212	0.9495	-0.0046	0.1610	0.1600	0.0259	0.9495
		12	0.0031	0.1309	0.1338	0.0171	0.9530	-0.0017	0.1430	0.1434	0.0204	0.9500	-0.0033	0.1596	0.1582	0.0255	0.9520
	P_{wO}	6	0.0029	0.1367	0.1394	0.0187	0.9510	-0.0028	0.1544	0.1535	0.0238	0.9450	-0.0045	0.1751	0.1748	0.0307	0.9550
		12	0.0030	0.1346	0.1375	0.0181	0.9520	-0.0019	0.1520	0.1518	0.0231	0.9460	-0.0031	0.1743	0.1736	0.0304	0.9540
stratified KM estimation: $\exp(\gamma) = 0.75,\ \exp(\delta) = 0.8$ and $\exp(\gamma_r) = 1$																	
0.15	GL		0.0006	0.1172	0.1161	0.0137	0.9415	-0.0015	0.1283	0.1250	0.0164	0.9415	-0.0004	0.1400	0.1396	0.0196	0.9480
	P_{wEE}	6	-0.0011	0.1365	0.1364	0.0186	0.9575	-0.0005	0.1460	0.1461	0.0213	0.9530	-0.0049	0.1618	0.1617	0.0262	0.9485
		12	-0.0009	0.1346	0.1342	0.0181	0.9550	-0.0001	0.1441	0.1439	0.0207	0.9525	-0.0050	0.1601	0.1596	0.0256	0.9500
	P_{wO}	6	-0.0011	0.1365	0.1399	0.0186	0.9615	-0.0005	0.1460	0.1542	0.0213	0.9670	-0.0049	0.1618	0.1764	0.0262	0.9685
		12	-0.0009	0.1346	0.1378	0.0181	0.9615	-0.0001	0.1441	0.1524	0.0207	0.9645	-0.0050	0.1601	0.1750	0.0256	0.9695
0.3	GL		0.0058	0.1158	0.1181	0.0134	0.9540	0.0033	0.1239	0.1267	0.0154	0.9585	0.0000	0.1407	0.1410	0.0198	0.9485
	P_{wEE}	6	0.0030	0.1331	0.1358	0.0177	0.9515	-0.0025	0.1455	0.1454	0.0212	0.9520	-0.0046	0.1609	0.1600	0.0259	0.9490
		12	0.0031	0.1309	0.1338	0.0171	0.9540	-0.0016	0.1428	0.1434	0.0204	0.9505	-0.0034	0.1594	0.1582	0.0254	0.9530
	P_{wO}	6	0.0030	0.1331	0.1394	0.0177	0.9615	-0.0025	0.1455	0.1535	0.0212	0.9625	-0.0046	0.1609	0.1748	0.0259	0.9660
		12	0.0031	0.1309	0.1375	0.0171	0.9615	-0.0016	0.1429	0.1519	0.0204	0.9640	-0.0034	0.1594	0.1737	0.0254	0.9660
Cox regression: $\exp(\gamma) = 0.75,\ \exp(\delta) = 0.8$ and $\exp(\gamma_r) = 1$																	
0.15	GL		0.0069	0.1182	0.1161	0.0140	0.9490	-0.0018	0.1285	0.1252	0.0165	0.9515	0.0062	0.1372	0.1394	0.0188	0.9525
	P_{wEE}	6	0.0046	0.1361	0.1365	0.0185	0.9530	-0.0033	0.1513	0.1464	0.0229	0.9460	0.0007	0.1621	0.1614	0.0263	0.9485
		12	0.0050	0.1343	0.1342	0.0181	0.9495	-0.0038	0.1492	0.1442	0.0223	0.9430	0.0009	0.1599	0.1594	0.0256	0.9470
	P_{wO}	6	0.0046	0.1370	0.1400	0.0188	0.9515	-0.0024	0.1524	0.1545	0.0232	0.9520	0.0003	0.1632	0.1760	0.0266	0.9600
		12	0.0050	0.1351	0.1379	0.0183	0.9520	-0.0030	0.1503	0.1526	0.0226	0.9545	0.0005	0.1606	0.1747	0.0258	0.9650
0.3	GL		0.0058	0.1158	0.1181	0.0134	0.9535	0.0032	0.1240	0.1267	0.0154	0.9590	0.0000	0.1409	0.1410	0.0198	0.9480
	P_{wEE}	6	0.0030	0.1331	0.1358	0.0177	0.9515	-0.0026	0.1456	0.1454	0.0212	0.9520	-0.0047	0.1609	0.1600	0.0259	0.9480
		12	0.0031	0.1309	0.1338	0.0171	0.9545	-0.0017	0.1429	0.1434	0.0204	0.9500	-0.0034	0.1595	0.1582	0.0254	0.9525
	P_{wO}	6	0.0029	0.1336	0.1394	0.0178	0.9595	-0.0028	0.1469	0.1535	0.0216	0.9600	-0.0050	0.1624	0.1748	0.0264	0.9630
		12	0.0030	0.1312	0.1375	0.0172	0.9615	-0.0019	0.1441	0.1519	0.0207	0.9615	-0.0037	0.1607	0.1737	0.0258	0.9645

Table 4.2: Empirical properties of treatment effect estimators $\widehat{\beta}$ from a GL model fitted under different estimation procedures; $\tau = 1, \mathbb{E}(N(\tau)|X = 0) = 0.8, \pi_D = P(T \leq \tau|X = 0) = 0.15$ and $0.3, \pi_r = P(C_r \leq \tau|X = 0) = 0.2, 0.4$ and $0.6, \exp(\gamma) = 0.75, \exp(\delta) = 0.8, \exp(\gamma_r) = 1, \widehat{G} \longrightarrow G, n = 500$ individuals, $n_{sim} = 2000$ simulation runs.

			$\pi_r = 0.2$					$\pi_r = 0.4$					$\pi_r = 0.6$				
π_D	Method	R	BIAS	ESE	ASE	MSE	CP	BIAS	ESE	ASE	MSE	CP	BIAS	ESE	ASE	MSE	CP
stratified KM estimation: $\exp(\gamma) = 0.75,\ \exp(\delta) = 1$ and $\exp(\gamma_r) = 0.85$																	
0.15	GL		0.0038	0.1194	0.1157	0.0143	0.9410	-0.0008	0.1252	0.1234	0.0157	0.9490	-0.0015	0.1358	0.1368	0.0184	0.9415
	P_{wEE}	6	-0.0002	0.1411	0.1357	0.0199	0.9485	-0.0029	0.1446	0.1442	0.0209	0.9570	-0.0043	0.1616	0.1589	0.0261	0.9460
		12	0.0005	0.1384	0.1335	0.0191	0.9490	-0.0023	0.1427	0.1420	0.0204	0.9525	-0.0040	0.1597	0.1568	0.0255	0.9460
	P_{wO}	6	-0.0001	0.1411	0.1390	0.0199	0.9540	-0.0027	0.1446	0.1517	0.0209	0.9675	-0.0040	0.1616	0.1725	0.0261	0.9615
		12	0.0005	0.1384	0.1369	0.0191	0.9535	-0.0022	0.1427	0.1498	0.0204	0.9650	-0.0037	0.1597	0.1710	0.0255	0.9615
0.3	GL		0.0008	0.1197	0.1179	0.0143	0.9475	-0.0027	0.1260	0.1256	0.0159	0.9485	-0.0032	0.1381	0.1381	0.0191	0.9530
	P_{wEE}	6	-0.0018	0.1358	0.1354	0.0184	0.9525	-0.0010	0.1435	0.1438	0.0206	0.9445	-0.0004	0.1548	0.1572	0.0239	0.9515
		12	-0.0015	0.1339	0.1334	0.0179	0.9515	-0.0012	0.1417	0.1418	0.0201	0.9450	-0.0007	0.1535	0.1554	0.0236	0.9525
	P_{wO}	6	-0.0017	0.1358	0.1387	0.0184	0.9570	-0.0009	0.1435	0.1513	0.0206	0.9595	-0.0001	0.1548	0.1709	0.0239	0.9705
		12	-0.0015	0.1339	0.1368	0.0179	0.9565	-0.0010	0.1417	0.1497	0.0201	0.9560	-0.0004	0.1535	0.1697	0.0236	0.9700
Cox regression: $\exp(\gamma) = 0.75,\ \exp(\delta) = 1$ and $\exp(\gamma_r) = 0.85$																	
0.15	GL		0.0006	0.1151	0.1157	0.0132	0.9550	-0.0030	0.1256	0.1237	0.0158	0.9465	-0.0043	0.1381	0.1367	0.0191	0.9485
	P_{wEE}	6	-0.0008	0.1350	0.1360	0.0182	0.9565	-0.0035	0.1463	0.1450	0.0214	0.9510	-0.0065	0.1594	0.1587	0.0254	0.9470
		12	-0.0001	0.1330	0.1337	0.0177	0.9530	-0.0033	0.1444	0.1427	0.0209	0.9495	-0.0069	0.1569	0.1566	0.0247	0.9505
	P_{wO}	6	-0.0005	0.1356	0.1393	0.0184	0.9580	-0.0040	0.1472	0.1524	0.0217	0.9620	-0.0066	0.1605	0.1722	0.0258	0.9640
		12	0.0002	0.1336	0.1371	0.0178	0.9595	-0.0037	0.1452	0.1505	0.0211	0.9610	-0.0069	0.1578	0.1707	0.0249	0.9635
0.3	GL		-0.0022	0.1203	0.1178	0.0145	0.9445	0.0022	0.1246	0.1258	0.0155	0.9555	0.0025	0.1428	0.1382	0.0204	0.9430
	P_{wEE}	6	-0.0050	0.1399	0.1355	0.0196	0.9345	0.0017	0.1429	0.1441	0.0204	0.9565	0.0040	0.1594	0.1569	0.0254	0.9460
		12	-0.0049	0.1376	0.1334	0.0190	0.9345	0.0016	0.1411	0.1421	0.0199	0.9550	0.0038	0.1579	0.1550	0.0249	0.9395
	P_{wO}	6	-0.0048	0.1408	0.1387	0.0198	0.9405	0.0013	0.1444	0.1516	0.0208	0.9660	0.0037	0.1604	0.1706	0.0257	0.9605
		12	-0.0046	0.1385	0.1369	0.0192	0.9405	0.0011	0.1423	0.1499	0.0203	0.9675	0.0035	0.1586	0.1694	0.0252	0.9615

Table 4.3: Empirical properties of treatment effect estimators $\widehat{\beta}$ from a GL model fitted under different estimation procedures; $\tau = 1, \mathbb{E}(N(\tau)|X = 0) = 0.8, \pi_D = P(T \leq \tau|X = 0) = 0.15$ and $0.3, \pi_r = P(C_r \leq \tau|X = 0) = 0.2, 0.4$ and $0.6, \exp(\gamma) = 0.75, \exp(\delta) = 1, \exp(\gamma_r) = 0.85, \widehat{G} \longrightarrow G, n = 500$ individuals, $n_{sim} = 2000$ simulation runs.

			$\pi_r = 0.2$					$\pi_r = 0.4$					$\pi_r = 0.6$				
π_D	Method	R	BIAS	ESE	ASE	MSE	CP	BIAS	ESE	ASE	MSE	CP	BIAS	ESE	ASE	MSE	CP
stratified KM estimation: $\exp(\gamma) = 0.75,\ \exp(\delta) = 0.8$ and $\exp(\gamma_r) = 0.85$																	
0.15	GL		0.0007	0.1153	0.1155	0.0133	0.9520	-0.0008	0.1248	0.1237	0.0156	0.9470	-0.0057	0.1384	0.1366	0.0192	0.9460
	P_{wEE}	6	-0.0007	0.1345	0.1358	0.0181	0.9585	-0.0002	0.1442	0.1446	0.0208	0.9535	-0.0064	0.1608	0.1585	0.0259	0.9445
		12	-0.0005	0.1317	0.1335	0.0173	0.9590	-0.0009	0.1423	0.1423	0.0202	0.9520	-0.0065	0.1590	0.1564	0.0253	0.9440
	P_{wO}	6	-0.0007	0.1345	0.1390	0.0181	0.9635	-0.0001	0.1442	0.1520	0.0208	0.9670	-0.0061	0.1608	0.1721	0.0259	0.9630
		12	-0.0005	0.1317	0.1369	0.0173	0.9645	-0.0007	0.1423	0.1502	0.0202	0.9675	-0.0062	0.1590	0.1706	0.0253	0.9630
0.3	GL		0.0006	0.1159	0.1176	0.0135	0.9520	0.0038	0.1258	0.1253	0.0158	0.9480	0.0021	0.1388	0.1381	0.0193	0.9550
	P_{wEE}	6	0.0031	0.1295	0.1353	0.0168	0.9605	-0.0025	0.1448	0.1438	0.0210	0.9470	-0.0028	0.1577	0.1570	0.0249	0.9455
		12	0.0032	0.1279	0.1332	0.0164	0.9620	-0.0017	0.1421	0.1419	0.0202	0.9465	-0.0019	0.1562	0.1552	0.0244	0.9455
	P_{wO}	6	0.0032	0.1295	0.1385	0.0168	0.9645	-0.0024	0.1448	0.1513	0.0210	0.9560	-0.0025	0.1577	0.1707	0.0249	0.9640
		12	0.0033	0.1279	0.1367	0.0164	0.9640	-0.0016	0.1421	0.1497	0.0202	0.9595	-0.0016	0.1562	0.1695	0.0244	0.9660
Cox regression: $\exp(\gamma) = 0.75,\ \exp(\delta) = 0.8$ and $\exp(\gamma_r) = 0.85$																	
0.15	GL		0.0064	0.1171	0.1155	0.0137	0.9515	-0.0022	0.1263	0.1238	0.0159	0.9460	0.0053	0.1359	0.1365	0.0185	0.9490
	P_{wEE}	6	0.0044	0.1351	0.1359	0.0183	0.9570	-0.0041	0.1512	0.1449	0.0229	0.9395	0.0009	0.1561	0.1584	0.0244	0.9535
		12	0.0047	0.1326	0.1336	0.0176	0.9580	-0.0043	0.1489	0.1426	0.0222	0.9400	0.0009	0.1539	0.1563	0.0237	0.9510
	P_{wO}	6	0.0045	0.1357	0.1391	0.0184	0.9605	-0.0037	0.1514	0.1524	0.0229	0.9525	0.0011	0.1588	0.1719	0.0252	0.9655
		12	0.0047	0.1331	0.1369	0.0177	0.9595	-0.0039	0.1491	0.1504	0.0222	0.9550	0.0010	0.1560	0.1704	0.0243	0.9685
0.3	GL		0.0060	0.1159	0.1176	0.0135	0.9525	0.0037	0.1258	0.1253	0.0158	0.9485	0.0021	0.1388	0.1381	0.0193	0.9530
	P_{wEE}	6	0.0031	0.1295	0.1353	0.0168	0.9605	-0.0026	0.1448	0.1438	0.0210	0.9480	-0.0029	0.1577	0.1570	0.0249	0.9455
		12	0.0032	0.1279	0.1332	0.0164	0.9620	-0.0018	0.1421	0.1418	0.0202	0.9455	-0.0019	0.1562	0.1552	0.0244	0.9450
	P_{wO}	6	0.0031	0.1297	0.1385	0.0168	0.9640	-0.0025	0.1456	0.1513	0.0212	0.9580	-0.0021	0.1598	0.1707	0.0255	0.9590
		12	0.0032	0.1279	0.1367	0.0164	0.9660	-0.0017	0.1427	0.1497	0.0204	0.9600	-0.0012	0.1576	0.1695	0.0248	0.9620

Table 4.4: Empirical properties of treatment effect estimators $\widehat{\beta}$ from a GL model fitted under different estimation procedures; $\tau = 1, \mathbb{E}(N(\tau)|X = 0) = 0.8, \pi_D = P(T \leq \tau|X = 0) = 0.15$ and $0.3, \pi_r = P(C_r \leq \tau|X = 0) = 0.2, 0.4$ and $0.6, \exp(\gamma) = 0.75, \exp(\delta) = 0.8, \exp(\gamma_r) = 0.85, \widehat{G} \longrightarrow G, n = 500$ individuals, $n_{sim} = 2000$ simulation runs.

4.4.1 The Ghosh-Lin estimator

Consider a quite general data-generating framework where the non-fatal event and death intensities (4.2.1) in Figure 1.1(e) may take the following form:

$$\lambda_{(k-1)k}(t|\mathcal{H}(t)) = \lambda_{k0}(t)\exp(\gamma_k X + \gamma_U \log(U)) \ , \qquad k = 1, 2, \dots \ , \tag{4.4.1}$$

$$\lambda_{kD_k}(t|\mathcal{H}(t)) = \rho_{k0}(t)\exp(\delta_k X + \delta_U \log(U)) \ , \qquad k = 0, 1, \dots \ , \tag{4.4.2}$$

where U is an unobserved gamma-distributed random effect representing heterogeneity across individuals with $\mathbb{E}(U) = 1$ and $\text{Var}(U) = \phi$, $\gamma_U \in \{0, 1\}$ and $\delta_U \in \{0, 1\}$. We let $G_U(u) = P(U \leq u)$ denote the cumulative distribution function of U, with associated density function $dG_U(u) = g_U(u)du$. The baseline intensities may vary according to the cumulative number of non-fatal events, but for simplicity we assume common treatment effects with $\gamma_k = \gamma$ and $\delta_k = \delta, k = 1, 2, \dots$. We further set $\rho_{k0}(t) = \rho_{k0}$, so the baseline death intensities are assumed constant over time for a specified k, $k = 0, 1, \dots$. If $\phi = 0$ or $\gamma_U = \delta_U = 0$, there is no random effect acting on the event and death intensities; this scenario is labelled as *GL-1* and considered in Sections 4.4.1.1 (*GL-1-1*) and 4.4.1.2 (*GL-1-2*). Scenarios where U acts multiplicatively on the event and death intensities (when $\gamma_U = 1, \delta_U = 1$ with $\phi > 0$) are studied in Section 4.4.1.3 and labelled as *GL-2*.

In all investigations that follow we set the administrative censoring time to $\tau = 1$ and the treatment effects on the intensities to $\exp(\gamma) = 0.75, 1$ and $\exp(\delta) = 0.8, 0.9, 1, 1.1$, respectively. We also restrict attention to the case where the adopted censoring model is correctly specified, i.e. $G^\star = G$.

4.4.1.1 Poisson processes for recurrent non-fatal events and death (GL-1-1)

The simplest data-generating process within framework (4.4.1)-(4.4.2) arises when $\gamma_U = \delta_U = \phi = 0$, $\rho_{k0} = \rho_0$ and $\lambda_{k0}(t) = \lambda_0(t)$, giving true intensities of the form

$$\lambda_{(k-1)k}(t|\mathcal{H}(t)) = \lambda_0(t)\exp(\gamma X) = \lambda(t|X)$$
$$\lambda_{kD_k}(t|\mathcal{H}(t)) = \rho_0\exp(\delta X) = \rho(t|X) \ .$$

Death is thus independent of the time-inhomogeneous Poisson process $\{N(t), 0 < t\}$, for which we assume a Weibull baseline intensity of the form $\lambda_0(t) = \lambda_0^\kappa \kappa t^{\kappa-1}$, $\kappa > 0$. In what follows we refer to this data-generating mechanism as *GL-1-1*.

We now look at the adequacy of the GL model (4.2.3) when the true process has intensities $\lambda(t|X)$ and $\rho(t|X)$, as specified above. Since $r(t|X) = \lambda(t|X)$ and $S(t|X) = \exp(-\rho_0 e^{\delta X} t)$ under these restrictions, we find

$$\mu(t|X) = \int_0^t \lambda_0(u) e^{\gamma X} \exp\big(-e^{\delta X} \rho_0 u\big) du \ , \tag{4.4.3}$$

and so the mean function ratio is

$$\frac{\mu(t|X=1)}{\mu(t|X=0)} = e^{\gamma} \cdot \frac{\int_0^t u^{\kappa-1} \exp\big(-e^{\delta} \rho_0 u\big) du}{\int_0^t u^{\kappa-1} \exp\big(-\rho_0 u\big) du} \ , \tag{4.4.4}$$

which is a function of time t unless $\delta = 0$. The GL model is thus misspecified in settings in which treatment affects the death intensity. It can be seen from (4.4.4) that the degree of misspecification depends in particular on the magnitude of $\delta, P(T \leq \tau|X = 0)$ (here through ρ_0) and κ. We explore this further by first computing the true values of $\mu(t|X)$ for $t \in (0, 1)$ and $X = 0, 1$ under some specific true processes, and then plotting the resulting mean function ratio (4.4.4) over time.

Results are displayed in Figure 4.1 ($\kappa = 1$) and Figure 4.2 ($\kappa \geq 1$), where departures from a horizontal line indicate misspecification of the GL model; that is, the ratio (4.4.4) of the mean function is not constant but rather varies with time. The mean function ratio is closer to constant with decreasing $P(T \leq \tau|X = 0)$ and as the magnitude of δ decreases. If there is no treatment effect on death (i.e. $\exp(\delta) = 1$), then $\mu(t|X = 1)/\mu(t|X = 0)$ is constant and the proportional means assumption is satisfied; in all other scenarios the ratio varies with time and the assumption is violated. The GL model can however be a reasonable approximation when the effect of treatment on mortality is small; see red line in Figure C.1 of Appendix C.3 which mimics the parameter settings of the bone metastases trial in Section 1.3.3. For the scenarios investigated in Figure C.1, the maximum deviation of the mean function ratio from constancy over $0 \leq t \leq 2$ is only about 8.2% $(= 100(0.595 - e^{\gamma})/e^{\gamma})$. Figure 4.2 demonstrates that adequacy of the GL model also varies with κ; the mean function ratio is closer to constant with decreasing κ. It is apparent that the dependency of κ on model adequacy becomes stronger with increasing $P(T \leq \tau|X = 0)$ and as the magnitude of δ increases.

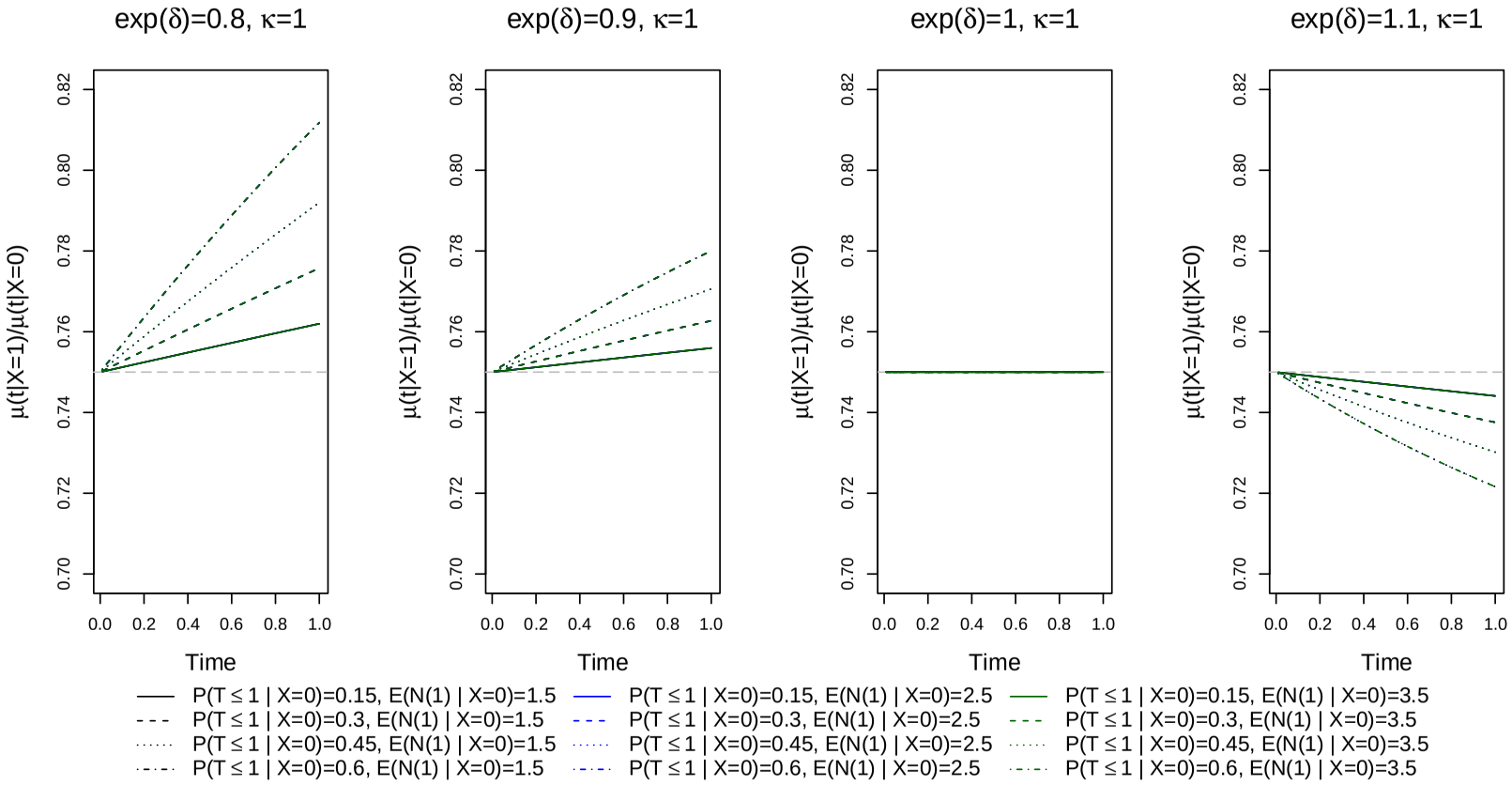


Figure 4.1: *GL-1-1:* Plots of $\mu(t|X=1)/\mu(t|X=0)$ for $e^{\gamma}=0.75$ and different values of $P(T\leq\tau|X=0), \delta$ and $\mathbb{E}(N(\tau)|X=0)$ when the true process has intensities $\lambda(t|X)=\lambda_0\exp(\gamma X)$ and $\rho(t|X)=\rho_0\exp(\delta X)$; $\tau=1, \kappa=1$.

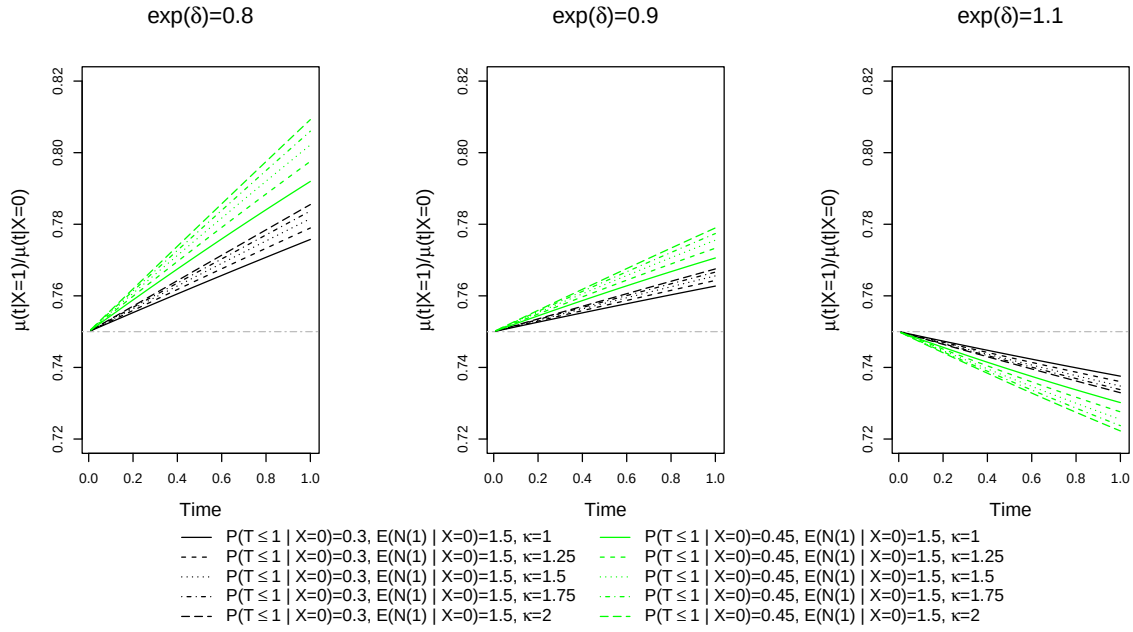


Figure 4.2: *GL-1-1:* Plots of $\mu(t|X=1)/\mu(t|X=0)$ for $\exp(\gamma)=0.75$ and different values of $\kappa \geq 1, P(T \leq \tau|X=0), \exp(\delta)$ and $\mathbb{E}(N(\tau)|X=0)$ when the true process has intensities $\lambda(t|X)=\lambda_0^\kappa \kappa t^{\kappa-1}\exp(\gamma X)$ and $\rho(t|X)=\rho_0\exp(\delta X)$; $\tau=1$.

We now consider the limiting estimands $\beta^\star$ for the GL model. If $\{Z(t), 0<t\} \perp\!\!\!\perp C_r|X$ and $G^\star = G$, the expectations needed to compute $\beta^\star$ are

$$\mathbb{E}(U_1(\alpha_r, G^\star)) = \sum_{x=0,1} \frac{P(X=x)h'(\alpha_r+\beta x)}{h(\alpha_r+\beta x)}\big(\mu(s_r|X=x) - h(\alpha_r+\beta x)\big)\ , \quad r=1,\ldots,R,$$

$$\mathbb{E}(U_1(\beta, G^\star)) = \sum_{r=1}^{R} \frac{P(X=1)h'(\alpha_r+\beta)}{h(\alpha_r+\beta)}\big(\mu(s_r|X=1) - h(\alpha_r+\beta)\big)$$

for the pseudo-Poisson approach in Section 4.3.2.1, with $\mu(s_r|X)$ given as in (4.4.3), and

$$s^{(l,\star)}(t,\beta) = \mathbb{E}_X\big(X^l\exp(\beta X)\big) = \sum_{x=0,1} P(X=x)x^l\exp(\beta x)$$

$$s^{(l,\star)}(t) = \mathbb{E}_X\big(X^l S(t|X)\mathbb{E}(dN(t)|T\geq t, X)\big) = \sum_{x=0,1} P(X=x)x^l S(t|X=x)r(t|X=x)$$

for the GL estimation method in Section 4.3.1, $l = 0, 1$. If $P(X = 1) = 0.5$ in addition to the assumptions above, we obtain for the latter estimation procedure

$$\beta^{\star}_{GL} = \gamma + \log\left(\frac{\int_0^{\tau} \lambda_0(u) \exp(-e^{\delta}\rho_0 u) du}{\int_0^{\tau} \lambda_0(u) \exp(-\rho_0 u) du} \right) . \tag{4.4.5}$$

For a given value of κ, we determined the baseline parameters λ_0 and ρ_0 by setting $P(T \leq \tau | X = 0) \in (0, 0.6]$ and $\mathbb{E}(N(\tau)|X = 0) = 1.5, 2.5$ and 3.5, and calculated $\beta^{\star}_{GL}$ and $\beta^{\star}_{P}$ for each set of regression coefficients γ and δ.

Figure 4.3 displays plots of the limiting values $\exp(\beta^{\star}_{GL})$ (left panels) and $\exp(\beta^{\star}_{P})$ (right panels) against $P(T \leq \tau | X = 0)$ when $\kappa = 1$, $G^{\star} = G$ and $R = 6$ for the pseudo-Poisson approach. If there is no effect of treatment on death (i.e. $\exp(\delta) = 1$), then $\exp(\beta^{\star}_{GL}) = \exp(\beta^{\star}_{P}) = \exp(\gamma)$ and interpretation of $\beta^{\star}$ is simple; see solid lines in upper and lower panels of Figure 4.3. In scenarios where $\exp(\delta) \neq 1$, $\exp(\beta^{\star}_{GL})$ and $\exp(\beta^{\star}_{P})$ are functions of the parameters of the underlying intensity-based process *GL-1-1* and the degree of model misspecification; see also Figure 4.1. Treatment effects on the death intensity impact $\beta^{\star}$ through a reduction ($\exp(\delta) > 1$) or increase ($\exp(\delta) < 1$) in the cumulative mean function for non-fatal events for the experimental treatment. If the mortality rate is low and the effect of treatment on mortality is relatively small, $\beta^{\star}$ appears close to the treatment effect γ on the non-fatal event intensities. It is also readily apparent from both (4.4.5) and Figure 4.3 that $\beta^{\star}$ does not depend on $\mathbb{E}(N(\tau)|X = 0)$, here through λ_0. Overall, since $\beta^{\star}_{P}$ is closer to γ than $\beta^{\star}_{GL}$, the estimand $\beta^{\star}_{P}$ appears less sensitive to misspecification of the GL model (4.2.3) compared to $\beta^{\star}_{GL}$. Similar findings were found in Chapter 3 when comparing CIF-based estimands of estimators arising from the Fine-Gray procedure and direct binomial regression (Fine and Gray, 1999; Scheike et al., 2008); see contour plots in Figure 3.1. Figure 4.4 where $\kappa \geq 1$ shows that $\beta^{\star}_{P}$ is also less sensitive to time-inhomogeneous non-fatal event intensities than $\beta^{\star}_{GL}$.

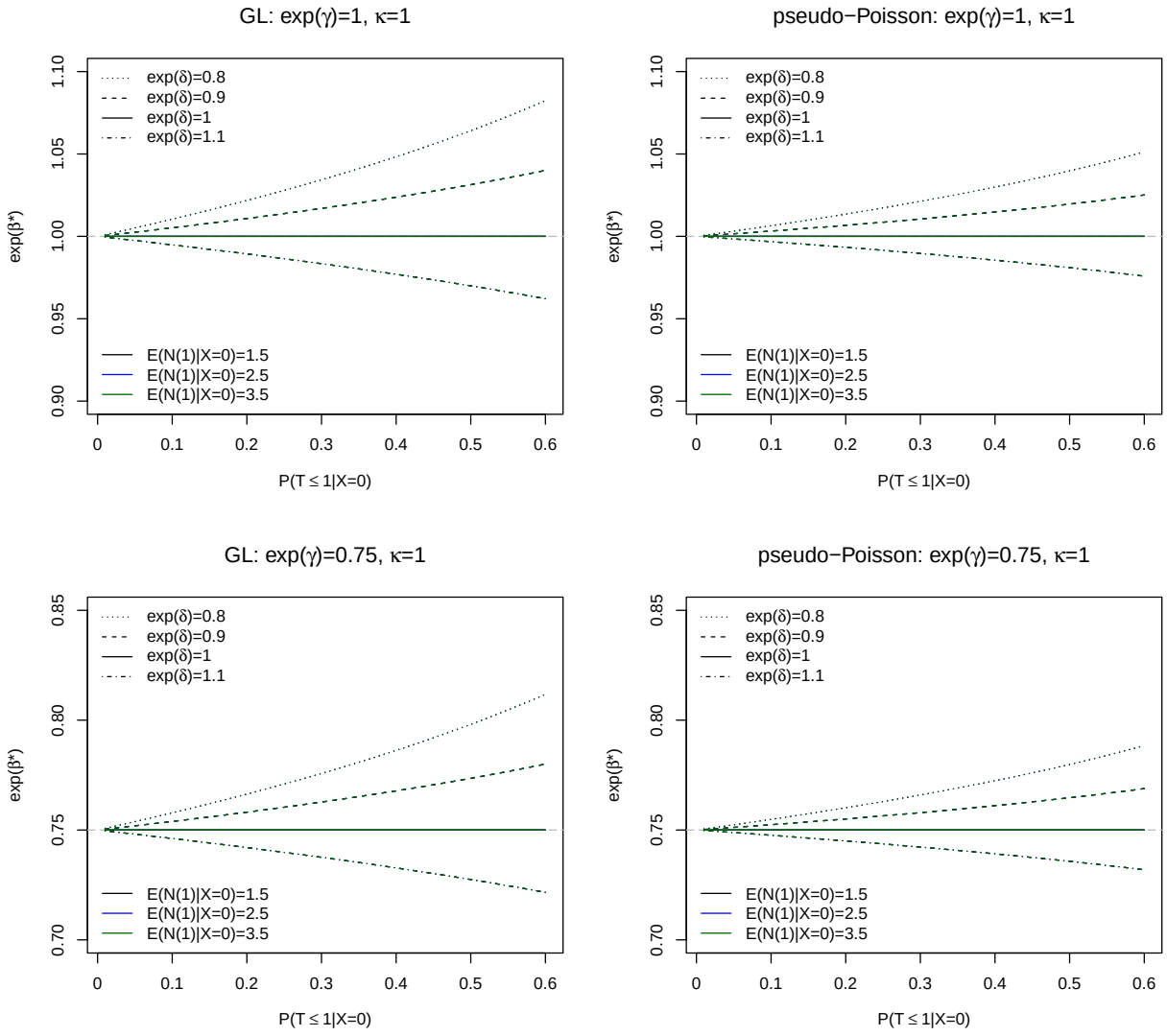


Figure 4.3: *GL-1-1:* Limiting values $\exp(\beta^\star_{GL})$ (left panels) and $\exp(\beta^\star_P)$ (right panels) as a function of $P(T \leq \tau|X=0)$ for different values of $\exp(\gamma), \exp(\delta)$ and $\mathbb{E}(N(\tau)|X=0)$ when the true process has time-homogeneous intensities $\lambda(t|X) = \lambda_0 \exp(\gamma X)$ and $\rho(t|X) = \rho_0 \exp(\delta X)$; $\kappa = 1, \tau = 1$, pseudo-Poisson estimation is based on $R = 6$ equi-spaced time points on $(0, \tau)$, $G^\star = G$.

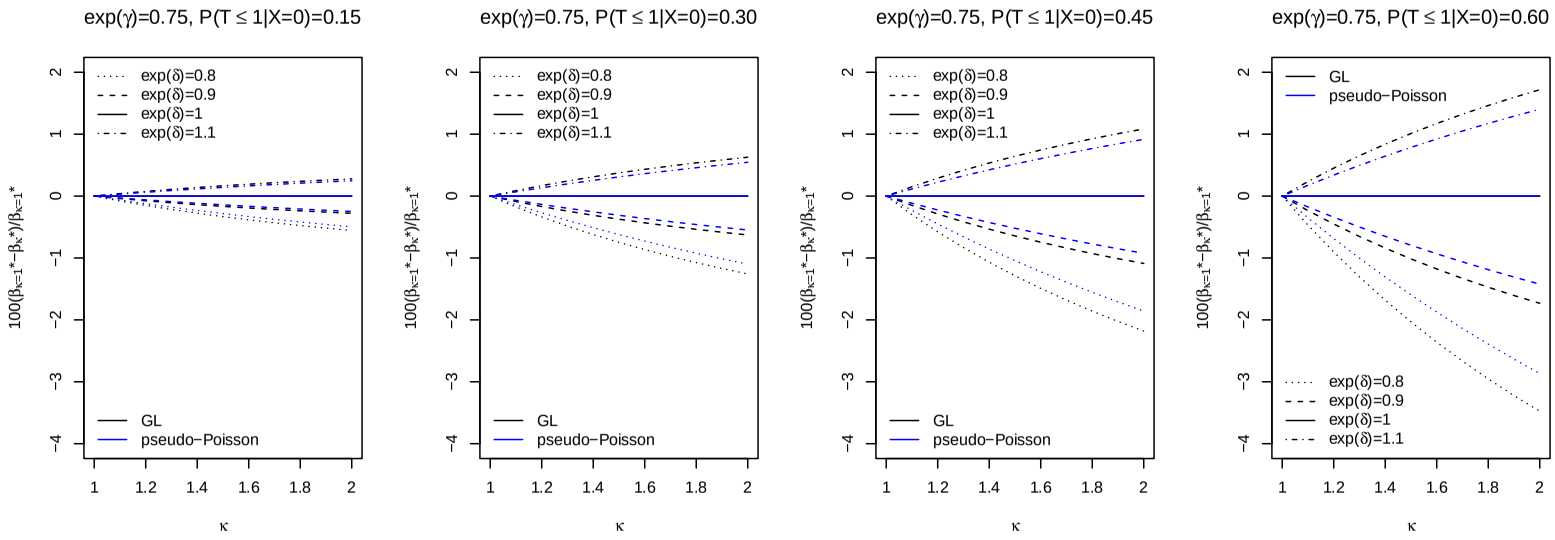


Figure 4.4: *GL-1-1:* Dependency of limiting values $\beta^{\star}_{GL}$ and $\beta^{\star}_{P}$ on κ for different values of $\exp(\delta)$ and $P(T \leq \tau|X=0)$; $\tau = 1$, pseudo-Poisson estimation is based on $R = 6$ equi-spaced time points on $(0, \tau)$, $G^{\star} = G$.

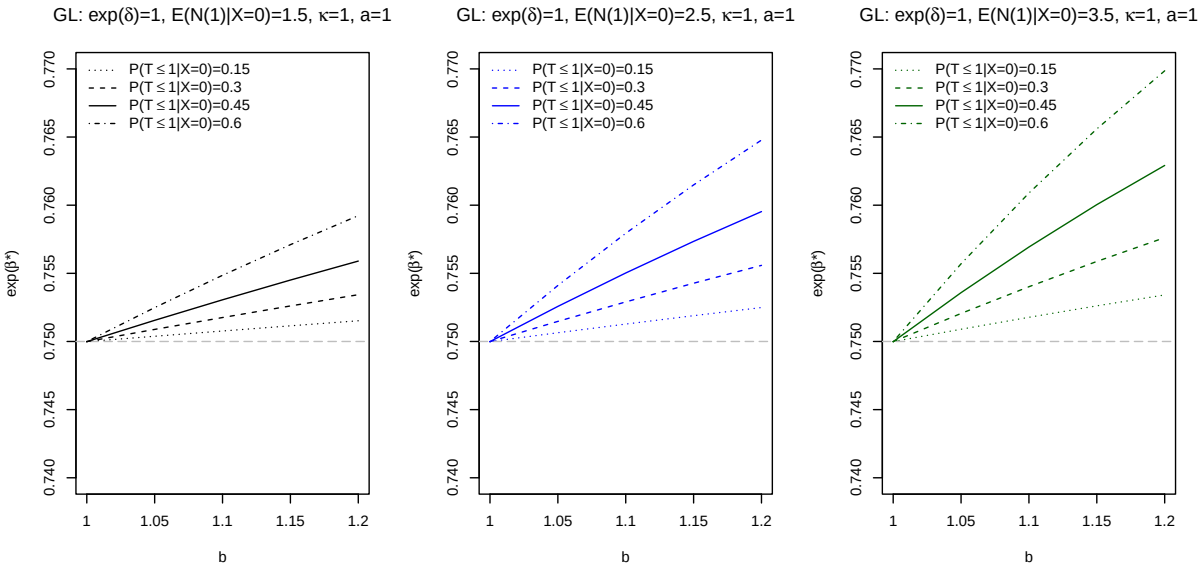


Figure 4.5: *GL-1-2:* Limiting values $\exp(\beta^{\star}_{GL})$ as a function of b for different values of $P(T \leq \tau|X=0)$ and $\mathbb{E}(N(\tau)|X=0) = 1.5$ (left panel), 2.5 (middle panel), 3.5 (right panel) when the true process has intensities $\lambda_k(t|X) = \lambda(t|X) = \lambda_0 \exp(\gamma X)$ and $\rho_k(t|X) = b^k \rho_0 \exp(\delta X)$ for $\exp(\gamma) = 0.75$ and $\exp(\delta) = 1$; $K = 20, \kappa = 1, \tau = 1$, $G^{\star} = G$.

4.4.1.2 Markov processes with event-specific baseline intensities (GL-1-2)

We next assume Markov intensities of the form

$$\lambda_{(k-1)k}(t|\mathcal{H}(t)) = \lambda_{(k-1)k}(t|Z(t^-) = k-1, X) = \lambda_{0k}(t)\exp(\gamma X) = \lambda_k(t|X)\ , \quad k = 1, 2, \dots\ ,$$
$$\lambda_{kD_k}(t|\mathcal{H}(t)) = \lambda_{kD_k}(t|Z(t^-) = k, X) = \rho_{0k}\exp(\delta X) = \rho_k(t|X)\ , \quad k = 0, 1, \dots\ ,$$

with $\lambda_{0k}(t) = a^k\lambda_0(t)$ and $\rho_{0k} = b^k\rho_0$ so that the occurrence of each non-fatal event increases the baseline intensities by a factor of $a \geq 1$ or $b \geq 1$, respectively. We refer to this data-generating mechanism as *GL-1-2*. Note that, if $a = b = 1$, *GL-1-2* coincides with *GL-1-1* from the previous section. Straightforward calculations in Appendix C.2 show that the cumulative mean function of

the "overall" counting process $\{N(t), 0 < t\}$ is given by

$$\mu(t|X) = \int_0^t r(u|X)S(u^-|X)du = \int_0^t \Big(\sum_{k=1}^{K} P(Z(u^-) = k-1|X)\lambda_k(u|X)\Big)du \ . \quad (4.4.6)$$

We focus on the special case where $a = 1$ but $b > 1$; this induces a dependency between $\{N(t), 0 < t\}$ and $\{N_D(t), 0 < t\}$. In the calculation of limiting values and mean function ratios $\mu(t|X = 1)/\mu(t|X = 0)$ under this data-generating mechanism, we consider a multistate process with a finite number of states by setting a maximum number of non-fatal events, denoted by K in what follows. This can be set sufficiently large so that it has little impact on insights regarding the properties of estimators; here we set $K = 20$.

The plots of $\mu(t|X = 1)/\mu(t|X = 0)$ in Figure C.2 of Appendix C.3 demonstrate that under *GL-1-2* the proportional means assumption in the GL model is violated even when $\exp(\delta) = 1$, with the extent of misspecification becoming larger as b increases and as the non-fatal and fatal event rates increase. For cases where $a > 1$, $\mu(0|X = 1)/\mu(0|X = 0) = \exp(\gamma)$ does no longer hold.

Under model misspecification the implied estimand under the GL estimation procedure is the limiting value $\beta^\star_{GL}$, the solution to (4.3.7) with

$$s^{(l,\star)}(t) = \sum_{k=1}^{K} \mathbb{E}_X\big(X^l P(Z(t^-) = k-1|X)\lambda_k(t|X)\big) \ , \qquad l = 0, 1 \ ,$$

and $s^{(l,\star)}(t, \beta)$ as in Section 4.4.1.1. The resulting limiting values $\exp(\beta^\star_{GL})$ are displayed in Figure 4.5 as a function of b when the effects of treatment in the true intensity-based process are $\exp(\gamma) = 0.75$ and $\exp(\delta) = 1$. Results in Figure 4.5 are aligned with those from Figure C.2 in that the event-specific baseline death intensities affect $\beta^\star_{GL}$ through an increase ($b > 1$) in $\mu(t|X = 1)$. As b becomes further away from 1 the treatment effect on $\mu(t|X)$ becomes more attenuated, resulting in a greater deviation from $\exp(\gamma)$. When the expected number of recurrent non-fatal events increases or the mortality rate increases, the impact of b becomes stronger; in this case $\beta^\star_{GL}$ is a function of several parameters of the data-generating process and its interpretation becomes challenging. We emphasize however that for relevant values of $b, P(T \leq 1|X = 0)$ and $\mathbb{E}(N(1)|X = 0)$, the GL estimand $\exp(\beta^\star_{GL})$ is fairly close to $\exp(\gamma)$. We note that, if $\exp(\gamma) = \exp(\delta) = 1$, then $\exp(\beta^\star_{GL}) = 1$ for all values of b (not shown here).

Plots of $\exp(\beta_P^\star)$ versus b are provided in Figure C.3 of Appendix C.3; the limiting values under the pseudo-Poisson procedure can be calculated based on the expressions of $\mathbb{E}(U_1(\alpha_r, G^\star))$ and $\mathbb{E}(U_1(\beta, G^\star))$ from Section 4.4.1.1, with (4.4.6) plugged in for $\mu(s_r|X)$. As previously noted, $\beta_P^\star$ appears less sensitive to the determinants of the true process than $\beta_{GL}^\dagger$.

4.4.1.3 Mixed Poisson processes for recurrent events and death (GL-2)

In this section we consider random effects models as "data-generating frameworks" (Cook and Lawless, 2007, Section 6.6.2), and let $\gamma_U = \delta_U = 1$ and $\phi > 0$. If in addition $\rho_{k0} = \rho_0$ and $\lambda_{k0}(t) = \lambda_0(t)$, then the recurrent event process and the terminating event process are independent conditional on the treatment group X and the random effect U. Given X and U, $\{N(t), 0 < t\}$ follows a time-inhomogeneous Poisson process with intensity function

$$\lambda_{(k-1)k}(t|\mathcal{H}(t)) = U\lambda_0(t)\exp(\gamma X) = \lambda(t|X, U) \ ,$$

and the terminating event (i.e. death) has a proportional intensity function of the form

$$\lambda_{kD_k}(t|\mathcal{H}(t)) = U\rho_0\exp(\delta X) = \rho(t|X, U) \ .$$

We derive the limiting values under this data-generation process shortly, but first provide insights into the forms of the following observable quantities (i) $\mathbb{E}(dN(t)|T \geq t, X)$, (ii) $\mathbb{E}(dN(t)|Z(s), 0 < s < t, X)$ and (iii) $\mathbb{E}(dN_D(t)|Z(s), 0 < s < t, X)$.

The conditional rate function (i) can be derived as

$$\begin{aligned}\mathbb{E}(dN(t)|T \geq t, X) &= \int_0^\infty \mathbb{E}(dN(t)|T \geq t, U = u, X)g_U(u|T \geq t, X)du \\ &= \lambda_0(t)e^{\gamma X}\mathbb{E}(U|T \geq t, X) \ ,\end{aligned}$$

where $dG_U(u|T \geq t, X) = g_U(u|T \geq t, X)du$ is the conditional density function of $U|T \geq t, X$ and

$$\begin{aligned}\mathbb{E}(U|T \geq t, X) &= \int_0^\infty uP(T \geq t, U = u|X)P(T \geq t|X)du \\ &= \int_0^\infty ug_U(u)\frac{S(t|X, U = u)}{S(t|X)}du \ ,\end{aligned}$$

since $X \perp\!\!\!\perp U$ due to randomization. To simplify notation in the derivations that follow, we let $\bar{N}(t) := \{dN(s), 0 < s < t\}$. If $\widetilde{\mathcal{H}}(t) = \{Z(s), 0 < s < t, X\} = \{\mathbb{1}(T \geq s), \bar{N}(t), X\} = \mathcal{H}(t) \setminus \{U\}$ denotes the process history up to time t^- excluding U, the intensity function (ii) for a non-fatal event is

$$\begin{aligned}\mathbb{E}(dN(t)|\widetilde{\mathcal{H}}(t)) &= \mathbb{E}_U\big(\mathbb{E}\big[dN(t)|\widetilde{\mathcal{H}}(t), U\big] \mid \widetilde{\mathcal{H}}(t)\big) = \mathbb{E}_U\big(\mathbb{E}\big[dN(t)|\mathcal{H}(t)\big] \mid \widetilde{\mathcal{H}}(t)\big) \\ &= \lambda_0(t)e^{\gamma X}\mathbb{E}(U \mid \bar{N}(t), T \geq t, X)\ ,\end{aligned}$$

where

$$\mathbb{E}(U \mid \bar{N}(t), T \geq t, X) = \int_0^\infty u g_U(u \mid \bar{N}(t), T \geq t, X)du$$

for

$$\begin{aligned}g_U(u \mid \bar{N}(t), T \geq t, X) &= P(U = u \mid \bar{N}(t), T \geq t, X) = \frac{P(\bar{N}(t), U = u \mid T \geq t, X)}{P(\bar{N}(t) \mid T \geq t, X)} \\ &= \frac{P(\bar{N}(t) \mid T \geq t, U = u, X)P(U = u \mid T \geq t, X)S(t|X)}{P(\bar{N}(t), T \geq t \mid X)} \\ &= \frac{P(\bar{N}(t),\ \mid T \geq t, U = u, X)S(t|X, U = u)g_U(u)}{P(\bar{N}(t), T \geq t \mid X)} \\ &= \frac{P(\bar{N}(t) \mid X, U = u)S(t|X, U = u)g_U(u)}{\mathbb{E}_U\big(P(\bar{N}(t) \mid X, U)S(t|X, U)\big)}\ .\end{aligned}$$

The last equality follows from the fact that $\{N(t), t > 0\} \perp\!\!\!\perp \{N_D(t), t > 0\} \mid X, U$. The death intensity (iii) can be likewise derived as

$$\begin{aligned}\mathbb{E}(dN_D(t)|\widetilde{\mathcal{H}}(t)) &= \frac{P(T \in [t, t + dt), \bar{N}(t) \mid X)}{P(T \geq t, \bar{N}(t) \mid X)} \\ &= \frac{\int_0^\infty P(T \in [t, t + dt) \mid X, U = u)P(\bar{N}(t) \mid X, U = u)g_U(u)du}{\int_0^\infty P(T \geq t \mid X, U = u)P(\bar{N}(t) \mid X, U = u)g_U(u)du}\ .\end{aligned}$$

Figure 4.6 contains plots of the quantities (i)-(iii) for four hypothetical sample paths $\big(\mathcal{H}_j(\infty)\big)_{j=1,\ldots,4}$ with varying number of recurrent non-fatal events. To help understand Figure 4.6, we illustrate $\mathbb{E}(U|\bar{N}(t), X)$ versus t in Figure 4.7 for different values of $\phi = \text{Var}(U)$.

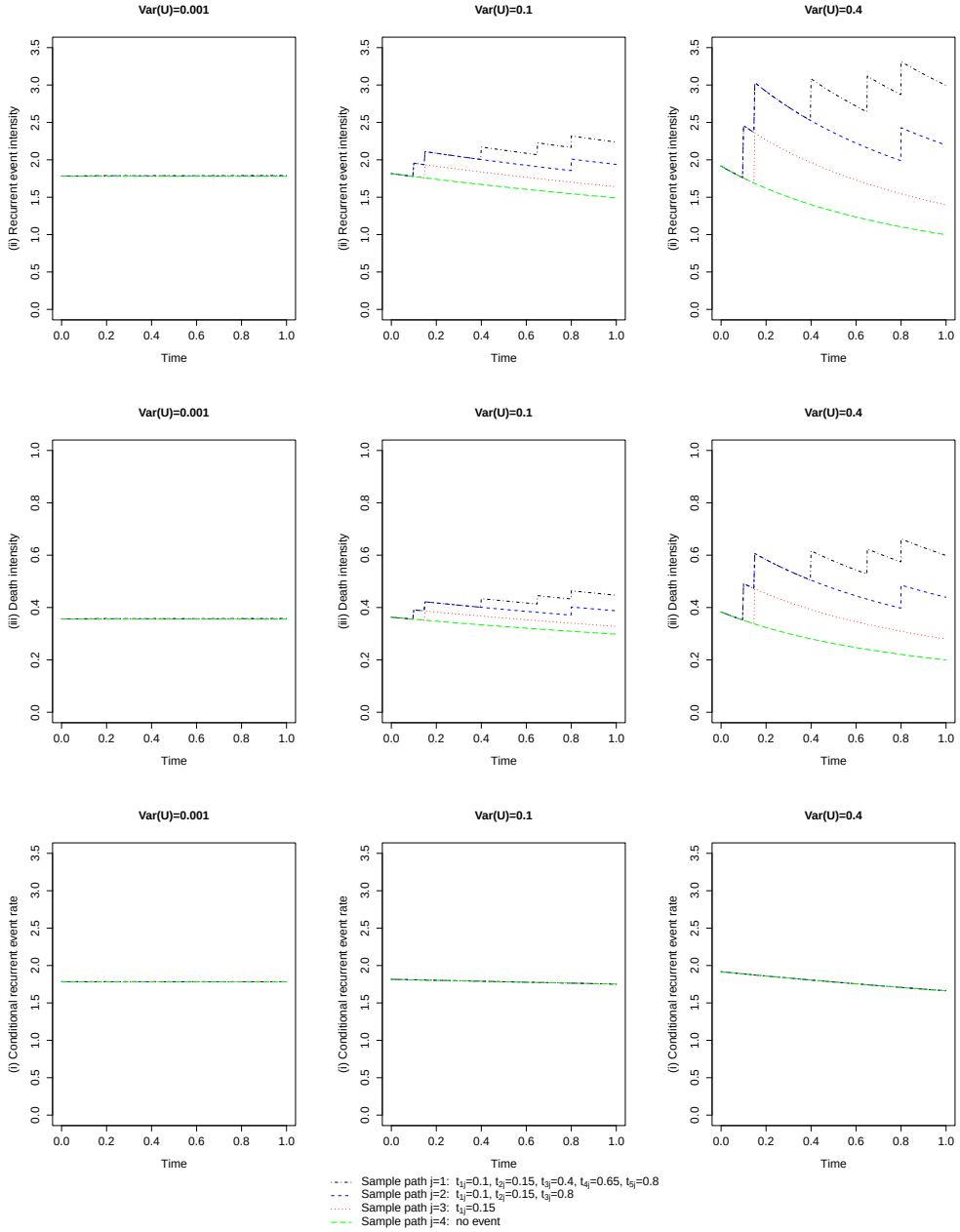


Figure 4.6: *GL-2*: Plots of the quantities (ii) $\mathbb{E}(dN(t)|\widetilde{\mathcal{H}}(t))$ (top panels), (i) $\mathbb{E}(dN(t)|T \geq t, X)$ (bottom panels) and (iii) $\mathbb{E}(dN_D(t)|\widetilde{\mathcal{H}}(t))$ (middle panels) for $X = 0$ and $\text{Var}(U) = 0.001, 0.1, 0.4$ when $P(T \leq \tau|X = 0) = 0.3$ and $\mathbb{E}(N(\tau)|X = 0) = 1.5$; $\tau = 1, \kappa = 1$.

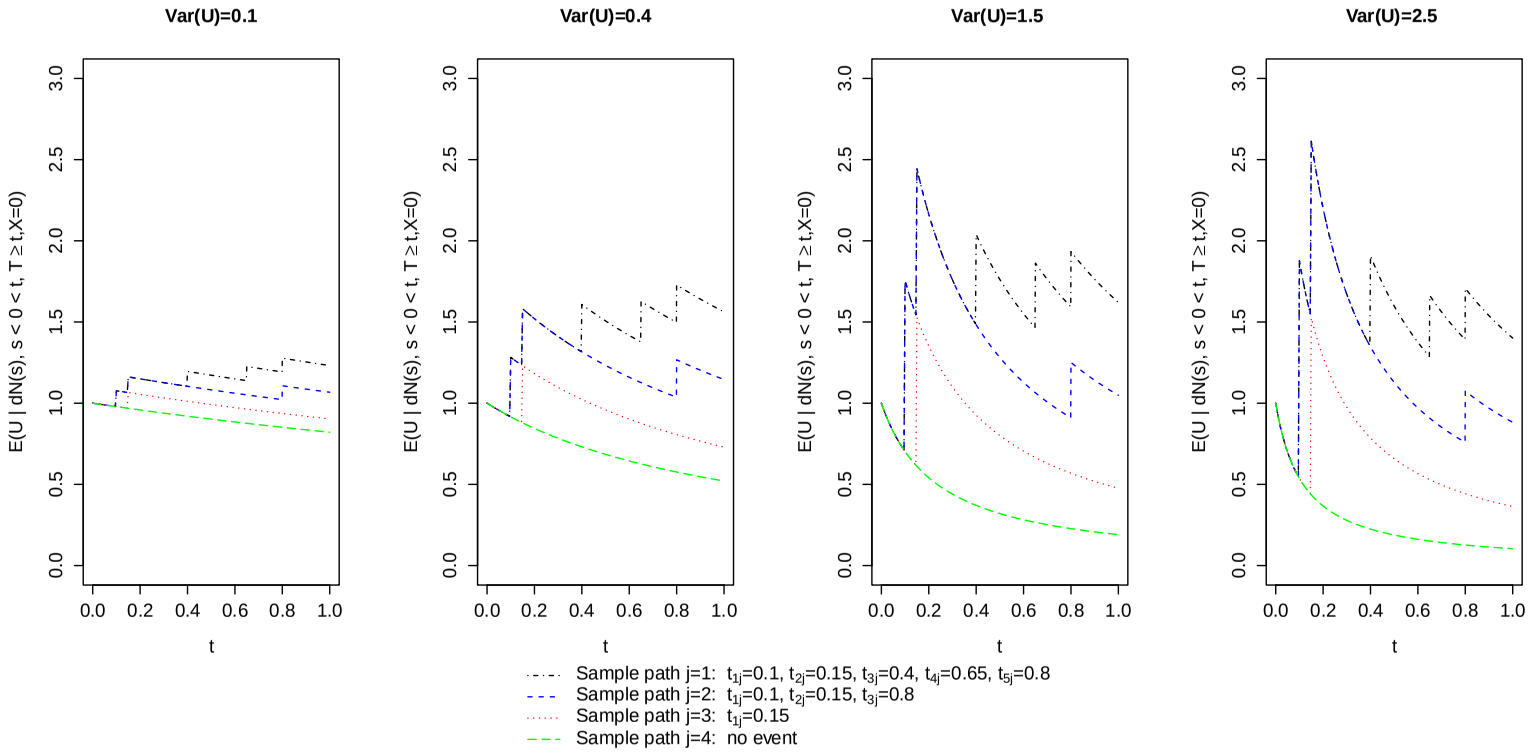


Figure 4.7: *GL-2*: Plots of $\mathbb{E}(U|\bar{N}(t), T \geq t, X = 0)$ versus time t for different values of Var(U); $\tau = 1, \kappa = 1, P(T \leq \tau|X = 0) = 0.3$ and $\mathbb{E}(N(\tau)|X = 0) = 1.5$.

The plots show that the intensity for recurrent non-fatal events increases immediately upon event occurrence, and declines over time between events; as more time passes since the last event the intensity may even drop below the population rate. Moreover, the extent of the jump and the rate of the decline between events are greater when ϕ is greater. This feature is characteristic of mixed Poisson models. The death intensities in the second row of Figure 4.6 also exhibit sharp increases when non-fatal events occur, a desirable feature if models aim to reflect transient increase in risk. Figure 4.7 shows why this is the case: both intensities are governed by the conditional expectation of the random effect U given the process history.

We next study the impact of ϕ on $\beta^\star_{GL}$ and $\beta^\star_P$. If $\{Z(u), 0 < u\} \perp\!\!\!\perp C_r|X, U$, $C_r \perp\!\!\!\perp U|X$ and $G^\star = G$, $\beta^\star_{GL}$ can be obtained using the following expectations ($l = 0, 1$):

$$\begin{aligned} s^{(l,\star)}(t,\beta) &= \mathbb{E}_{X|U}\big(X^l \exp(\beta X)\big) = \sum_{x=0,1} P(X=x)x^l \exp(\beta x)\ , \\ s^{(l,\star)}(t) &= \mathbb{E}\big(X^l S(t|X,U)\mathbb{E}(dN(t)|T \geq t, X, U)\big) \\ &= \sum_{x=0,1} P(X=x)x^l \lambda_0(t)\exp(\gamma x)\mathbb{E}_U\big(US(t|X=x,U)\big)\ . \end{aligned}$$

Similarly, $\beta^\star_P$ can be obtained by solving the following equation system:

$$\begin{aligned} \mathbb{E}(U_1(\alpha_r, G^\star)) &= \sum_{x=0,1} \frac{P(X=x)h'(\alpha_r+\beta x)}{h(\alpha_r+\beta x)}\big(\mathbb{E}_U(\mu(s_r|X=x,U)) - h(\alpha_r+\beta x)\big)\ , \\ \mathbb{E}(U_1(\beta, G^\star)) &= \sum_{r=1}^{R} \frac{P(X=1)h'(\alpha_r+\beta)}{h(\alpha_r+\beta)}\big(\mathbb{E}_U(\mu(s_r|X=1,U)) - h(\alpha_r+\beta)\big)\ , \end{aligned}$$

where $d\mu(s_r|X,U) = r(s_r|X,U)S(s_r|X,U) = \mathbb{E}(dN(s_r)|T \geq s_r, X, U)S(s_r|X,U)$.

We assume a Weibull baseline function $\lambda_0(t)$ as before and, for fixed values of κ and ϕ, set λ_0 and ρ_0 so that $P(T \leq \tau|X=0) = \mathbb{E}_U(P(T \leq \tau|X=0,U)) \in (0, 0.6]$ and $\mathbb{E}(N(\tau)|X=0) = \mathbb{E}_U(\mathbb{E}(N(\tau)|X=0,U)) = 1.5, 2.5$ and 3.5. Figure 4.8 contains contour plots depicting the dependency of $\exp(\beta^\star_{GL})$ and $\exp(\beta^\star_P)$ on $P(T \leq \tau|X=0)$ and $\exp(\delta)$ for $\phi = 0.1$ and 0.4; the dashed gray lines in the left panels correspond to the setting $\phi = 0$, which is equivalent to the scenario *GL-1-1* in Section 4.4.1.1. The higher the variance of the random effect U, the more attenuated the effect of treatment on the cumulative mean function for non-fatal events.

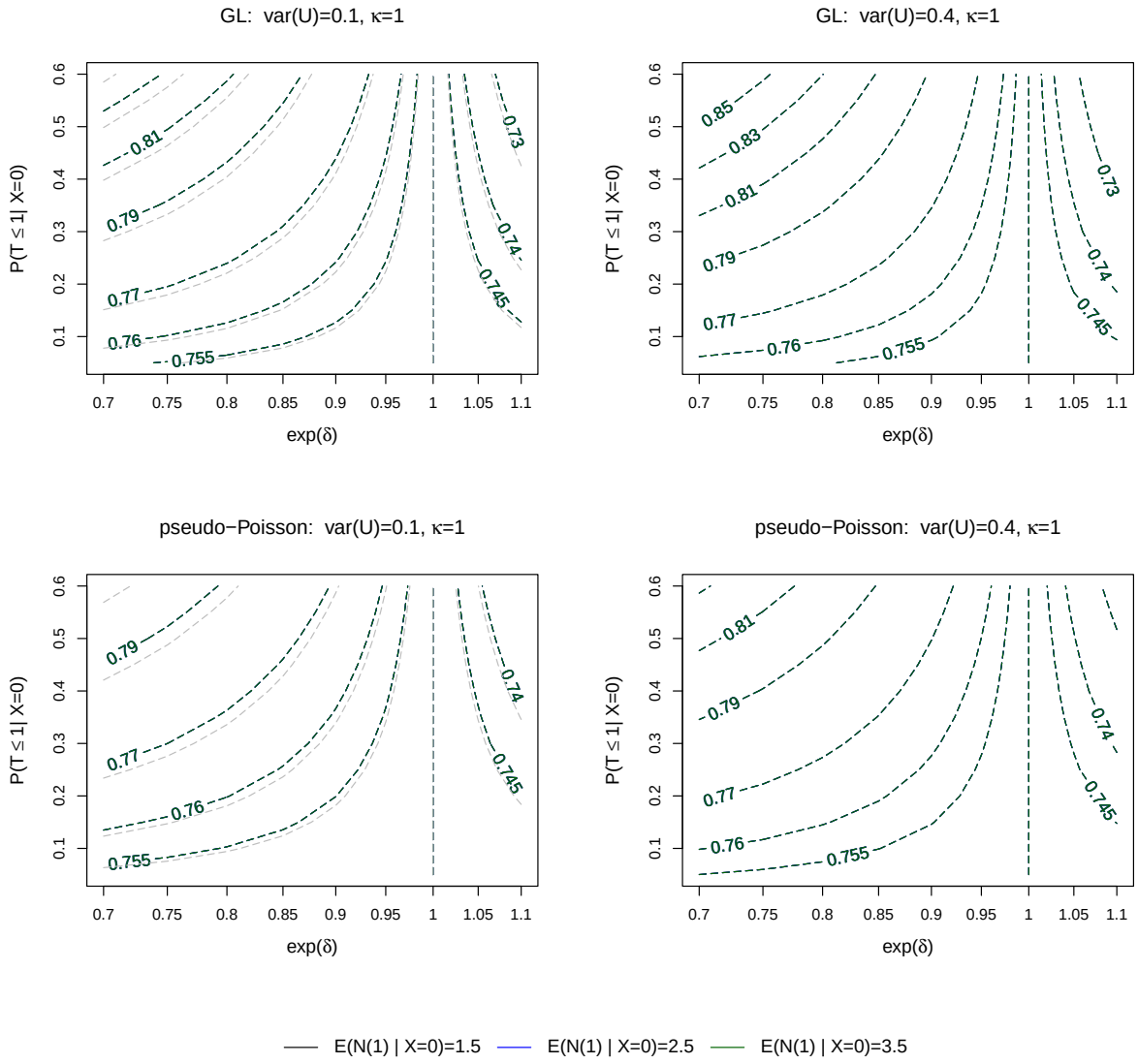


Figure 4.8: *GL-2:* Contour plots showing the dependence of the limiting values $\exp(\beta^{\star}_{GL})$ (upper panels) and $\exp(\beta^{\star}_{P})$ (lower panels) on $\exp(\delta)$ and $P(T \leq \tau | X = 0)$ for $\phi = \text{Var}(U) = 0.1$ (left panels) and 0.4 (right panels) when the true process has intensities $\lambda(t|X, U) = U\lambda_0(t)\exp(\gamma X)$ and $\rho(t|X, U) = U\rho_0 \exp(\delta X)$ for $\exp(\gamma) = 0.75$ and $\kappa = 1$, $G^{\star} = G$.

4.4.2 The Mao-Lin estimator

This section builds on the work of Toenges et al. (2021) and studies the limiting values of the Mao-Lin estimator in settings where disease-related death is part of the composite event process $\{N^\dagger(t) = N(t) + N_D(t), 0 < t\}$, but non-disease related death is not (Furberg et al., 2022). In addition to the $k-1 \longrightarrow k$ and $k \longrightarrow D_k$ intensities within framework (4.4.1)-(4.4.2), we assume the $k \longrightarrow D'_k$ intensities in Figure 1.1(f) to take the form

$$\lambda_{kD'_k}(t|\mathcal{H}(t)) = \psi(t|X) = \psi_0 \ , \qquad k = 0, 1, \dots \ . \tag{4.4.7}$$

We consider the simplest data-generating mechanism and let $\gamma_U = \delta_U = 0, \phi = 0$, $\rho_{k0} = \rho_0$ and $\lambda_{k0}(t) = \lambda_0(t) = \lambda_0^\kappa \kappa t^{\kappa-1}$ so that

$$\begin{aligned}
\lambda_{(k-1)k}(t|\mathcal{H}(t)) &= \lambda_0(t)\exp(\gamma X) = \lambda(t|X) \\
\lambda_{kD_k}(t|\mathcal{H}(t)) &= \rho_0 \exp(\delta X) = \rho(t|X) \\
\lambda_{kD'_k}(t|\mathcal{H}(t)) &= \psi_0 = \psi(t|X) \ .
\end{aligned}$$

Since $r^\dagger(t|X) = \lambda(t|X) + \rho(t|X)$ and $S(t|X) = \exp(-(\rho_0 e^{\delta X} + \psi_0)t)$ under these restrictions, the mean function for the composite event process $\{N^\dagger(t), 0 < t\}$ is given by

$$\mu^\dagger(t|X) = \int_0^t \big(\lambda_0^\kappa \kappa u^{\kappa-1} e^{\gamma X} + \rho_0 e^{\delta X}\big)\exp\big(-(\rho_0 e^{\delta X} + \psi_0)u\big)du \ , \tag{4.4.8}$$

leading to a mean function ratio of

$$\frac{\mu^\dagger(t|X=1)}{\mu^\dagger(t|X=0)} = \frac{\int_0^t (\lambda_0^\kappa \kappa u^{\kappa-1} e^{\gamma} + \rho_0 e^{\delta})\exp\big(-(\rho_0 e^{\delta} + \psi_0)u\big)du}{\int_0^t (\lambda_0^\kappa \kappa u^{\kappa-1} + \rho_0)\exp\big(-(\rho_0 + \psi_0)u\big)du} \ . \tag{4.4.9}$$

We now explore the adequacy of the ML model in more detail and consider the LEADER trial of Section 1.3.1 for illustration. To reflect the LEADER trial results found in Furberg et al. (2022), we took $\tau = 5$ and set $\lambda_0 = 0.023$, $\rho_0 = 0.016$ and $\psi_0 = 0.010$, giving $P(T \le \tau|X = 0) = 0.12, P(T \le \tau, T_D < T_{D'}|X = 0) = 0.074, P(T_D < T_{D'}|T \le \tau, X = 0) = 0.607$ and $\mathbb{E}(N(\tau)|X = 0) = 0.11$; see Figures 1 and 5 in Furberg et al. (2022). Treatment effects on the recurrent MI intensity and the hazard for time to CV death were observed to be roughly $e^\gamma = 0.85$ and $e^\delta = 0.8$; there was no treatment effect observed on non-CV death. In numerical

investigations that follow we refer to this parameter setting as the LEADER trial scenario. In addition to this scenario, we further set $e^{\delta} = 1.0$ to represent the case where treatment does not affect the CV death intensity, and determined λ_0, ρ_0 and ψ_0 so that $P(T \leq \tau | X = 0) = 0.12$ and 0.3, $P(T_D < T_{D'} | T \leq \tau, X = 0) = 0.4, 0.61$ and 0.8, and $\mathbb{E}(N(\tau)|X = 0) = 0.11$ or 1.1. Figure 4.9 displays plots of $\mu^{\dagger}(t|X = 1)/\mu^{\dagger}(t|X = 0)$ based on (4.4.9) when $\kappa = 1$. Compared to the GL model within the framework *GL-1-1*, adequacy of the ML model additionally varies with $\mathbb{E}(N(\tau)|X = 0)$ through λ_0. When treatment affects CV death the model is misspecified, but the maximum deviation of $\mu^{\dagger}(t|X = 1)/\mu^{\dagger}(t|X = 0)$ from constancy over the observation period is only about 2.7% $(= 100(0.838 - 0.816)/0.816$ or $100(0.864 - 0.841)/0.841)$, observed for the scenarios with $\delta = \log(0.8), P(T \leq \tau | X = 0) = 0.3$ and $P(T_D < T_{D'} | T \leq \tau, X = 0) = 0.8$. When $\delta = 0$ and $\kappa = 1$, the mean function ratio (4.4.9) is equal to $(\lambda_0 e^{\gamma} + \rho_0)/(\lambda_0 + \rho_0)$ and thus constant over time; see right panels of Figure 4.9. If $\lambda_0(t)$ varies with time (i.e. when $\kappa \neq 1$), the proportional means assumption does no longer hold. On the other hand, when $\delta = 0$, the mean ratio (4.4.4) for $\mu(t)$ under the GL model is constant over time regardless of whether $\lambda(t)$ or $\rho(t)$ are time-homogeneous or not. Validity of the ML model and its one-dimensional treatment effect thus requires the recurrent and terminal event process to be time-homogeneous, whereas validity of the GL model and estimand does not.

If $\{Z(t), 0 < t\} \perp\!\!\!\perp C_r | X$ and $G^{\star} = G$, expectations needed to calculate the limiting value $\beta_P^{\dagger\star}$ are

$$\mathbb{E}(U_1(\alpha_r^{\dagger}, G^{\star})) = \sum_{x=0,1} \frac{P(X = x) h'(\alpha_r^{\dagger} + \beta^{\dagger} x)}{h(\alpha_r^{\dagger} + \beta^{\dagger} x)} \big(\mu^{\dagger}(s_r | X = x) - h(\alpha_r^{\dagger} + \beta^{\dagger} x) \big) \, ,$$

$$\mathbb{E}(U_1(\beta^{\dagger}, G^{\star})) = \sum_{r=1}^{R} \frac{P(X = 1) h'(\alpha_r^{\dagger} + \beta^{\dagger})}{h(\alpha_r^{\dagger} + \beta^{\dagger})} \big(\mu^{\dagger}(s_r | X = 1) - h(\alpha_r^{\dagger} + \beta^{\dagger}) \big) \, ,$$

with $\mu^{\dagger}(s_r|X)$ as given in (4.2.7). For the ML estimation method of Section 4.3.1, the limiting value $\beta_{ML}^{\dagger\star}$ can be computed based on the following expectations $(l = 0, 1)$:

$$s^{(l,\star)}(t, \beta^{\dagger}) = \mathbb{E}_X \big(X^l \exp(\beta^{\dagger} X) \big) = \sum_{x=0,1} P(X = x) x^l \exp(\beta^{\dagger} x)$$

$$\begin{aligned} s^{(l,\star)}(t) &= \mathbb{E}_X \big(X^l S(t|X) \mathbb{E}(dN^{\dagger}(t) | T \geq t, X) \big) \\ &= \sum_{x=0,1} P(X = x) x^l S(t|X = x) r^{\dagger}(t|X = x) \, . \end{aligned}$$

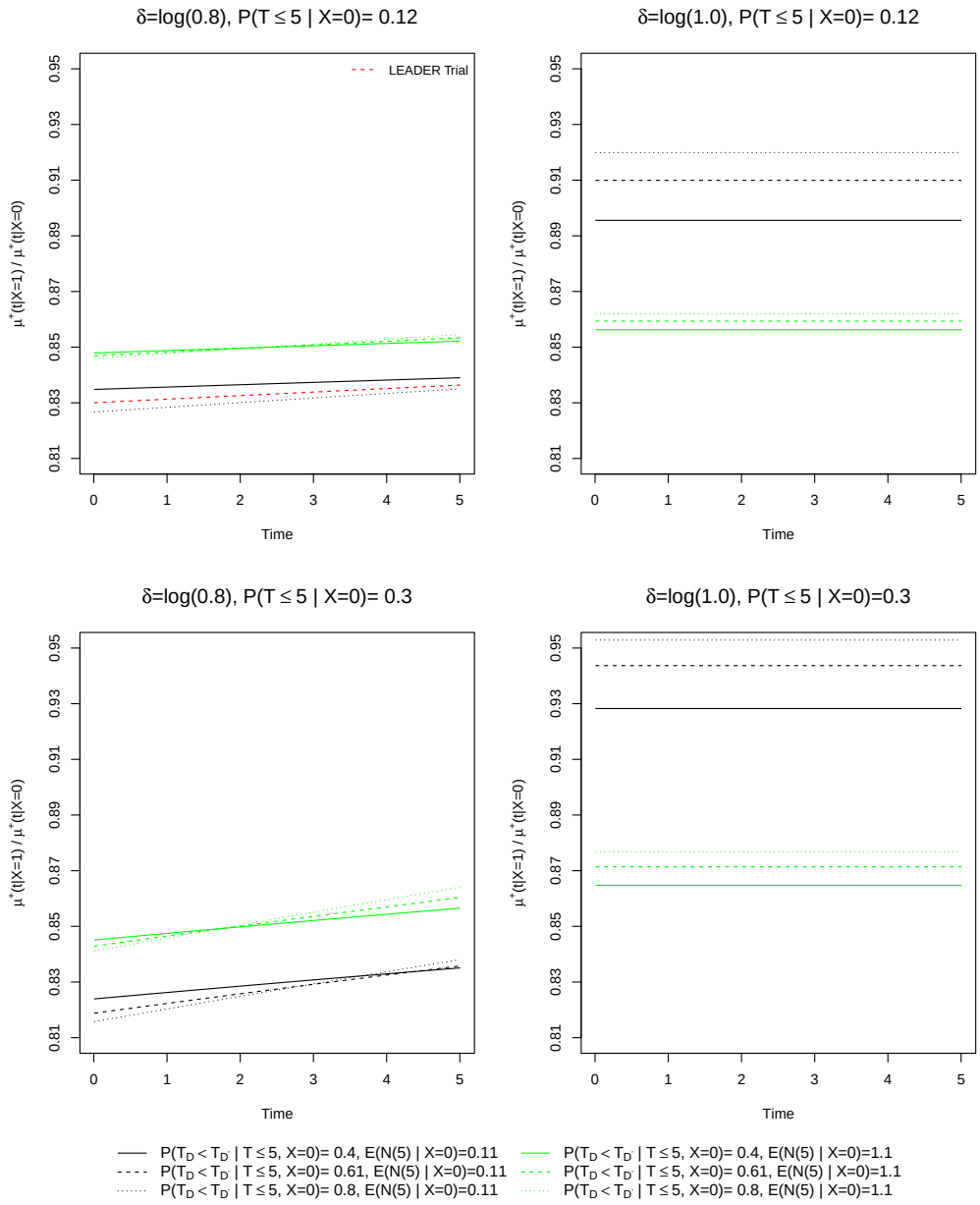


Figure 4.9: *ML-1-1:* Plots of $\mu^{\dagger}(t|X=1)/\mu^{\dagger}(t|X=0)$ for $\exp(\gamma) = 0.85$ and different values of $P(T \leq \tau|X=0), P(T_D < T_D'|T \leq \tau, X=0), \delta$ and $\mathbb{E}(N(\tau)|X=0)$ when the true process has intensities $\lambda(t|X) = \lambda_0 \exp(\gamma X), \rho(t|X) = \rho_0 \exp(\delta X)$ and $\psi(t|X) = \psi_0$; the red line in the upper left panel corresponds to the parameter settings in the LEADER trial (Marso et al., 2016); $\tau = 5, \kappa = 1$.

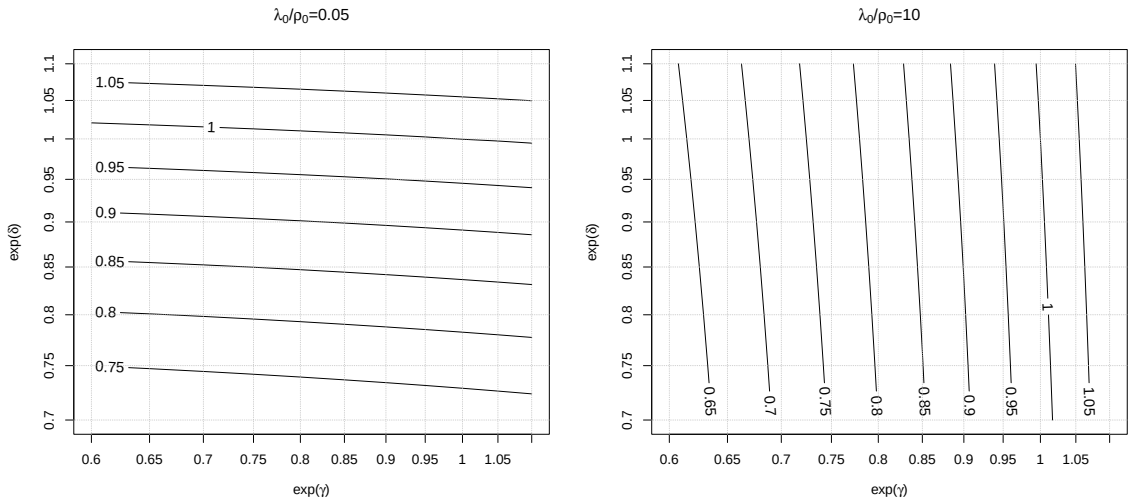
Assuming $P(X = 1) = 0.5$, we obtain

$$\beta_{ML}^{\dagger\star} = \log\left(\frac{\int_0^\tau \left[\lambda_0(u)e^\gamma + \rho_0 e^\delta\right] \exp(-(e^\delta\rho_0 + \psi_0)u)du}{\int_0^\tau \left[\lambda_0(u) + \rho_0\right] \exp(-(\rho_0 + \psi_0)u)du}\right). \tag{4.4.10}$$

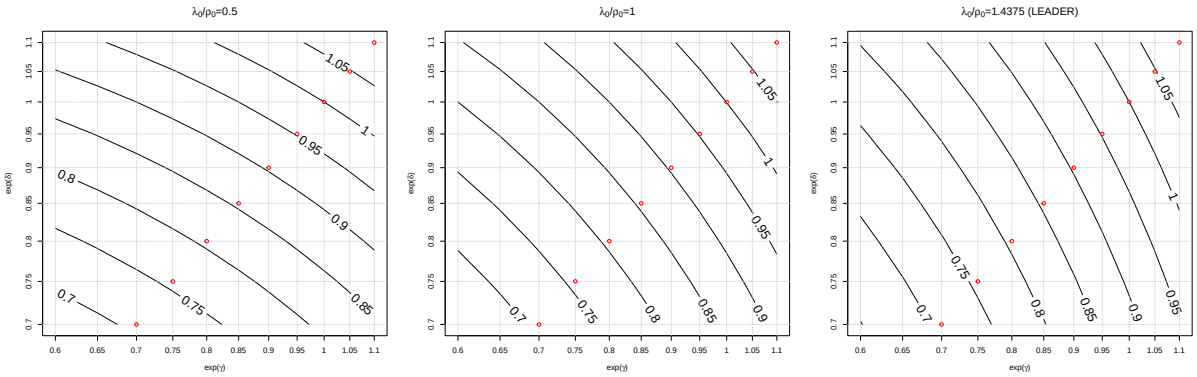
Figure 4.10 contains contour plots depicting the dependence of the limiting value $\exp(\beta_{ML}^{\dagger\star})$ on treatment and the process intensities for the case where the adopted censoring model is correctly specified ($G^\star = G$) and $\kappa = 1$; see Figure C.4 of Appendix C.3.2 for the pseudo-Poisson approach. Here we took $\rho_0 = 0.016$ and $\psi_0 = 0.010$, and determined λ_0 so that $\lambda_0/\rho_0 = 0.05, 0.5, 1, 1.4375$ (LEADER) and 10.

The estimand of the treatment effect estimator in the ML model (4.2.8) is interpretable only in some particular settings. First, the left panel of Figure 4.10(a) illustrates that $\exp(\beta_{ML}^{\dagger\star}) \approx \exp(\delta)$ if the non-fatal event rate is close to 0 (see horizontal contours); the right panel of Figure 4.10(a) indicates that $\exp(\beta_{ML}^{\dagger\star}) \approx \exp(\gamma)$ if the mortality rate is close to 0 (see vertical contours). Second, we see from the panels in Figure 4.10(b) where λ_0/ρ_0 takes more realistic values, that $\exp(\beta_{ML}^{\star}) = \exp(\gamma) = \exp(\delta)$ if $\gamma = \delta$, λ_0/ρ_0 close to 1, and γ, δ close to 0. In all other scenarios, $\exp(\beta_{ML}^{\dagger\star})$ is a function of the intensities of the underlying process including the treatment effects.

We can draw the following conclusions from the numerical investigations in this chapter. The GL estimand can be interpretable when the mortality rate is relatively small and the treatment effect on the death intensity is negligible; in this case the GL estimand is close to γ, the treatment effect on the non-fatal event intensity. The ML estimand is, however, less easy to interpret. The calculations and plots show that the ML estimand is a complex function of the intensities for the recurrent events and death as well as the treatment effects on the respective events and their association. As a result it is difficult to characterize the implications to patients of using the experimental treatment over the control treatment; the effects on event risk and death are conflated since it is based on a composite of a recurrent and terminal event.



(a) Settings where the Mao-Lin estimand is easy to interpret.
Left panel: $\lambda_0/\rho_0 \approx 0$; right panel: $\lambda_0/\rho_0 \nearrow \infty$.



(b) Settings where the Mao-Lin estimand is less easy to interpret.
Left panel: $\lambda_0/\rho_0 = 0.5$; middle panel: $\lambda_0/\rho_0 = 1$ and right panel: $\lambda_0/\rho_0 \approx 1.5$.

Figure 4.10: *ML-1-1:* Plots showing the dependency of the limiting values $\exp(\beta^{\dagger\star}_{ML})$ on treatment and the time-homogeneous process intensities $\lambda(t|X) = \lambda_0 \exp(\gamma X), \rho(t|X) = \rho_0 \exp(\delta X)$ and $\psi(t|X) = \psi_0$; $\kappa = 1, \tau = 5, P(T \leq \tau|X = 0) = 0.12, G^\star = G$; the red circles indicate the points on the grid where $\exp(\gamma) = \exp(\delta)$.

4.5 Prevention of recurrent skeletal complications in metastatic breast cancer

Women with advanced stage IV breast cancer often have skeletal metastases, which can cause fractures, bone pain and other skeletal complications (Coleman, 2001, 2006). Here we consider analysis of data from the palliative trial introduced in Section 1.3.3 and assess whether the bisphosphonate pamidronate is effective in preventing recurrent skeletal-related events (SREs) in these women, as compared to placebo (Hortobagyi et al., 1996). We estimate the effect of pamidronate on the expected number of SREs by fitting the Ghosh-Lin (GL) model to the data. Adequacy of the GL model will be assessed by comparing the resulting semiparametric estimates of the cumulative mean function, $\widehat{\mu}(t|X) = \widehat{\mu}(t)\exp(\widehat{\beta}X)$, with their nonparametric counterparts, including the Aalen-Johansen (AJ) estimates of (4.2.2) and Cook-Lawless (CL) estimates of (4.2.4) based on the multistate model in Figure 1.1(e). Any discrepancy between the semiparametric and the nonparametric estimates may indicate model inadequacy.

Over the 24 month follow-up period, 125 of the n=195 placebo patients had at least one skeletal event, 84 had at least two events, 49 had at least three events and 31 had at least four events; in the pamidronate arm (n=185) these numbers were 85, 46, 29 and 17, respectively; here K, the upper limit on the number of recurrent non-fatal events that can occur, was set to 9. There were 28 deaths in the placebo arm and 34 in the pamidronate arm. Plots of the Nelson-Aalen (NA) estimates of Markov cumulative intensity functions for the occurrence of new skeletal complications, death and censoring stratified by $N(t^-)$ are displayed in Figure 4.11 for placebo patients. There is an increasing risk of both a new skeletal complication and death with the occurrence of each skeletal event. The plots of the cumulative intensities for censoring with no, one, two or three SREs reveal some degree of event-dependent censoring, as patients with a higher number of skeletal complications are at an increased risk of early withdrawal from the study. Figure 4.12 provides further evidence of event-dependent censoring in the placebo arm: the Kaplan-Meier (KM) and AJ estimates for the cumulative distribution function of the time to death begin to separate at 12 months; the former estimate is sensitive to event-dependent censoring while the latter is not. This in turn explains the differences in the unweighted versions of the KM-based and AJ-based CL estimates of $\mu(t|X = 0)$ in the left panel of Figure 4.13.

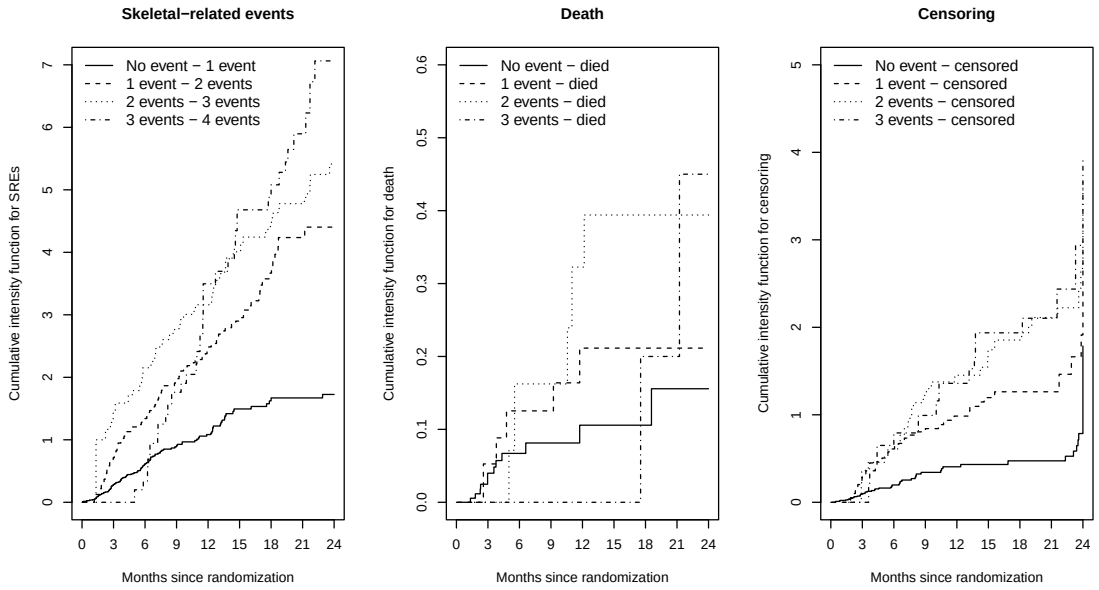


Figure 4.11: Nonparametric Nelson-Aalen estimates of the cumulative intensity functions for the occurrence of new skeletal-related events (left panel), death (middle panel) and censoring (right panel) for placebo patients in Hortobagyi et al. (1996).

We used the inverse probability of censoring weighting techniques in Cook et al. (2009) to protect against event- or state-dependent censoring in the estimation of the cumulative mean functions (4.2.2) and (4.2.4); see Datta and Satten (2002) for further details on this approach. Based on the findings from Figure 4.11, we adopt an intensity for the random censoring time C_r in a given treatment arm of the form

$$\begin{aligned}\lambda^r(t|\mathcal{H}_i(t)) &= \lim_{\Delta t \downarrow 0} \frac{P(t < C_{ri} \leq t + \Delta t \mid C_{ri} \geq t, T_i \geq t, N_i(t^-))}{\Delta t} \\ &= \lambda^r(t|N_i(t^-), T_i \geq t) = \sum_{k=0}^{K} \mathbb{1}(N_i(t^-) = k, T_i \geq t)\lambda_k^r(t) \ .\end{aligned}$$

For simplicity we suppress the dependence on X in the notation. The right panel of Figure C.5 in Appendix C.4 shows that there is also a slight increase in the censoring intensity in the pamidronate group with each additional skeletal-related event, which further motivates the use of this censoring model. The censoring intensity $\lambda_k^r(t)$ for individuals with k SREs can be estimated

non-parametrically via

$$\widehat{\lambda}_k^r(t) = \frac{\sum_{i=1}^n \mathbb{1}(C_{ri} = t, T_i \geq t, N_i(t^-) = k)}{\sum_{i=1}^n \mathbb{1}(C_{ri} \geq t, T_i \geq t, N_i(t^-) = k)}, \qquad k = 0, 1, \ldots, K \ .$$

Consequently, by product integration, we obtained the individual-specific censoring survival probabilities as

$$\widehat{G}_i(t) = \widehat{P}(C_{ri} > t | N_i(t^-), T_i \geq t) = \prod_{u \leq t} \big(1 - \widehat{\lambda}^r(u | N_i(u^-), T_i \geq t)\big) \ .$$

The weighted AJ and CL estimates of the mean function for SREs were calculated based on Eqs. (4.2)-(4.10) in Cook et al. (2009) and are given in the right panel of Figure 4.13, along with the semiparametric estimates implied by the GL model. We obtained $\widehat{\beta}$ and $d\widehat{\mu}_0(t)$ under the GL estimation procedure using the `coxph()` function with $\widehat{w}_i(t) = \mathbb{1}(C_{ri} \geq \min(T_i, t))/\widehat{G}_i(t)$ accommodated with the weights argument (Therneau, 2024) and robust standard errors. Simulation studies of Section 4.3.3 demonstrated the validity of the robust standard errors provided by `coxph()` (Therneau, 2024). Alternatively, the GL procedure is implemented in the `R` packages `mets` and `reReg` (Holst et al., 2016; Chiou et al., 2023). The corresponding `R` functions `recreg()` and `reReg()` do not allow for more general censoring models and so are not applicable to the present data example; specifically, they are not flexible enough to accommodate time-varying internal covariates in the censoring process. We obtained $\widehat{\beta}$ and $\widehat{\mu}_0(t) = \exp(\widehat{\alpha}_r(t))$ under the pseudo-Poisson procedures using the `geese()` function (Halekoh et al., 2006), with estimation subject to the constraints $\alpha_r = \alpha_{r-1} + \exp(\psi_r)$, $r = 2, \ldots, R$. Constrained estimation is required to ensure monotonicity of the baseline mean function (Sun and Kalbfleisch, 1995; Sun and Zhao, 2013).

The GL estimation procedure gives a treatment effect estimate of $\widehat{\beta} = -0.586$ (SE=0.174, 95% CI: $-0.927, -0.246$, p=0.000742). The pseudo-Poisson procedures yield slightly different treatment effect estimates with larger standard errors: $\widehat{\beta}_{wEE} = -0.558$ (SE=0.208, 95% CI: $-0.965, -0.151$) and $\widehat{\beta}_{wO} = -0.605$ (SE=0.266, 95% CI: $-1.126, -0.083$), respectively. The mean function estimates for SREs, $\widehat{\mu}(t|X) = \widehat{\mu}_0(t)\exp(\widehat{\beta}X)$, implied by the GL procedure are in close agreement with the weighted nonparametric AJ and CL estimates; see right panel of Figure 4.13. The plot exhibits the largest differences between the semi- and nonparametric estimates early on in the trial; this can be explained by the proportionality assumption inherent in the GL model (4.2.3). Overall, the GL model and the GL estimation procedures provide a satisfactory fit

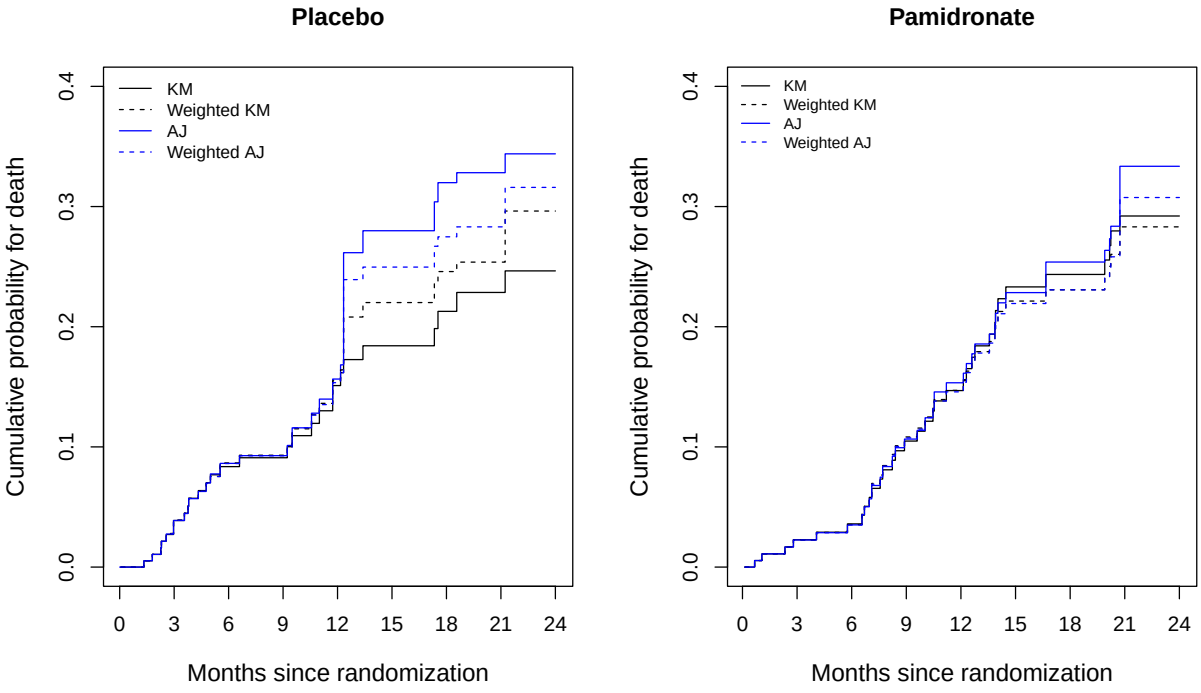


Figure 4.12: Unweighted and weighted nonparametric Kaplan-Meier and Aalen-Johansen estimates of the cumulative probability for death; IPC weighted estimates were calculated based on Eqs. (4.3)-(4.5) and (4.10)-(4.11) of Cook et al. (2009).

to the data. The pseudo-Poisson procedures offer less reasonable approximations to the data than the GL procedure.

The primary analysis suggests that pamidronate has a significant effect on the cumulative mean function for SREs compared to placebo. Figure 4.13 shows that control patients experience about 1.7 as many skeletal-related events as patients on pamidronate, with estimates of roughly $\widehat{\mu}(24|X = 1) = 1.85$ and $\widehat{\mu}(24|X = 0) = 3.19$. We complement the primary analysis with secondary intensity-based analyses to provide a fuller understanding of this marginal treatment effect. Plots of the nonparametric NA estimates of the cumulative intensity functions for one, two, three or four SREs are given in Figure C.6 of Appendix C.4. A Cox analysis of the effect of pamidronate on the time to the first SRE yields $\widehat{\gamma}_1 = -0.528$ (SE=0.141, 95% CI: $-0.251, -0.804$,

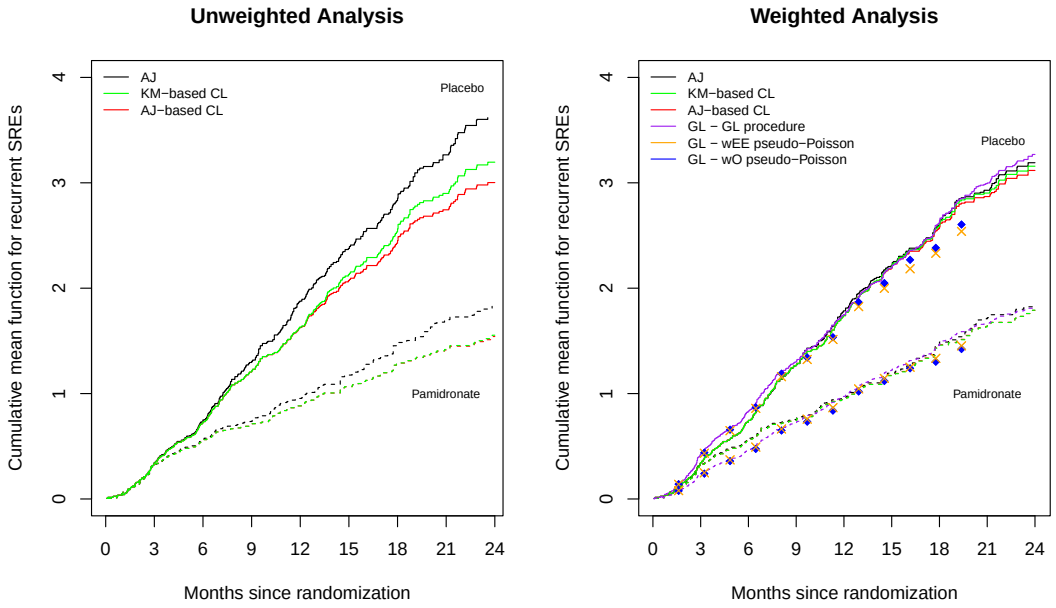


Figure 4.13: Unweighted and weighted estimates of the cumulative mean function for skeletal-related events (SREs); the weighted AJ and CL estimates were calculated based on Eqs. (4.10)-(4.12) of Cook et al. (2009), pseudo-Poisson estimation is based on R=12 equi-spaced time points on $(0, 21)$, $s_{12} = 19.4$.

p=0.000185, log-rank test: p=0.0002), so patients treated with pamidronate have a lower risk for a first SRE as compared to placebo patients; see left panel in upper row of Figure C.6. The right panel in the upper row indicates that patients in the placebo group have a roughly constant risk of a second skeletal-related event, following a first event. Patients on pamidronate who have already experienced one event are at the highest risk of a second event during the first 6 months of follow-up. The slope of the estimated $1 \longrightarrow 2$ cumulative intensity function is much lower beyond 6 months, and so the NA estimates of the two treatment groups appear to cross at about 10 months; a Cox analysis would not be suitable in that case because of the crossing intensities. Following a second event, there is a high risk of experiencing a third SRE in both treatment groups. Specifically, the NA estimates of the cumulative intensities for $2 \longrightarrow 3$ transitions in the two treatment groups begin to separate at about 9 months, with patients on pamidronate being at a

slightly higher risk later on in the trial than placebo patients; see left panel in the bottom row of Figure C.6. Following a third event, patients in the placebo group have a roughly constant risk of a fourth skeletal-related event; the corresponding risk for patients in the pamidronate group decreases over time.

An Andersen-Gill (AG) analysis of the effect of pamidronate on the conditional rate function for SREs, $r(t|X) = r_0(t)\exp(\xi X)$ (see Section 4.2.1), gives $\widehat{\xi} = -0.652$ (SE=0.090, 95% CI: $-0.475, -0.829$, p< 0.0001) for the unweighted approach and $\widehat{\xi} = -0.594$ (SE=0.175, 95% CI: $-0.250, -0.938$, p< 0.000714) for the weighted approach (Cook et al., 2009, Section 6.2). Figure 4.12 and Figure C.7 of Appendix C.4 indicate that pamidronate does not have any significant effect on mortality. An unweighted Cox analysis of the effect of pamidronate on the time to death gives $\widehat{\zeta} = 0.084$ (SE=0.256, 95% CI: $-0.417, 0.585$, p=0.742), whereas a weighted analysis yields $\widehat{\zeta} = -0.051$ (SE=0.304, 95% CI: $-0.646, 0.545$, p=0.868).

The secondary analyses demonstrate that the differences in $\widehat{\mu}(t|X = 1)$ and $\widehat{\mu}(t|X = 0)$ are mainly due to the reduction in the intensity for a first SRE (i.e. $0 \longrightarrow 1$ transition) by the use of pamidronate. The risk of a first SRE is fairly small compared to the risk of a second, third, ... SRE for patients experiencing a first, second, ... event, but the difference for the first SREs "drives" who is at risk for second and later events.

4.6 Discussion and recommendations

The proportional mean function models of Ghosh and Lin (2002) and Mao and Lin (2016) are routinely used in clinical trials involving recurrent and terminal events (Solomon et al., 2019; Vaduganathan et al., 2024). Recent advancements in the efficiency of associated estimators (Cortese and Scheike, 2022; Genetti et al., 2024) and power calculations used for trial design (Furberg et al., 2024) have further facilitated their use. This thesis chapter has provided important insights into the relationship between mean functions and process intensities, and into the consequences of model misspecification on the interpretability of treatment effects.

We considered various "data-generating frameworks" including intensity-based models and random effects models, and provided numerical studies to assess the adequacy of the Ghosh-Lin (GL) and Mao-Lin (ML) models under such true processes. The proportional means assumption is

satisfied only when the true model is rather simple, e.g. when there is no effect of treatment on the fatal event intensities in scenario GL-1-1; see Section 4.4.1.1. In general, recurrent and terminal event processes affected by treatment do not lead to mean functions of GL or ML form. Even if the ML model is satisfactory and adequately describes the observed data, it is not recommended for treatment comparisons in clinical trials because of the difficulty of interpretation. The GL model's estimand is more readily interpretable (but see below); the model should however be checked in any given setting.

Model adequacy can be easily assessed by model expansion. A test of $H_0 : \nu = 0$ in the expanded model $\mu(t|X) = \mu_0(t)\exp((\beta + \nu b(t))X)$ for a function $b(t)$ of time corresponds to testing whether the assumption of a constant treatment effect over time is adequate. Goodness of fit tests analogous to those discussed in Section 3.3.3 can be considered here. An alternative model checking procedure is to compare the model-based estimates of $\mu(t|X)$ with their nonparametric counterparts, as illustrated in the application of Section 4.5.

We demonstrated that the GL estimand $\beta^\star$ is a function of the degree of misspecification and the intensities of the underlying process including their treatment effects. This makes its interpretation challenging in certain situations and means that $\beta^\star$ on its own is insufficient for portraying the effect of the experimental treatment in randomized trials. We however support the use of GL models in trials (II) where the mortality rate is low and treatment has a fairly small effect on the death intensities, provided the GL analysis is supplemented with secondary intensity-based analyses of survival. An alternative strategy was proposed by Furberg et al. (2023).

Rather than considering a one-dimensional estimand, Furberg et al. (2023) proposed a two-dimensional estimand involving the expected number of non-fatal events via the GL model and the survival function for the time to death through a Cox model of the form (2.2.6), with estimation based on pseudo-observations (Overgaard et al., 2023). In CV trials where patients may die from CV and non-CV related causes, their approach can be extended to a three-dimensional estimand by considering two separate cloglog transformation models for the cumulative cause-specific death probabilities. Furberg et al. (2024) discussed a simulation-based procedure for power estimation based on the GL model for recurrent non-fatal events and a Cox model for the time to death.

Analyses of the expected number of recurrent non-fatal events and the time to death may not be sufficiently informative in disease processes involving non-fatal events with a duration. In

the HF-ACTION trial (O'Connor et al., 2009), for instance, heart failure patients are admitted and discharged from hospital repeatedly; while hospitalized, patients are naturally not at risk for another hospital admission. In the absence of terminal events, Lee and Cook (2019) investigated the impact of ignoring the non at-risk periods (i.e. keeping patients in the risk set while being in hospital) in semiparametric rate-based recurrent event analyses. Disease processes with so-called recurrent episodes and terminal events are in general better represented by the reversible illness-death model (see e.g. Cook and Lawless, 2018, Section 2.1, Andersen et al., 2019), with the time spent in the intermediate state at the kth occasion corresponding to the length of the kth hospital stay. A marginal feature of interest in this setting is the expected total number of days spent in hospital; we revisit this example in Section 6.3.2.

Chapter 5

Methods based on utility functions and ranked outcomes

5.1 Introduction

In clinical trials involving complex disease processes individuals are typically at risk of several types of events. In trials of patients with metastatic cancer, for instance, individuals are at risk of cancer progression, skeletal metastases, related complications, and death (Hortobagyi et al., 1996; Theriault et al., 1999). The NASCET trial evaluating the effectiveness of a surgical intervention in patients with carotid stenosis may record times of events such as strokes ipsilateral to the surgical site, contralateral strokes and death (Barnett et al., 1998). Traditional hazard-based models for the time to a particular single-type event are generally not aligned with causal statements (Hernán, 2010; Aalen et al., 2015). Moreover, it is hard to pick a single endpoint when interest lies in preventing each of the recorded adverse events. Use of multiple endpoints raises multiplicity issues and makes trial design more difficult (EMA, 2017). Composite endpoints that combine several event types of interest are widely adopted in cardiovascular (CV) trials, but interpretability of estimands in such analyses is problematic (Wu and Cook, 2012). For example, diabetic patients in the LEADER trial (Marso et al., 2016) may experience non-fatal myocardial infarction (MI), non-fatal stroke, and CV-related death. An analysis of the time to a first stroke, MI or CV death treats all components of the composite endpoint as of equal importance, but they clearly differ in

severity and clinical relevance. In addition, treatment may affect each component in a different way. Another longstanding concern in the use of composite endpoints is that less important events may exert undue influence in the evidence of the treatment effect. This is also an issue with rate and mean function analyses of recurrent and terminal events. Some such methods make use of all non-fatal and fatal events occurring over the course of follow-up, but the Mao and Lin (2016) model yields estimands with undesirable interpretations; see Chapter 4. The time-to-event and count frameworks (1.2.6) for estimand specification are often inadequate in type (I) trial settings. Recall from Section 1.1.1 that the focus in type (I) trial settings is on both death and other relevant non-fatal outcomes (i.e. on more than one type of event).

Over recent years there has been a strong desire to base efficacy evaluations in type (I) trial settings on individuals' entire trajectories, while taking severity and clinical relevance of different disease stages and conditions into account; see Evans and Follmann (2016) who aim "to use outcomes to analyze patients rather than patients to analyze outcomes". In the IBCSG breast cancer trial (Ludwig Breast Cancer Study Group, 1988), for instance, it is not only relevant whether short- or long-term chemotherapy prolongs survival, but also how each chemotherapy program affects time spent in different health states to assess the impact on quality of life. The marginal estimands discussed earlier in this thesis (see Chapters 3 and 4) do not address this. Moreover, quality of life accumulates over time and should be assessed based on the totality of an individual's disease course. Gelber et al. (1989) point out that analysis strategies in such settings aiming to provide a one-dimensional estimand require some form of ranking or weighting. The purpose of ranking approaches is to allow comparisons of treatment and control subjects, with estimands defined in terms of the probability that a treated subject has a more favorable sample path (or process history $\mathcal{H}(\infty)$) than a control subject.

Various types of ranking and weighting methods have been proposed in the literature for different disease process settings. These methods can be classified according to whether they (i) rank individuals' sample paths $\mathcal{H}(\infty)$, (ii) rank outcomes or event types, or (iii) weight event types or time units spent with a specific disease condition. An example of (i) is the win ratio (WR), proposed by Pocock et al. (2012), which has received a great deal of attention (Bebu and Lachin, 2015; Oakes, 2016; Mao, 2018; Finkelstein and Schoenfeld, 2019; Mao and Wang, 2021; Dong et al., 2020b,c,a; Parner and Overgaard, 2023, 2024). As an example of (ii), Mao (2023) developed the so-called restricted mean time in favor of treatment (RMT-IF), which extends the restricted

mean survival time (RMST) to more general multistate processes and orders event types in terms of their favorability. Examples of (iii) include quality-adjusted life years (QALY) and the Q-TWiST measure introduced in Section 1.3.4 in the context of the IBCSG trial.

The objectives of this chapter are two-fold. First, we provide a comprehensive review of the ranking methods (i)-(ii), evaluate their limitations and argue for the usefulness of methods (iii) based on utility functions. Second, we carefully describe and utilize the infinitesimal jackknife procedure to develop a robust variance estimator for the expected cumulative utility score (1.2.4).

This chapter is organized as follows. In Section 5.2, we introduce notation for progressive multistate models and define utility functions. Section 5.3 provides a critical discussion of various ranking and weighting methods for comparing experimental treatments to standard care/placebo with respect to their effects on time spent in different health states and event occurrence, including the win ratio (Section 5.3.1), the RMT-IF (Section 5.3.2) and utility-based analyses (Section 5.3.3). We highlight in particular the advantages and disadvantages of each approach and carefully outline how they can be viewed in a unified way. A multistate model-based formulation for the win ratio is proposed in Section 5.3.1.1. Section 5.4 focuses on utility-based estimands and explores topics related to variance estimation and testing in the utility-based framework (1.2.3.2). In Section 5.4.1.2, we derive an infinitesimal jackknife variance estimator for the expected cumulative utility score in a general multistate process. Section 5.4.2 returns to the setting of the commonly used illness-death model, and examines the validity of the proposed variance estimator and associated tests in extensive simulation studies. An application to a platelet transfusion trial in chemotherapy-induced thrombocytopenia is provided in Section 5.5. Concluding remarks are given in Section 5.6.

5.2 Notation and utility functions

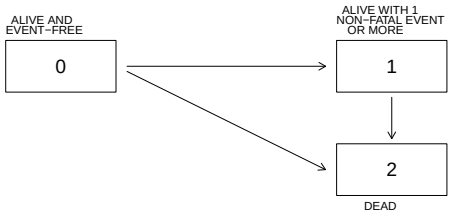
Consider a continuous-time multistate process $\{Z(t), 0 \leq t\}$ with a (partially) ordered state space $\mathcal{S} = \{0, 1, \ldots, K, K+1\}$, with an initial state 0 (e.g. alive and event-free) occupied by individuals at the time of randomization ($t = 0$) and state $K+1$ entered upon death. The transient states $k = 1, \ldots, K$ represent non-fatal disease conditions of increasing severity. For example, state k may represent a condition of some impairment due to a mild stroke, and $k+1$ could

correspond to a state of greater impairment due to a severe stroke. State-space diagrams for five different progressive multistate models are displayed in Figure 5.1. The diagram in Figure 5.1(a) illustrates the well-known illness-death model (Beyersmann et al., 2012), which is widely used in oncology trials with state 1 entered upon relapse or progression. Figure 5.1(b) extends the illness-death process to a progressive four-state model, with state 2 entered when further complications arise following cancer progression. For the recurrent and terminal event process in Figure 5.1(c) (Cook and Lawless, 2007), the states $k = 0, \ldots, K$ correspond to the cumulative number of non-fatal events with state $K+1$ entered upon death. In disease process settings (a)-(c) individuals can only transition from state k to either $k+1$ (i.e. the next highest non-fatal state) or $K+1$ (i.e. death). The multistate model in Figure 5.1(d) is motivated by trials of coronary heart diseases, where states 1 and 2 may be entered upon mild and severe strokes respectively (O'Connor et al., 2009). Figure 5.1(e) conveys the structure of the model for the IBCSG breast cancer trial described in Section 1.3.4, with states 0: toxicity, 1: toxicity- and symptom-free, 2: relapse, and 3: death (Ludwig Breast Cancer Study Group, 1988); here state $k+1$ can only be entered from state k, $k = 0, \ldots, 2$. We restrict attention to progressive processes here where each state can be entered at most once, but the methods are easily adopted to more general processes. We let $T_{kl} = \inf\{t > 0 : Z(t) = l, Z(t^-) = k\}$ and $T_k = \inf\{t > 0 : Z(t) = k\}$ represent the $k \longrightarrow l$ transition time and the entry time to state k, respectively.

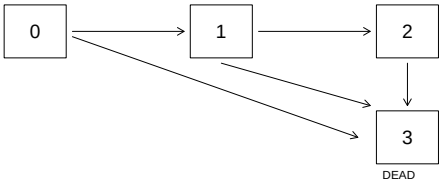
We let X denote the binary treatment indicator as before, and use $i = 1, \ldots, n_1$ to index individuals in the experimental treatment group $(X = 1)$ and $j = 1, \ldots, n_0$ for individuals in the control group $(X = 0)$; hence, $Z_i(t) \in \mathcal{S}$ represents the state occupied by an individual with $X = 1$ at time t, and $Z_j(t) \in \mathcal{S}$ is used for individuals with $X = 0$. Suppose all $n = n_0 + n_1$ individuals are observed over the time period $[0, \tau]$, where $\tau > 0$ denotes the administrative censoring time. We further assign a numerical utility value $u_k \in [0, 1]$ to each state $k \in \mathcal{S}$ such that the utility values $u_0, u_1, \ldots, u_{K+1}$ have the same (partial) ordering as the states, i.e.

$$1 = u_0 \geq u_1 \geq \ldots \geq u_{K+1} = 0 \, .$$

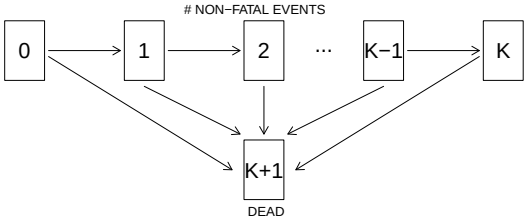
Naturally, if u_k measures quality of life associated with state k, the highest utility is assigned to the initial state 0 and the lowest utility to the absorbing state $K+1$ of death. Quality of life utilities used by Glasziou et al. (1990) for the IBCSG trial of Section 1.3.4 are $0.1, 0.5, 0.1$ and 0



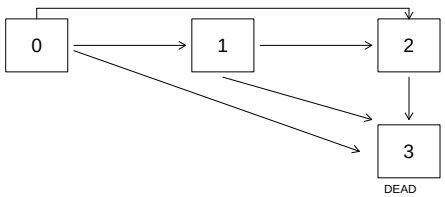
(a) A progressive illness-death process with one intermediate event.



(b) A progressive process with two intermediate events.



(c) A recurrent event process terminated by death.



(d) A progressive process with two intermediate events.



(e) A progressive process with two intermediate events.

Figure 5.1: Illustrative multistate diagrams for progressive disease processes.

for the states $0, 1, 2$ and 3, and are thus not ordered the same way as the states. If

$$u(t) = \sum_{k \in \mathcal{S}} u_k \mathbb{1}(Z(t) = k)$$

is an individual's utility (or quality of life score) at time t (or over $[t, t + dt)$), the cumulative utility score over the trial duration $[0, \tau]$ is defined as

$$U(\tau) = \int_0^\tau u(t)dt = \sum_{k \in \mathcal{S}} u_k \int_0^\tau \mathbb{1}(Z(t) = k)dt = \sum_{k \in \mathcal{S}} u_k S_k(\tau) \ , \tag{5.2.1}$$

where $S_k(\tau)$ is the sojourn time in state k over $[0, \tau]$. The expectation of (5.2.1), termed the expected cumulative utility score, was introduced in (1.2.4) as a potential target marginal process feature for estimand specification in type (I) clinical trial settings (Cook et al., 2003; Cook and Lawless, 2018, Section 4.2.3):

$$D(\tau) = \mathbb{E}(U(\tau)) = \sum_{k \in \mathcal{S}} u_k \int_0^\tau P(Z(t) = k)dt = \sum_{k \in \mathcal{S}} u_k RMT_k(\tau) \ , \tag{5.2.2}$$

where $RMT_k(\tau)$ is the restricted mean time spent in state k over $[0, \tau]$; see also (1.2.5). Note that (5.2.2) can be viewed as the "weighted" sum of restricted mean times $RMT_k(\tau)$ in states $k = 0, 1, \ldots, K, K + 1$. More generally time- and state-dependent utilities could be considered, in which case $u(t) = \sum_{k \in \mathcal{S}} u_k(t) \mathbb{1}(Z(t) = k)$, but for simplicity we assume a time-constant utility value u_k for state k.

5.3 A critical discussion of estimands based on utility functions and ranked outcomes

We next discuss some current approaches based on ranking sample paths or states of health in more detail, and describe a framework with which they can be unified and generalized using utilities.

5.3.1 Win ratio

The win ratio (WR) was introduced by Pocock et al. (2012) as an approach to jointly consider first hospitalizations and cardiovascular (CV) death in chronic heart failure trials. More generally it has become a popular method for the analysis of disease processes involving multiple events of varying clinical importance (Ferreira et al., 2020; Beal et al., 2022; Kondo et al., 2023; Pocock et al., 2023). Tafamidis, a drug for the treatment of cardiomyopathy, was approved by the US Food and Drug Administration (FDA) based on a primary analysis using the WR based on non-fatal CV-related hospitalizations and death (Maurer et al., 2018; Pocock and Collier, 2019).

The standard approach to formulating the WR is based on pairwise comparisons (Finkelstein and Schoenfeld, 1999; Buyse, 2010): all possible pairs (i, j) of individuals with $X_i = 1, X_j = 0$ and their process histories $(\mathcal{H}_i(\infty), \mathcal{H}_j(\infty))$ are compared according to a pre-specified hierarchical rule and "winners" are determined. The WR is defined as the proportion of pairs where individual i (from the experimental arm) "wins" over individual j (from the control arm) over $[0, \tau]$ divided by its complement (i.e. the proportion of pairs where j "wins" over i). In the absence of censoring it is estimated by

$$\widehat{\beta}_c^{\dagger}(\tau) = \frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_0} \mathbb{1}(i \text{ wins over } j \text{ over } [0, \tau])}{\sum_{i=1}^{n_1} \sum_{j=1}^{n_0} \mathbb{1}(j \text{ wins over } i \text{ over } [0, \tau])} .$$

It can be shown that if all individuals are followed over $[0, \tau]$ (i.e. there are no premature dropouts), the simple "counting" estimator $\widehat{\beta}_c^{\dagger}(\tau)$ is consistent for the estimand

$$\beta^{\dagger}(\tau) = \frac{p_{ij}(\tau)}{p_{ji}(\tau)} ,$$

where $p_{ij}(\tau) = P(i \text{ wins over } j \text{ over } [0, \tau])$ and $p_{ji}(\tau) = P(j \text{ wins over } i \text{ over } [0, \tau])$ (Bebu and Lachin, 2015; Luo et al., 2017). In the presence of censoring and unequal follow-up times, however, the ordinary counting approach cannot be implemented and restriction to pairs for which it can be evaluated yields biased estimates of the win probabilities $p_{ij}(\tau)$ and $p_{ji}(\tau)$; this is a major limitation of the traditionally used WR. Matched WR versions that compare only pairs (i, j) with similar lengths of follow-up have been proposed to overcome this limitation. Other adjustments have been proposed by Dong et al. (2020c), Dong et al. (2021), Parner and Overgaard (2023) and Wang et al. (2024).

We next discuss the WR estimand in terms of joint probabilities regarding the event times.

The WR estimand $\beta^{\dagger}(\tau)$ relies on the ranking rule by which "winners" and "losers" are determined. For the illness-death process in Figure 5.1(a) individuals are typically first compared based on death (i.e. the *worst* outcome): if $T_{2j} \leq t, T_{2j} < T_{2i}$ for $X_i = 1$ and $X_j = 0$, then individual i is declared as the winner. If no decision can be made on death, individuals i and j are compared with regard to the non-fatal event, which is the second worst outcome in the present context: i wins over j on the non-fatal event by time t if $T_{2j} > t, T_{2i} > t, T_{1j} \leq t$ and $T_{1j} < T_{1i}$. A so-called tie is declared if a decision cannot made regarding the available data on death and the non-fatal event. The probability that i wins over j at time t can then be expressed as

$$p_{ij}^{ID}(t) = \underbrace{P(T_{2j} \leq t, T_{2j} \leq T_{2i})}_{win\ on\ survival} + \underbrace{P(T_{2j} > t, T_{2i} > t, T_{1j} \leq t, T_{1j} < T_{1i})}_{win\ on\ non-fatal\ event} . \tag{5.3.1}$$

Oakes (2016) noted that $p_{ij}^{ID}(t)$ is likewise given by

$$p_{ij}^{ID}(t) = \int_0^t f_{2j}(u)P(T_{2i} > u)du + P(T_{2i} > t)P(T_{2j} > t) \int_0^t g_j(u|t)P(T_{1i} > u|T_{2i} > t)du , \tag{5.3.2}$$

where $f_{2j}(u) = \partial P(T_{2j} \leq u)/\partial u$ is the marginal density function of the overall survival time and $g_j(u|t) = \partial P(T_{1j} \leq u|T_{2j} \geq t)/\partial u$ is the improper conditional density function of the entry time to state 1 at time u given survival up to time t. Oakes (2016) referred to (5.3.2) as the "integral" formulation of $p_{ij}^{ID}(t)$. Within the multistate model framework, i wins over j by time t if i and j occupy the same state at t but j has entered the state earlier than i, i.e.

$$\begin{aligned} &Z_i(t) = 2, Z_j(t) = 2,\ T_{2j} < T_{2i} \\ &Z_i(t) = 1, Z_j(t) = 1,\ T_{1j} < T_{1i} , \end{aligned}$$

or if i is in a more favorable (and less severe) state than j at t, i.e.

$$\begin{aligned} &Z_i(t) = 0, Z_j(t) = 2 \\ &Z_i(t) = 1, Z_j(t) = 2 \\ &Z_i(t) = 0, Z_j(t) = 1 . \end{aligned}$$

There is no winner if both processes are in the initial state 0 at t, i.e. $Z_i(t) = 0$ and $Z_j(t) = 0$. As

a result, the win probability can be expressed alternatively as

$$p_{ij}^{ID}(t) = \sum_{k=1}^{2}\left[P(Z_i(t) = Z_j(t) = k, T_{kj} < T_{ki}) + \sum_{r<k} P(Z_i(t) = r, Z_j(t) = k)\right], \quad (5.3.3)$$

where

$$P(Z_i(t) = Z_j(t) = k, T_{kj} < T_{ki}) = \int_0^t P(Z_i(t) = k, Z_j(t) = k, T_{kj} = u, T_{ki} > u)du, \quad k = 1, 2,$$
$$P(Z_i(t) = r, Z_j(t) = k) = P(Z_i(t) = r)P(Z_j(t) = k), \quad r < k,$$

provided that $\{Z_i(t), 0 < t\} \perp\!\!\!\perp \{Z_j(t), 0 < t\}$. Explicit expressions for $P(Z_i(t) = Z_j(t) = k, T_{kj} < T_{ki}), k = 1, 2$, are given in Appendix D.1.2. Simple algebra in Appendix D.1.1 proves the equality of the win probabilities (5.3.1), (5.3.2) and (5.3.3). In light of (5.3.3), a straightforward approach to estimation is to use nonparametric Nelson-Aalen (NA) and Aalen-Johansen (AJ) estimates of the cumulative transition intensities and transition probabilities for each treatment group to obtain estimates of (5.3.3) and $p_{ji}^{ID}(t)$. Although justified under a Markov assumption, Aalen et al. (2001) point out that the AJ estimator provides a consistent estimate of $P(Z(t) = k)$ for non-Markov processes. Variance estimation can be carried out based on a nonparametric bootstrap.

5.3.1.1 A general multistate model formulation of the win ratio

We extend Oakes (2016) integral-based WR formulation to more general disease processes and reexpress the WR estimand $\beta^{\dagger}(\tau)$ through win and loss probabilities implied by the multistate model representation. A particular appeal of this reformulation is that the proposed approach does not require pairing of individuals and accounts for incomplete data due to right-censoring by nonparametric NA and AJ methods. A small simulation study in Appendix D.1.3 shows that the multistate modeling approach provides estimates of $p_{ij}(\tau)$ and $p_{ji}(\tau)$ with negligible bias even when the censoring rate is quite substantial, whereas the ordinary counting approach does not. Moreover, the multistate model formulation facilitates comparisons of $\beta^{\dagger}(\tau)$ to other approaches such as the RMT-IF (Mao, 2023) and utility functions (Cook et al., 2003), as we will see in Section 5.3.3.

We define the WR in a general progressive multistate process by

$$\beta^{\dagger}(\tau) = \frac{\sum_{k=1}^{K+1}\left[P(Z_i(\tau) = Z_j(\tau) = k, T_{kj} < T_{ki}) + \sum_{r<k} P(Z_i(\tau) = r, Z_j(\tau) = k)\right]}{\sum_{k=1}^{K+1}\left[P(Z_j(\tau) = Z_i(\tau) = k, T_{ki} < T_{kj}) + \sum_{r<k} P(Z_j(\tau) = r, Z_i(\tau) = k)\right]} . \quad (5.3.4)$$

Although the estimand (5.3.4) can be estimated consistently from right-censored data (Aalen et al., 2001), expressions involved such as $P(Z_i(\tau) = Z_j(\tau) = k, T_{kj} < T_{ki})$ can become quite complicated for large $|S|$ and estimability issues may arise. Progressive multistate models with relatively few states are less problematic. For the four-state process in Figure 5.1(b), for example, there are 9 conditions under which the treatment group wins over the control group.

Hierarchical ranking of sample paths $\mathcal{H}(\infty)$ becomes more challenging for more complex processes. There is, for example, no consensus on how to rank sample paths in the recurrent and terminal event setting of Figure 5.1(c). Mao et al. (2022) used the "counting" formulation to construct the first-event-assisted and the last-event-assisted WR, without addressing the censoring issue. Caution is thus warranted when using the `R` package `WR`; see also simulation results in Appendix D.1.3. In their approach the $n_0 n_1$ pairs of individuals are initially compared with respect to death, followed by the cumulative number of non-fatal events. If both i and j are alive at time t with the same number of non-fatal events experienced over $[0, t]$ (i.e. $N_i(t) = N_j(t) = k$), the first-event assisted WR breaks that tie by taking the time to the first non-fatal event as the basis for a third comparison. Hence, Mao et al. (2022)'s ranking rule would declare that i's path is preferable over j's path if $T_{(K+1)j} \leq t, T_{(K+1)j} \leq T_{(K+1)i}$, or $T_{(K+1)j} > t, T_{(K+1)i} > t, N_i(t) < N_j(t)$, or $T_{(K+1)j} > t, T_{(K+1)i} > t, N_i(t) = N_j(t) = k, T_{1j} < T_{1i}$; for the last-event assisted WR, the condition $T_{1j} < T_{1i}$ is replaced by comparisons of the time of the last non-fatal event: $T_{kj} < T_{ki}$.

Equations (1) and (2) in Parner and Overgaard (2023) give win probabilities of the form

$$P(Z_i(t) = Z_j(t) = K+1, T_{(K+1)j} < T_{(K+1)i}) + \sum_{k=1}^{K+1}\sum_{r<k} P(Z_i(t) = r, Z_j(t) = k) ,$$

with $Z_i(t) = Z_j(t) = k$ (or $N_i(t) = N_j(t) = k$) treated as a tie. Thus, Parner and Overgaard (2023)'s rule would declare that there is no definite winner although j, for example, might have reached state k much earlier than i. The WR (5.3.4) takes this into account and breaks such ties

by comparisons of the time to the kth non-fatal event, which is similar to the last-event assisted WR. An issue with the WR is that trade-offs between different outcomes depend on what they represent and on judgments or decisions about ranking paths.

Another limitation with the WR is its dependence on τ (Finkelstein and Schoenfeld, 2019; Dong et al., 2020c). This means that designation of winners in pairs (i, j) may change with follow-up. Individual i may be declared the winner at $\tau_0 < \tau_1$, but as time passes this designation could change, making individual j the winner at a later time point τ_1. The time point τ determines the priority of outcomes (or the components of $p_{ij}(\tau)$) based on which $\beta^{\dagger}(\tau)$ is calculated. This is graphically exemplified in Figures 5.2-5.4, where the illness-death process with time-homogeneous intensities λ_{kl} was used for illustration. For the moderate mortality scenario in Figure 5.2 (i.e. when $\lambda_{01} = 1, \lambda_{02} = 0.75$ and $\lambda_{12} = 0.5$), the win probability $p_{ij}^{ID}(\tau)$ is determined by $P(Z_i(\tau) = 0, Z_j(\tau) = 1)$ and $P(Z_i(\tau) = 0, Z_j(\tau) = 2)$ for earlier choices of τ, and by $P(Z_i(\tau) = 1, Z_j(\tau) = 2)$ and $P(Z_i(\tau) = 2, Z_j(\tau) = 2, T_{2j} < T_{2i})$ for later choices of τ. For the extreme scenarios of low or high mortality in Figures 5.3 and 5.4, $p_{ij}^{ID}(\tau)$ is mostly determined by comparisons on the non-fatal event or death for all choices of τ, respectively. An advantage of the multistate model WR formulation lies in its facilitation of investigations of this kind.

The WR in (5.3.4) uses the state $k = 1, \ldots, K + 1$ occupied at time τ and the associated entry time T_k to rank individuals. This ranking rule follows the original papers on WR methods (Oakes, 2016), but is quite restrictive and fails to capture an individual's entire sample path. If both individuals i and j are in state k at τ, their paths before entering state k (i.e. $\mathcal{H}_i(t_{ik})$ and $\mathcal{H}_j(t_{jk})$) do not matter for the comparison of (i, j). More generally, the pre-defined ranking rule can depend on arbitrary features of the process history $\mathcal{H}(\tau)$. For instance, the time spent in a more favorable state over $[0, \tau]$ could be used to determine a ranking (Troendle et al., 2024). Alternatively, individuals can be ranked according to their cumulative utility scores (5.2.1):

$$\beta^{\dagger}(\tau) = \frac{P(U_i(\tau) > U_j(\tau))}{P(U_j(\tau) > U_i(\tau))} \,. \tag{5.3.5}$$

The underlying ranking rule in (5.3.5) makes fuller use of individuals' sample paths and in particular the time spent in more desirable health states defined by their relative utilities.

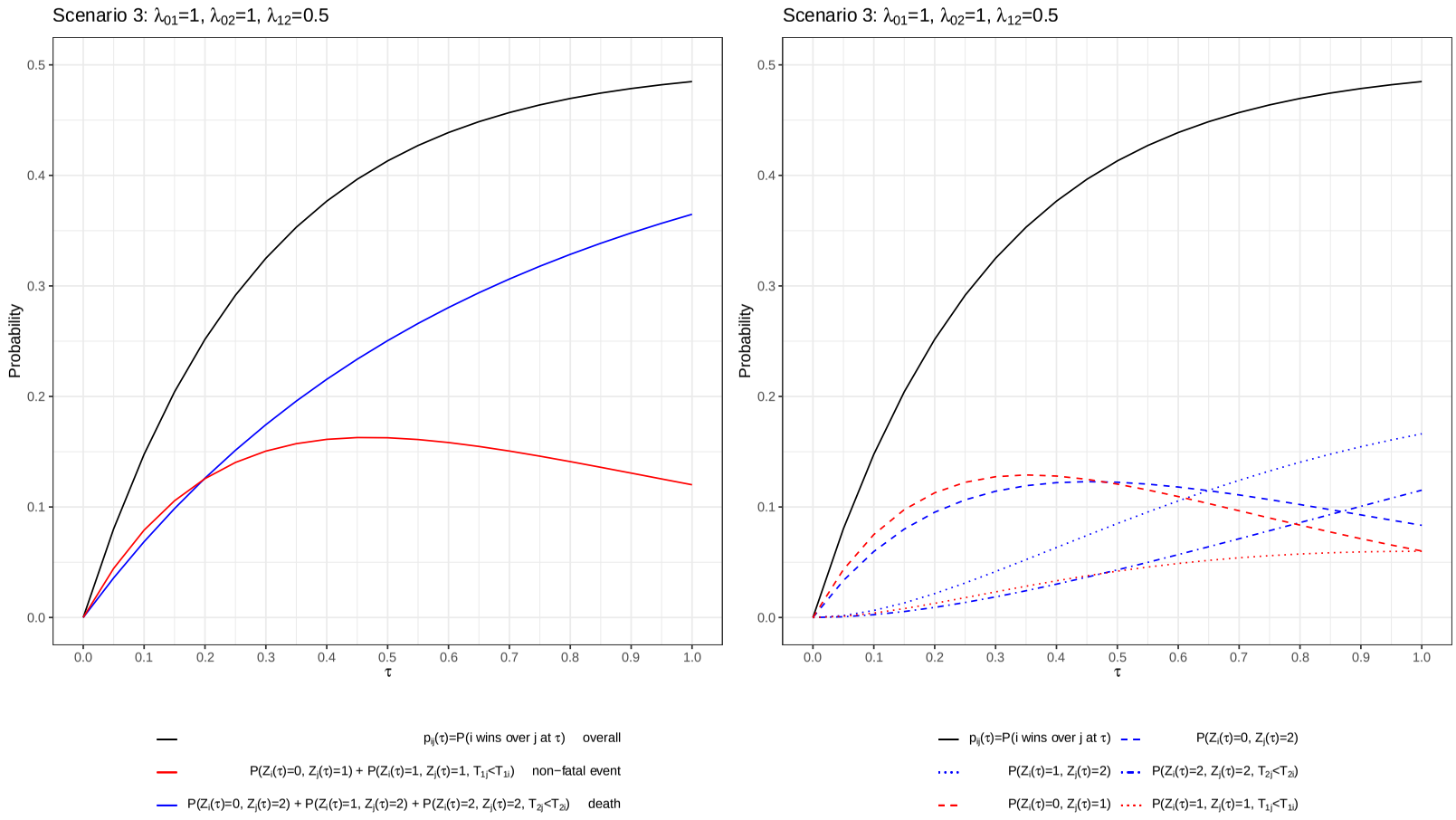


Figure 5.2: Plots of the win probability $p_{ij}^{ID}(\tau)$ and its 2 (left panel) or 5 components (right panel) versus τ when the true illness-death process has time-homogeneous intensities $\lambda_{01} = 1, \lambda_{02} = 0.75$ and $\lambda_{01} = 0.5$ (moderate mortality scenario).

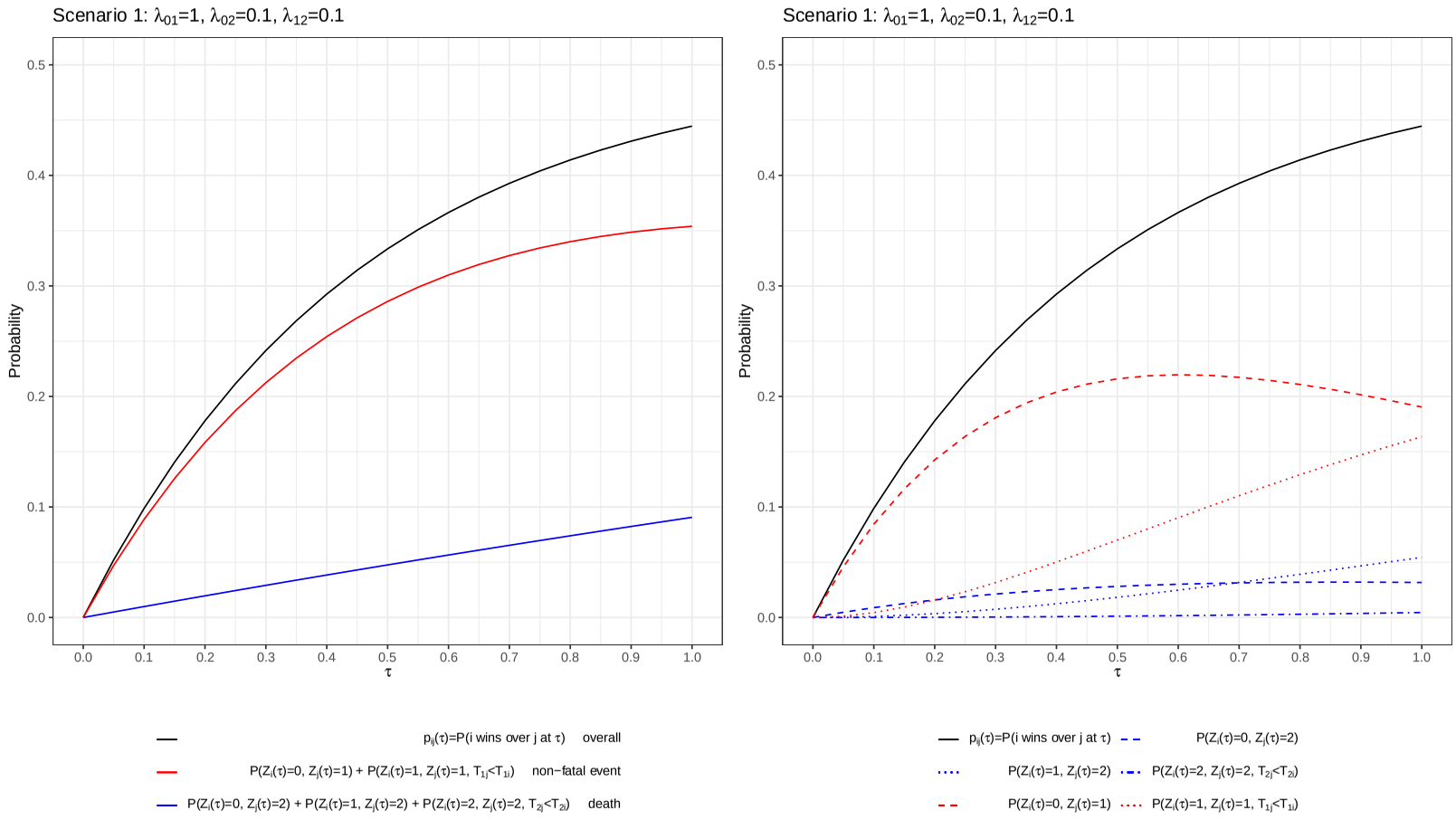


Figure 5.3: Plots of the win probability $p_{ij}^{ID}(\tau)$ and its 2 (left panel) or 5 components (right panel) versus τ when the true illness-death process has time-homogeneous intensities $\lambda_{01} = 1, \lambda_{02} = 0.1$ and $\lambda_{01} = 0.1$ (low mortality scenario).

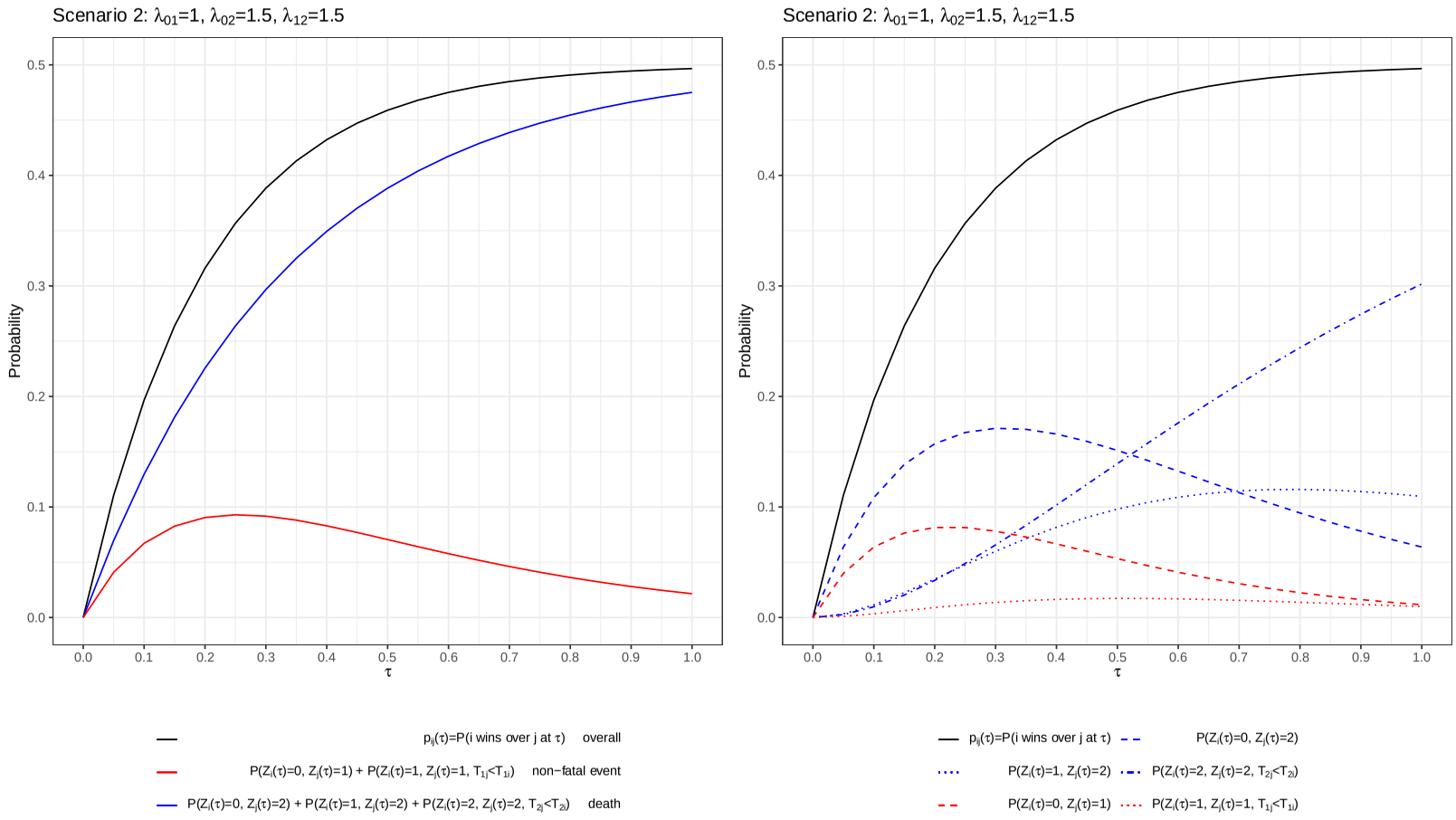


Figure 5.4: Plots of the win probability $p_{ij}^{ID}(\tau)$ and its 2 (left panel) or 5 components (right panel) versus τ when the true illness-death process has time-homogeneous intensities $\lambda_{01} = 1, \lambda_{02} = 1.5$ and $\lambda_{01} = 1.5$ (high mortality scenario).

5.3.2 Restricted mean time in favor of treatment

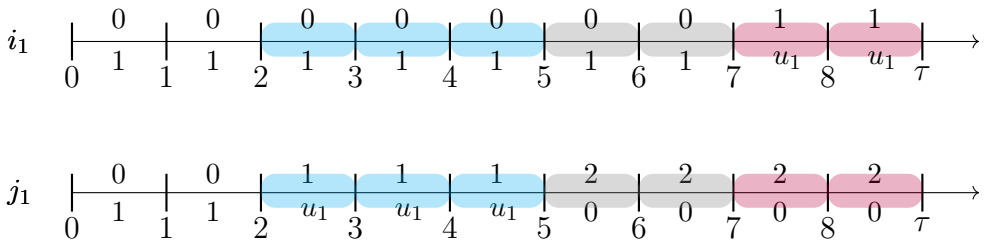
Mao (2023) defines the so-called RMT-IF, which quantifies the restricted mean time (RMT) over the time interval $[0, \tau]$ that an individual in the experimental treatment group is expected to spend in a "better" state than an individual in the control group. In a similar spirit to the WR it considers all possible pairs (i, j) of individuals with $X_i = 1$ and $X_j = 0$, and determines the total time over $[0, \tau]$ that i spent in a lower (or more favorable and less severe) state than j, defined as

$$V_{ij}(\tau) = \sum_{k=1}^{K+1} \int_0^\tau \mathbb{1}(Z_i(t) < k, Z_j(t) = k)dt \ , \tag{5.3.6}$$

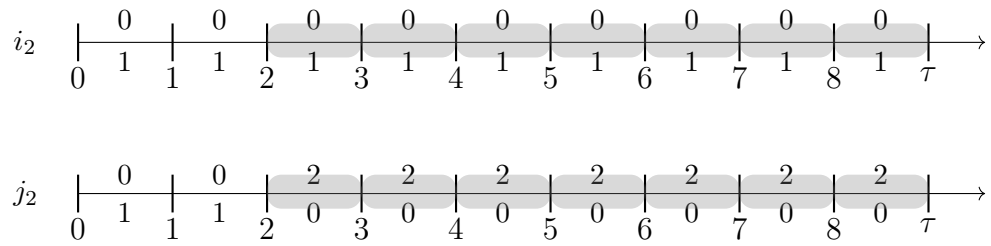
and vice versa. The following example provides an illustration of the marginal process feature (5.3.6) in the context of a three-state illness-death process, where state 0 is considered "better" than states $1, 2$ and state 1 is "better" than state 2.

Example 1 *Consider a hypothetical trial in which individuals experience illness-death processes (see Figure 5.1(a)), with no premature dropouts before the end of follow-up time $\tau = 9$. Recall that, in the context of Figure 5.1(a), T_{ki} denotes the entry time to state $k = 1, 2$ for individual i; we set $T_{ki} = \infty$ if i has not entered state k before time τ.*
Consider Figure 5.5(a), where the upper panel (or row) represents individual i_1's sample path $\mathcal{H}_{i_1}(\tau)$ and the bottom panel corresponds to j_1's path $\mathcal{H}_{j_1}(\tau)$; the numbers above the x-axis denote the state k an individual is in over $[s, s + 1), s = 0, \ldots, \tau - 1$, $k = 0, 1, 2$. For now, we disregard the numbers below the x-axis as well as the colored areas. If individual i_1 with $X_{i_1} = 1$ has $T_{1i_1} = 7, T_{2i_1} = \infty$ and individual j_1 with $X_{j_1} = 0$ has $T_{1j_1} = 2, T_{2j_1} = 5$, then the total time (5.3.6) with i_1 in a better state is 7 and the total time with j_1 in a better state is 0, respectively. That is, $V_{i_1 j_1}(\tau) = 7$ and $V_{j_1 i_1}(\tau) = 0$.
Consider now Figure 5.5(b). If individual i_2 with $X_{i_2} = 1$ has $T_{1i_2} = T_{2i_2} = \infty$ and individual j_2 with $X_{j_2} = 0$ has $T_{1j_2} = \infty, T_{2j_2} = 2$, then the total time (5.3.6) with i_2 in a better state is 7 and the total time with j_2 in a better state is 0. That is, $V_{i_2 j_2}(\tau) = 7$ and $V_{j_2 i_2}(\tau) = 0$.

Taking the expectation of (5.3.6) over all pairs (i, j) gives the RMT over $[0, \tau]$ where an individual from the experimental treatment group is expected to occupy a lower (or more favorable and less



(a) Individual i_1 vs. j_1: $T_{1i_1} = 7, T_{2i_1} = \infty, T_{1j_1} = 2, T_{2j_1} = T_{12j_1} = 5$.



(b) Individual i_2 vs. j_2: $T_{1i_2} = T_{2i_2} = \infty, T_{1j_2} = \infty, T_{2j_2} = T_{02j_2} = 2$.

Figure 5.5: Illustration of the RMT-IF (5.3.7) and the utility-based estimand (5.3.8) in a three-state illness-death process using sample paths of 4 hypothetical individuals; $X_{i_1} = X_{i_2} = 1, X_{j_1} = X_{j_2} = 0$.

severe) state than an individual from the control treatment group (Mao and Wang, 2024):

$$\sum_{k=1}^{K+1} \int_0^\tau P(Z_i(t) < k, Z_j(t) = k)dt = \sum_{k=1}^{K+1} \sum_{r<k} RMT_{rk}(\tau) \ ,$$

where

$$RMT_{rk}(\tau) := \int_0^\tau P(Z_i(t) = r, Z_j(t) = k)dt \ .$$

If

$$RMT_{kr}(\tau) := \int_0^\tau P(Z_j(t) = r, Z_i(t) = k)dt$$

respectively, the RMT-IF by Mao (2023) considers the difference

$$\widetilde{\beta}(\tau) = \sum_{k=1}^{K+1} \sum_{r<k} \big(RMT_{rk}(\tau) - RMT_{kr}(\tau) \big) \tag{5.3.7}$$

as the target marginal estimand. Assuming $\{Z_i(t), t > 0\} \perp\!\!\!\perp \{Z_j(t), t > 0\}$, (5.3.7) can be estimated by nonparametric Aalen-Johansen estimation of the state occupancy probabilities $P_{0k}(t|X)$ in each treatment group.

Compared to the win ratio (WR), specification of the estimand (5.3.7) does not require ranking of individuals' sample paths $\mathcal{H}(\infty)$ but ranking or ordering of states in a multistate model. It requires, in particular, a clear distinction of the states in terms of their favorability; this is rather simple in a two-state survival process (i.e. alive vs. dead) or in an illness-death process with only one non-fatal event (i.e. alive and event-free vs. non-fatal event vs. dead), but becomes more complicated for other disease processes. We note that in the case of a two-state survival process the RMT-IF reduces to the well-known restricted mean survival time (Royston and Parmar, 2013). Careful thought should therefore go into how best to rank the states defined by event occurrence. In CV trials, for example, it is debatable whether a stroke or a myocardial infarction (MI) is more favorable and it will in practice depend on the severity of the events and their impact on quality of life; this is also an issue for analyses based on utility functions. That is, the definitions of the states need to reflect health following an event like a stroke or MI. This is a topic of future research mentioned in Section 6.4.

The estimand (5.3.7) contains some components that are also involved in the WR (5.3.4), such as $P(Z_i(t) = r, Z_j(t) = k), r < k$, but (5.3.7) accumulates them over $[0, \tau]$. Although the RMT-IF approach makes fuller use of individuals' trajectories than any WR method, it does not always provide a reasonable assessment of two sample paths $\mathcal{H}_i(\infty)$ and $\mathcal{H}_j(\infty)$. Specifically, severity of different event types (or states) is only partially addressed in (5.3.7). In Example 1, for instance, the time unit over which individual i_1 is in state 0 and j_1 in state 1 is equally treated as the time unit over which i_1 is in state 1 and j_1 in state 2. This leads to the fact that the pair (i_1, j_1) is equally treated to the pair (i_2, j_2) despite their differing sample paths. More generally, there is no distinction between 0 vs. 1, 0 vs. 2 or 1 vs. 2 comparisons in an illness-death process. There is no further refinement of "better" or "worse", which is a drawback of the RMT-IF.

This has lead to the consideration of more general utility function and utility-based estimands. In the next section we propose general utility-based methods as a preferable alternative to WRs and related ranking methods. First, we point out that utility functions provide a unification of the estimands in (5.3.4) and (5.3.7).

5.3.3 Unification of estimands through utilities

Recall from Section 5.2 that $u_i(t)$ defines individual i's utility at time t. The fundamental idea of Mao (2023)'s restricted mean time in favor of treatment (RMT-IF) is to quantify the treatment effect on the (cumulative) expected time units $[t, t+dt)$ over $[0, \tau]$ at which the experimental treatment group is in a "better" (or lower ranked) state than the control group. Since a lower ranked state has a higher utility, the condition $u_i(t) > u_j(t)$ is equivalent to $Z_i(t) < Z_j(t)$. Consequently, we obtain

$$\mathbb{1}(u_i(t) > u_j(t)) = \sum_{k=1}^{K+1} \sum_{r<k} \mathbb{1}(u_i(t) = u_r, u_j(t) = u_k) = \sum_{k=1}^{K+1} \sum_{r<k} \mathbb{1}(Z_i(t) = r, Z_j(t) = k) \ ,$$

and the estimand $\widetilde{\beta}(\tau)$ in (5.3.7) can be rewritten in terms of utility functions as

$$\begin{aligned}
\widetilde{\beta}(\tau) &= \int_0^\tau \big(P(u_i(t) > u_j(t)) - P(u_j(t) > u_i(t))\big) dt \\
&= \sum_{k=1}^{K+1} \sum_{r<k} \int_0^\tau \big[P(Z_i(t) = r, Z_j(t) = k) - P(Z_j(t) = r, Z_i(t) = k)\big] dt \\
&= \sum_{k=1}^{K+1} \sum_{r<k} \big(RMT_{rk}(\tau) - RMT_{kr}(\tau)\big) \ .
\end{aligned}$$

This estimand considers only whether the utility in the experimental treatment group at time t is higher than the utility in the control group, but fails to distinguish between specific utility scores. Merely ranking event types (or states) does not reflect whether the sample paths are similar or highly dissimilar; see also Figure 5.5. Utility-based methods have the advantage of providing a metric in addition to the ranking, as we will see next.

5.3.3.1 Utility-based estimand

The fundamental idea of utility-based analyses is to "penalize" each time unit $[t, t+dt)$ over $[0, \tau]$ at which an individual is not in the initial state 0 (Cook et al., 2003). The value $u_k \in [0, 1)$ by which the time unit is multiplied depends on the state k occupied by an individual at t, $k \neq 0$. If an individual i stays in the most favorable state 0 over $[0, \tau]$, the cumulative utility score is $U_i(\tau) = u_0\tau = \tau$ for $u_0 = 1$. Otherwise, $U_i(\tau) < \tau$.

Example 2 *Reconsider Example 1 of Figure 5.5, and recall that in the upper panel (a)* $X_{i_1} = 1, X_{j_1} = 0$ *and* $T_{1i_1} = 7, T_{2i_1} = \infty, T_{1j_1} = 2, T_{2j_1} = 5$ *and in the lower panel (b)* $X_{i_2} = 1, X_{j_2} = 0$ *and* $T_{1i_2} = T_{2i_2} = \infty, T_{1j_2} = \infty, T_{2j_2} = 2$. *The numbers below the x-axis in each panel represent the utility scores* u_k *for* $k = 0, 1, 2$. *By assuming* $u_0 = 1$ *and* $u_2 = 0$, *the cumulative utility scores (5.2.1) over* $[0, \tau]$ *for individuals* i_1 *and* j_1 *are* $U_{i_1}(\tau) = 7u_0 + 2u_1 = 7 + 2u_1 < 9$ *and* $U_{j_1}(\tau) = 2u_0 + 3u_1 = 2 + 3u_1 < 9$*; similarly,* $U_{i_2}(\tau) = 9u_0 = 9$ *and* $U_{j_2}(\tau) = 2u_0 = 2 < 9$. *This demonstrates that the utility-based approach allows for a more nuanced comparison of sample paths compared to the RMT-IF, since* $U_{i_1}(\tau) - U_{j_1}(\tau) = 5u_0 - u_1 \neq 7u_0 = U_{i_2}(\tau) - U_{j_2}(\tau)$ *and* $V_{i_1 j_1}(\tau) - V_{j_1 i_1}(\tau) = 7 = V_{i_2 j_2}(\tau) - V_{j_2 i_2}(\tau)$.

The utility-based estimand can be expressed in one of the following ways:

$$
\begin{aligned}
\beta(\tau) &= D_i(\tau) - D_j(\tau) = \mathbb{E}(U_i(\tau)) - \mathbb{E}(U_j(\tau)) \\
&= \mathbb{E}\left(\int_0^\tau u_i(t)dt\right) - \mathbb{E}\left(\int_0^\tau u_j(t)dt\right) \\
&= \sum_{k=1}^{K+1} \sum_{r<k} \int_0^\tau u_r \big[P(Z_i(t) = r, Z_j(t) = k) - P(Z_j(t) = r, Z_i(t) = k)\big] dt \\
&\qquad - \sum_{k=1}^{K} \sum_{r<k} \int_0^\tau u_k \big[P(Z_i(t) = r, Z_j(t) = k) - P(Z_j(t) = r, Z_i(t) = k)\big] dt \\
&= \sum_{k=1}^{K+1} \sum_{r<k} (u_r - u_k)\big(RMT_{rk}(\tau) - RMT_{kr}(\tau)\big) \ .
\end{aligned}
\tag{5.3.8}
$$

The second last equality follows from the fact that

$$
U_i(\tau) - U_j(\tau) = \sum_{k \in \mathcal{S}} u_k \int_0^\tau \big(\mathbb{1}(Z_i(t) = k) - \mathbb{1}(Z_j(t) = k)\big) dt \ ,
$$

where $\mathbb{1}(Z_i(t) = k) - \mathbb{1}(Z_j(t) = k)$ equals 0 if $Z_i(t) = Z_j(t) = k$, 1 if $Z_i(t) = k, Z_j(t) \neq k$, and -1 if $Z_i(t) \neq k, Z_j(t) = k$. A thorough proof is provided in Appendix D.2.

5.4 More general use of utilities

Tests for treatment effects play a central role in the primary analysis of clinical trials. As in Section 3.1 where we consider cumulative incidence function regression, here we maintain the position that tests of no treatment effect should be defined in terms of the primary target estimand. Within the framework (1.2.3.2) of utility-based analyses it is thus important for tests of no treatment effect to be associated with the utility-based estimand (5.3.8):

$$\beta(\tau) = D(\tau|X = 1) - D(\tau|X = 0) \ .$$

Our objective here is to develop valid hypothesis tests for

$$H_0 : \beta(\tau) \leq 0 \ \text{ vs. } \ H_1 : \beta(\tau) > 0, \quad \tau > 0 \tag{5.4.1}$$

using test statistics of the form

$$\frac{\widehat{\beta}(\tau)}{\sqrt{\widehat{\text{var}}\big(\widehat{\beta}(\tau)\big)}} = \frac{\widehat{D}(\tau|X = 1) - \widehat{D}(\tau|X = 0)}{\sqrt{\widehat{\text{var}}\big(\widehat{D}(\tau|X = 1)\big) + \widehat{\text{var}}\big(\widehat{D}(\tau|X = 0)\big)}} \overset{H_0}{\sim} N(0, 1) \quad , \tag{5.4.2}$$

where

$$\widehat{D}(\tau|X) = \sum_{k \in \mathcal{S}} u_k \int_0^\tau \widehat{P}_{0k}(0, t|X)dt \ ,$$

is a nonparametric Aalen-Johansen-based estimator for $\widehat{D}(\tau|X)$ (Aalen et al., 2001; Cook et al., 2003) and $\widehat{\text{var}}\big(\widehat{D}(\tau|X)\big)$ is a suitable variance estimator for $\widehat{D}(\tau|X)$, $X = 0, 1$. We next introduce notation and rewrite $\widehat{D}(\tau|X)$ to facilitate derivation of the variance for $\widehat{D}(\tau|X)$.

Suppose individuals under study experience one of the multistate processes of Figure 5.1 over the time period $[0, C_a]$, where C_a is a common administrative censoring time. We let $Y_{ik}(t) = \mathbb{1}(Z_i(t^-) = k)$ indicate whether individual i is in state k at time t^- or not. We further let $dN_{ikl}(t) = \mathbb{1}(Z_i(t) = l, Z_i(t^-) = k) = 1$ if a $k \longrightarrow l$ transition occurs for individual i at time

t and, $dN_{ikl}(t) = 0$ otherwise; we denote the associated right-continuous counting process by $\{N_{ikl}(t), 0 < t\}$. Individual i may be prematurely lost to follow-up at a random time $C_{ir} < C_a$, and so we define the net censoring time as $C_i = \min(C_{ir}, C_a)$. Finally, if $Y_i(t) = \mathbb{1}(C_i \geq t)$, $\overline{Y}_{ik}(t) = Y_i(t)Y_{ik}(t)$ is the indicator that individual i is under observation and at risk of transitions out of state k at t.

We now let $0 < t_{(1)} < \ldots < t_{(S)} \leq \tau < C_a$ be the S distinct ordered observed transition (or event) times less than τ across all individuals i in the sample; thus, $\overline{Y}_{ik}(t_{(s)})dN_{ikl}(t_{(s)}) = 1$ for any $s = 1, \ldots, S$, $i = 1, \ldots, n$ and $k, l \in \mathcal{S}$. Since the nonparametric Nelson-Aalen estimator for $\Lambda_{kl}(t) = \int_0^t d\Lambda_{kl}(u)$ based on

$$d\widehat{\Lambda}_{kl}(t) = \frac{\sum_{i=1}^n \overline{Y}_{ik}(t)dN_{ikl}(t)}{\sum_{i=1}^n \overline{Y}_{ik}(t)}, \qquad k < l \ , \tag{5.4.3}$$

is a step function with jumps only at times $t_{(s)}$ for which a $k \longrightarrow l$ transition was observed, $\widehat{D}(\tau|X)$ can be written as

$$\widehat{D}(\tau|X) = \sum_{k\in\mathcal{S}} u_k \sum_{s:t_{(s)}\leq\tau} \widehat{P}_{0k}(0, t_{(s)}|X)\Delta t_{(s)} \ , \tag{5.4.4}$$

where $\Delta t_{(s)} := t_{(s+1)} - t_{(s)}$ for $s = 0, 1, \ldots, S$ with $t_{(0)} := 0$ and $t_{(S+1)} := \tau$. Simple algebra gives

$$\begin{aligned}\text{var}\big(\widehat{D}(\tau|X)\big) = &\sum_{k\in\mathcal{S}} u_k^2 \,\text{var}\bigg(\sum_{s:t_{(s)}\leq\tau} \widehat{P}_{0k}(0, t_{(s)}|X)\Delta t_{(s)}\bigg) \\ &+ \sum_{k\in\mathcal{S}} \sum_{\widetilde{k}\in\mathcal{S},\widetilde{k}\neq k} u_k u_{\widetilde{k}} \,\text{cov}\bigg(\sum_{s:t_{(s)}\leq\tau} \widehat{P}_{0k}(0, t_{(s)}|X)\Delta t_{(s)}, \sum_{s:t_{(s)}\leq\tau} \widehat{P}_{0\widetilde{k}}(0, t_{(s)}|X)\Delta t_{(s)}\bigg) \ .\end{aligned} \tag{5.4.5}$$

The first expression of the right hand side in (5.4.5) can be further derived as

$$\begin{aligned}\text{var}\bigg(\sum_{s:t_{(s)}\leq\tau} \widehat{P}_{0k}(0, t_{(s)}|X)\Delta t_{(s)}\bigg) = &\sum_{s:t_{(s)}\leq\tau} \Delta t_{(s)}^2 \text{var}\big(\widehat{P}_{0k}(0, t_{(s)}|X)\big) \\ &+ \sum_{s:t_{(s)}\leq\tau} \sum_{\widetilde{s}:t_{(\widetilde{s})}\leq\tau,\widetilde{s}\neq s} \Delta t_{(s)}\Delta t_{(\widetilde{s})} \,\text{cov}\big(\widehat{P}_{0k}(0, t_{(s)}|X), \widehat{P}_{0k}(0, t_{(\widetilde{s})}|X)\big) \ .\end{aligned}$$

Similarly,

$$\begin{aligned}
&\text{cov}\Big(\sum_{s:t_{(s)}\le\tau}\widehat{P}_{0k}(0,t_{(s)}|X)\Delta t_{(s)},\sum_{s:t_{(s)}\le\tau}\widehat{P}_{0\widetilde{k}}(0,t_{(s)}|X)\Delta t_{(s)}\Big)\\
&\quad=\sum_{s:t_{(s)}\le\tau}\Delta t_{(s)}^2\text{cov}\big(\widehat{P}_{0k}(0,t_{(s)}|X),\widehat{P}_{0\widetilde{k}}(0,t_{(s)}|X)\big)\\
&\qquad\qquad+\sum_{s:t_{(s)}\le\tau}\sum_{\widetilde{s}:t_{(\widetilde{s})}\le\tau,\widetilde{s}\ne s}\Delta t_{(s)}\Delta t_{(\widetilde{s})}\,\text{cov}\big(\widehat{P}_{0k}(0,t_{(s)}|X),\widehat{P}_{0\widetilde{k}}(0,t_{(\widetilde{s})}|X)\big)\ .
\end{aligned}$$

Andersen et al. (1993) derived an expression for the covariance of estimates of the transition probabilities at $u < t$ for pairs of states under a Markov process, but robust variance estimates (that do not assume the process is Markov) are preferred. The Aalen-Johansen estimates for $P_{0k}(0,t)$ are consistent under non-Markov processes, provided censoring is not state-dependent (Glidden, 2002).

Bootstrap and jackknife procedures offer alternative nonparametric approaches to variance estimation (Efron, 1981; Efron and Gong, 1983). Bootstrapping was used by Cook et al. (2003) and Cook and Lawless (2018, Section 4.2.3) to obtain standard errors for $\widehat{D}(t|X)$ in the analysis of the IBCSG trial. We explore the jackknife procedure in the next sections.

5.4.1 Variance estimation using the jackknife procedure

We first provide a brief review on the jackknife procedure in the context of event history data. Consider one of the multistate processes $\{Z(t), 0 \le t\}$ of Figure 5.1 as before, and let

$$\mathcal{O}_i = \{dN_{ikl}(\cdot), \overline{Y}_{ik}(\cdot), k < l, X\}$$

denote observed data for individual i, $i = 1,\ldots,n$. We further let $w = (w_1, w_2, \ldots, w_n)'$ be an $n \times 1$ vector of individual-specific, time-constant weights $w_i \ge 0$, and define

$$d\widehat{\Lambda}_{kl}(t;w) = \frac{\sum_{i=1}^n w_i \overline{Y}_{ik}(t) dN_{ikl}(t)}{\sum_{i=1}^n w_i \overline{Y}_{ik}(t)}\ , \qquad k < l \tag{5.4.6}$$

as the "weighted" Nelson-Aalen (NA) estimator for a $k \longrightarrow l$ transition. If $\{Z(t), 0 \le t\}$ is Markov, the "weighted" nonparametric Aalen-Johansen (AJ) estimator for the transition probability

matrix $\mathbb{P}(u,t) = \left[P_{kl}(u,t)\right]_{k,l\in\mathcal{S}}$ can in turn be defined as

$$\widehat{\mathbb{P}}(u,t;w) = \prod_{u\leq s\leq t}\left(\mathbb{I} + d\widehat{\mathbb{A}}(s;w)\right) , \tag{5.4.7}$$

where $d\widehat{\mathbb{A}}(s;w) = \left[d\widehat{\Lambda}_{kl}(s;w)\right]_{k,l\in\mathcal{S}}$ and $\mathbb{I}$ denotes the identity matrix (Cox and Miller, 1965). Throughout this chapter we let $\widetilde{w} := (1,1,\ldots,1)'$ denote the $n\times 1$ weight vector of ones for notational convenience. The standard NA and AJ estimators are obtained from the more general representations (5.4.6) and (5.4.7) when $w = \widetilde{w}$:

$$\widehat{\mathbb{P}}(u,t) = \widehat{\mathbb{P}}(u,t;w=(1,1,\ldots,1)) = \widehat{\mathbb{P}}(u,t;\widetilde{w}) ,$$

i.e., when estimation is based on all observations $\mathcal{O} = (\mathcal{O}_1,\ldots,\mathcal{O}_{i-1},\mathcal{O}_i,\mathcal{O}_{i+1},\ldots,\mathcal{O}_n)$. The idea behind the ordinary jackknife (OJK) technique is to exclude $\mathcal{O}_i$ from the sample $\mathcal{O}$ and estimate $d\mathbb{A}(t)$ only based on the $n-1$ remaining observations $\mathcal{O}^{(i)} = (\mathcal{O}_1,\ldots,\mathcal{O}_{i-1},\mathcal{O}_{i+1},\ldots,\mathcal{O}_n)$ (Miller, 1974), which gives

$$d\widehat{\Lambda}_{kl}^{(i)}(t) = \frac{\sum_{i'=1,i'\neq i}^{n} w_{i'}\overline{Y}_{i'k}(t)dN_{i'kl}(t)}{\sum_{i'=1,i'\neq i}^{n} w_{i'}\overline{Y}_{i'k}(t)} , \qquad k<l ,$$

and

$$\widehat{\mathbb{P}}^{(i)}(u,t) = \prod_{u\leq s\leq t}\left(\mathbb{I} + d\widehat{\mathbb{A}}^{(i)}(s;w)\right)$$

for $d\widehat{\mathbb{A}}^{(i)}(s;w) = \left[d\widehat{\Lambda}_{kl}^{(i)}(s;w)\right]_{k,l\in\mathcal{S}}$. Omitting $\mathcal{O}_i$ from estimation is equivalent to assigning a weight of 0 to individual i (i.e. setting $w_i = 0$). Thus, if $w^{(i)} = (1,\ldots,1,0,1,\ldots,1)'$ is the $n\times 1$ vector of weights with 0 at the ith entry and 1 in all other entries, we obtain $d\widehat{\Lambda}_{kl}^{(i)}(t) = d\widehat{\Lambda}_{kl}(t;w^{(i)})$ and

$$\widehat{\mathbb{P}}^{(i)}(u,t) = \widehat{\mathbb{P}}(u,t;w^{(i)}) , \qquad i=1,\ldots,n .$$

The (k,l) entry of $\widehat{\mathbb{P}}^{(i)}(u,t)$ is given by $\widehat{P}_{kl}^{(i)}(u,t) = \widehat{P}_{kl}(u,t;w^{(i)})$.

5.4.1.1 Ordinary jackknife variance estimator for $\widehat{D}(\tau)$

Individual i's jackknife influence value on the estimated state k occupancy probability at time $t_{(s)}$, defined as

$$\widehat{P}_{0k}^{(i)}(0, t_{(s)}) - \widehat{P}_{0k}(0, t_{(s)}) \quad \text{or} \quad \widehat{P}_{0k}(0, t_{(s)}; w^{(i)}) - \widehat{P}_{0k}(0, t_{(s)}; \widetilde{w}) \ , \tag{5.4.8}$$

can be used to derive the ordinary jackknife (OJK) variance estimator for $\widehat{D}(\tau)$. For convenience we suppress the dependence on the treatment indicator X and simply write

$$\widehat{D}(\tau) = \sum_{k \in \mathcal{S}} u_k \sum_{s: t_{(s)} \leq \tau} \widehat{P}_{0k}(0, t_{(s)}) \Delta t_{(s)} = \sum_{k \in \mathcal{S}} u_k \widehat{\mu}_k(\tau) \ ,$$

where $\widehat{\mu}_k(\tau) := \sum_{s: t_{(s)} \leq \tau} \widehat{P}_{0k}(0, t_{(s)}) \Delta t_{(s)}$ is the area under the estimated $\widehat{P}_{0k}(0, t)$ curve over $0 \leq t \leq \tau$. Since $\widehat{\mu}_k(\tau)$ is simply the areas of rectangles, individual i's jackknife influence value on $\widehat{\mu}_k(\tau)$ is

$$\begin{aligned} \widehat{\mu}_k^{(i)}(\tau) - \widehat{\mu}_k(\tau) &= \widehat{\mu}_k(\tau; w^{(i)}) - \widehat{\mu}_k(\tau; \widetilde{w}) \\ &= \sum_{s: t_{(s)} \leq \tau} \big(\widehat{P}_{0k}(0, t_{(s)}; w^{(i)}) - \widehat{P}_{0k}(0, t_{(s)}; \widetilde{w}) \big) \Delta t_{(s)} \ , \end{aligned} \tag{5.4.9}$$

with resulting ordinary jackknife variance and covariance estimates for $\widehat{\mu}_k(\tau)$ obtained as

$$\begin{aligned} \widehat{\text{var}}_{\text{OJK}}\big(\widehat{\mu}_k(\tau)\big) &= \sum_{i=1}^{n} \big(\widehat{\mu}_k(\tau; w^{(i)}) - \widehat{\mu}_k(\tau; \widetilde{w}) \big)^2 \\ &\overset{(5.4.9)}{=} \sum_{i=1}^{n} \Bigg(\sum_{s: t_{(s)} \leq \tau} \Delta t_{(s)} \big[\widehat{P}_{0k}(0, t_{(s)}; w^{(i)}) - \widehat{P}_{0k}(0, t_{(s)}; \widetilde{w}) \big] \Bigg)^2 \ , \end{aligned}$$

and

$$\begin{aligned} \widehat{\text{cov}}_{\text{OJK}}\big(\widehat{\mu}_k(\tau), \widehat{\mu}_{\widetilde{k}}(\tau)\big) &= \sum_{i=1}^{n} \big(\widehat{\mu}_k(\tau; w^{(i)}) - \widehat{\mu}_k(\tau; \widetilde{w}) \big) \big(\widehat{\mu}_{\widetilde{k}}(\tau; w^{(i)}) - \widehat{\mu}_{\widetilde{k}}(\tau; \widetilde{w}) \big) \\ &\overset{(5.4.9)}{=} \sum_{i=1}^{n} \Bigg(\sum_{s: t_{(s)} \leq \tau} \Delta t_{(s)} \big[\widehat{P}_{0k}(0, t_{(s)}; w^{(i)}) - \widehat{P}_{0k}(0, t_{(s)}; \widetilde{w}) \big] \Bigg) \cdot \\ &\qquad \Bigg(\sum_{s: t_{(s)} \leq \tau} \Delta t_{(s)} \big[\widehat{P}_{0\widetilde{k}}(0, t_{(s)}; w^{(i)}) - \widehat{P}_{0\widetilde{k}}(0, t_{(s)}; \widetilde{w}) \big] \Bigg) \ . \end{aligned}$$

Finally,

$$\widehat{\text{var}}_{\text{OJK}}\big(\widehat{D}(\tau)\big) = \sum_{k\in\mathcal{S}} u_k^2 \widehat{\text{var}}_{\text{OJK}}\big(\widehat{\mu}_k(\tau)\big) + \sum_{k\in\mathcal{S},k\neq\widetilde{k}} u_k u_{\widetilde{k}}\, \widehat{\text{cov}}_{\text{OJK}}\big(\widehat{\mu}_k(\tau), \widehat{\mu}_{\widetilde{k}}(\tau)\big)\ . \tag{5.4.10}$$

Calculation of (5.4.10) requires re-estimation of the Aalen-Johansen estimates for $\mathbb{P}(0, t_{(s)})$, or more specifically, the Nelson-Aalen estimates for $d\mathbb{A}(t_{(s)})$ for each leave-one-out sample $\mathcal{O}^{(i)} = (\mathcal{O}_1, \ldots, \mathcal{O}_{i-1}, \mathcal{O}_{i+1}, \ldots, \mathcal{O}_n), i = 1, \ldots, n$. This can be computing-intensive, especially when the sample size n is large and the state space $\mathcal{S}$ of the multistate model has a high cardinality. To overcome this limitation, the terms (5.4.8) in $\widehat{\text{var}}_{\text{OJK}}\big(\widehat{\mu}_k(\tau)\big)$ and $\widehat{\text{cov}}_{\text{OJK}}\big(\widehat{\mu}_k(\tau), \widehat{\mu}_{\widetilde{k}}(\tau)\big)$ can be approximated by first-order Taylor series expansion. This is the idea of Jaeckel (1972)'s infinitesimal jackknife (IJK), which we consider in the next section.

The IJK procedure has also gained popularity in other areas of survival and event history analysis. Parner and Overgaard (2023), for instance, approximate the OJK pseudo-observations by IJK pseudo-observations when estimating unknown parameters from generalized estimating equations in regression analyses of marginal process features, such as restricted mean times.

5.4.1.2 Infinitesimal jackknife variance estimator for $\widehat{D}(\tau)$

Recall from the previous sections that $w^{(i)} = (1, \ldots, 1, 0, 1, \ldots, 1)'$ and $\widetilde{w} = (1, \ldots, 1, 1, 1, \ldots, 1)'$. By first-order Taylor series expansion of $\widehat{P}_{0k}(0, t_{(s)}; w^{(i)})$ about $\widetilde{w}$ and letting

$$\Delta w_i := w^{(i)} - \widetilde{w} = (0, \ldots, 0, -1, 0, \ldots, 0)'\ ,$$

we have

$$\widehat{P}_{0k}(0, t_{(s)}; w^{(i)}) \simeq \widehat{P}_{0k}(0, t_{(s)}; \widetilde{w}) + \left.\frac{\partial \widehat{P}_{0k}(0, t_{(s)}; w)}{\partial w}\right|_{w=\widetilde{w}} \cdot \Delta w_i\ .$$

But since

$$\frac{\partial \widehat{P}_{0k}(0, t_{(s)}; w)}{\partial w} = \left(\frac{\partial \widehat{P}_{0k}(0, t_{(s)}; w)}{\partial w_i} : i = 1, \ldots, n\right)$$

and Δw_i is a $n \times 1$ vector with zeros in $n-1$ entries and -1 in the ith entry (Giordano et al., 2020), we obtain

$$\widehat{P}_{0k}(0, t_{(s)}; w^{(i)}) \simeq \widehat{P}_{0k}(0, t_{(s)}; \widetilde{w}) - \left.\frac{\partial \widehat{P}_{0k}(0, t_{(s)}; w)}{\partial w_i}\right|_{w=\widetilde{w}}\ . \tag{5.4.11}$$

If

$$\widehat{\mu}_k(\tau; w) = \sum_{s:t_{(s)}\le\tau} \widehat{P}_{0k}(0, t_{(s)}; w)\Delta t_{(s)} \ ,$$

then we find

$$\begin{aligned}\widehat{\mu}_k(\tau; w^{(i)}) &\overset{(5.4.11)}{\simeq} \sum_{s:t_{(s)}\le\tau} \widehat{P}_{0k}(0, t_{(s)}; \widetilde{w})\Delta t_{(s)} - \sum_{s:t_{(s)}\le\tau} \left[\frac{\partial \widehat{P}_{0k}(0, t_{(s)}; w)}{\partial w_i}\bigg|_{w=\widetilde{w}}\right]\Delta t_{(s)} \\ &= \widehat{\mu}_k(\tau; \widetilde{w}) - \sum_{s:t_{(s)}\le\tau} \left[\frac{\partial \widehat{P}_{0k}(0, t_{(s)}; w)}{\partial w_i}\bigg|_{w=\widetilde{w}}\right]\Delta t_{(s)} \end{aligned} \tag{5.4.12}$$

Consequently,

$$\widehat{\text{var}}_{\text{IJK}}\big(\widehat{D}(\tau)\big) = \sum_{k\in\mathcal{S}} u_k^2 \widehat{\text{var}}_{\text{IJK}}\big(\widehat{\mu}_k(\tau)\big) + \sum_{k\in\mathcal{S}, k\neq\widetilde{k}} u_k u_{\widetilde{k}}\, \widehat{\text{cov}}_{\text{IJK}}\big(\widehat{\mu}_k(\tau), \widehat{\mu}_{\widetilde{k}}(\tau)\big) \ , \tag{5.4.13}$$

where

$$\begin{aligned}\widehat{\text{var}}_{\text{IJK}}\big(\widehat{\mu}_k(\tau)\big) &= \sum_{i=1}^{n} \big(\widehat{\mu}_k(\tau; w^{(i)}) - \widehat{\mu}_k(\tau; \widetilde{w})\big)^2 \\ &\overset{(5.4.12)}{\approx} \sum_{i=1}^{n} \bigg(\sum_{s:t_{(s)}\le\tau} \Delta t_{(s)} \left[\frac{\partial \widehat{P}_{0k}(0, t_{(s)}; w)}{\partial w_i}\bigg|_{w=\widetilde{w}}\right]\bigg)^2 \ , \end{aligned} \tag{5.4.14}$$

and

$$\begin{aligned}&\widehat{\text{cov}}_{\text{IJK}}\big(\widehat{\mu}_k(\tau), \widehat{\mu}_{\widetilde{k}}(\tau)\big) = \sum_{i=1}^{n} \big(\widehat{\mu}_k(\tau; w^{(i)}) - \widehat{\mu}_k(\tau; \widetilde{w})\big)\big(\widehat{\mu}_{\widetilde{k}}(\tau; w^{(i)}) - \widehat{\mu}_{\widetilde{k}}(\tau; \widetilde{w})\big) \\ &\overset{(5.4.12)}{\approx} \sum_{i=1}^{n} \bigg(\sum_{s:t_{(s)}\le\tau} \Delta t_{(s)} \left[\frac{\partial \widehat{P}_{0k}(0, t_{(s)}; w)}{\partial w_i}\bigg|_{w=\widetilde{w}}\right]\bigg)\bigg(\sum_{s:t_{(s)}\le\tau} \Delta t_{(s)} \left[\frac{\partial \widehat{P}_{0\widetilde{k}}(0, t_{(s)}; w)}{\partial w_i}\bigg|_{w=\widetilde{w}}\right]\bigg) \ . \end{aligned}$$

We let

$$\widehat{I}_{ik}(t_{(s)}) := \frac{\partial \widehat{P}_{0k}(0, t_{(s)}; w)}{\partial w_i}\bigg|_{w=\widetilde{w}} , \qquad k \in \mathcal{S} \ ,$$

be individual i's so-called influence value on $\widehat{P}_{0k}(0,t_{(s)})$ at time $t_{(s)}$, $i=1,\ldots,n$, evaluated at $w=\widetilde{w}=(1,\ldots,1,1,1,\ldots,1)'$ (i.e. $w_i=1$), in which case (5.4.13) reduces to

$$\widehat{\text{var}}_{\text{IJK}}\big(\widehat{D}(\tau)\big)=\sum_{k\in\mathcal{S}}u_k^2\sum_{i=1}^{n}\Big(\sum_{s:t_{(s)}\le\tau}\widehat{I}_{ik}(t_{(s)})\Delta t_{(s)}\Big)^2 + \sum_{k\in\mathcal{S},k\neq\widetilde{k}}u_k u_{\widetilde{k}}\sum_{i=1}^{n}\Big(\sum_{s:t_{(s)}\le\tau}\widehat{I}_{ik}(t_{(s)})\Delta t_{(s)}\Big)\Big(\sum_{s:t_{(s)}\le\tau}\widehat{I}_{i\widetilde{k}}(t_{(s)})\Delta t_{(s)}\Big)\,.$$

The IJK procedure has the advantage that it is not necessary to recalculate $d\widehat{\mathbb{A}}(t_{(s)};w^{(i)})$ for each choice of $w^{(i)}$. Specifically calculation of $\widehat{\text{var}}_{\text{IJK}}\big(\widehat{D}(\tau)\big)$ "only" requires estimation of $n\times(S+1)$ influence matrices of the form

$$\widehat{\mathbb{J}}_k=\big[\widehat{I}_{ik}(t_{(s)})\big]=\begin{pmatrix}\widehat{I}_{1k}(t_{(0)}) & \widehat{I}_{1k}(t_{(1)}) & \widehat{I}_{1k}(t_{(2)}) & \ldots & \ldots & \widehat{I}_{1k}(t_{(S)})\\ \widehat{I}_{2k}(t_{(0)}) & \widehat{I}_{2k}(t_{(1)}) & \widehat{I}_{2k}(t_{(2)}) & \ldots & \ldots & \widehat{I}_{2k}(t_{(S)})\\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\ \widehat{I}_{nk}(t_{(0)}) & \widehat{I}_{nk}(t_{(1)}) & \widehat{I}_{nk}(t_{(2)}) & \ldots & \ldots & \widehat{I}_{nk}(t_{(S)})\end{pmatrix}\,,\qquad k\in\mathcal{S}\,,$$

in the full sample $\mathcal{O}$.

Since $\widehat{\mathbb{P}}(0,t_{(s)};w)=\widehat{\mathbb{P}}(0,t_{(s-1)};w)(\mathbb{I}+d\widehat{\mathbb{A}}(t_{(s)};w))$, we obtain the recursive relation

$$\begin{aligned}\frac{\partial\widehat{\mathbb{P}}(0,t_{(1)};w)}{\partial w_i}&=\frac{\partial d\widehat{\mathbb{A}}(t_{(1)};w)}{\partial w_i}\,,\\ \frac{\partial\widehat{\mathbb{P}}(0,t_{(2)};w)}{\partial w_i}&=\frac{\partial\widehat{\mathbb{P}}(0,t_{(1)};w)}{\partial w_i}(\mathbb{I}+d\widehat{\mathbb{A}}(t_{(2)};w))+\widehat{\mathbb{P}}(0,t_{(1)};w)\frac{\partial d\widehat{\mathbb{A}}(t_{(2)};w)}{\partial w_i}\,,\\ &\vdots\\ &\vdots\\ \frac{\partial\widehat{\mathbb{P}}(0,t_{(s)};w)}{\partial w_i}&=\frac{\partial\widehat{\mathbb{P}}(0,t_{(s-1)};w)}{\partial w_i}(\mathbb{I}+d\widehat{\mathbb{A}}(t_{(s)};w))+\widehat{\mathbb{P}}(0,t_{(s-1)};w)\frac{\partial d\widehat{\mathbb{A}}(t_{(s)};w)}{\partial w_i}\,,\end{aligned}$$

where

$$\frac{\partial}{\partial w_i}\big(d\widehat{\mathbb{A}}(t;w)\big)=\Big(\frac{\partial}{\partial w_i}d\widehat{\Lambda}_{kl}(t;w)\Big)_{k,l\in\mathcal{S}}\tag{5.4.15}$$

is the $|S| \times |S|$ derivative matrix of $d\widehat{\mathbb{A}}(t;w)$ with respect to w_i. Note that the $(1,k)$ entry of the matrix $\partial\widehat{\mathbb{P}}(0,t_{(s)};w)/\partial w_i$ evaluated at $w=\widetilde{w}$ provides an estimate of $I_{ik}(t_{(s)})$:

$$\left[\left.\frac{\partial\widehat{\mathbb{P}}(0,t_{(s)};w)}{\partial w_i}\right|_{w=\widetilde{w}}\right]_{1k} = \left(\widehat{I}_{i1}(t_{(s)}) \quad \widehat{I}_{i2}(t_{(s)}) \quad \widehat{I}_{i3}(t_{(s)}) \quad \ldots \quad \widehat{I}_{iK}(t_{(s)}) \quad \widehat{I}_{i(K+1)}(t_{(s)})\right) ,$$

When there are no ties and the overall counting process $dN(t) = (dN_{kl}(t) : k < l)$ only jumps once at time t, the (k,l) entry of the derivative matrix (5.4.15) can take one of the following expressions ($k < l$):

$$\left.\frac{\partial d\widehat{\Lambda}_{kl}(t;w)}{\partial w_i}\right|_{w=\widetilde{w}} = \frac{\overline{Y}_{ik}(t)(dN_{ikl}(t) - d\widehat{\Lambda}_{kl}(t;\widetilde{w}))}{\overline{Y}_{\cdot k}(t)}$$

$$= \begin{cases} \dfrac{1 - d\widehat{\Lambda}_{kl}(t;\widetilde{w})}{\overline{Y}_{\cdot k}(t)} = \dfrac{\overline{Y}_{\cdot k}(t) - 1}{\overline{Y}_{\cdot k}(t)^2} & , \text{ if } \overline{Y}_{ik}(t) = 1, dN_{ikl}(t) = 1, dN_{i'kl}(t) = 0 \\ \dfrac{-d\widehat{\Lambda}_{kl}(t;\widetilde{w})}{\overline{Y}_{\cdot k}(t)} = \dfrac{-1}{\overline{Y}_{\cdot k}(t)^2} & , \text{ if } \overline{Y}_{ik}(t) = 1, dN_{ikl}(t) = 0, dN_{i'kl}(t) = 1 \\ 0 & , \text{ if } \overline{Y}_{ik}(t) = 1, dN_{ikl}(t) = 0, dN_{i'kl}(t) = 0 \\ 0 & , \text{ if } \overline{Y}_{ik}(t) = 0 \end{cases} , \quad (5.4.16)$$

where $\overline{Y}_{\cdot k}(t) = \sum_{i=1}^{n} \overline{Y}_{ik}(t)$ and for some individual $i' \neq i$. For case 1, $\overline{Y}_{ik}(t) = 1, dN_{ikl}(t) = 1$ and $dN_{i'kl}(t) = 0$, so there is an observed $k \longrightarrow l$ transition at time t for individual i. For case 2, $\overline{Y}_{ik}(t) = 1, dN_{ikl}(t) = 0$ and $dN_{i'kl}(t) = 1$, so the $k \longrightarrow l$ transition at t has not been made by individual i, but by individual i'; i is however at-risk for a $k \longrightarrow l$ transition. For case 3, no $k \longrightarrow l$ transition happens at t, in which case $dN_{ikl}(t) = 0$ and $d\widehat{\Lambda}_{kl}(t;w) = 0$. For case 4, $\overline{Y}_{ik}(t) = 0$ so individual i is no longer at-risk for a $k \longrightarrow l$ transition at t.

Appendix D.3 derives explicit expressions for the influence values $I_{ik}(t_{(s)})$ based on (5.4.16) for the setting of an illness-death process. Appendix D.4 demonstrates how to implement the IJK procedure by standard software (Therneau, 2024).

5.4.2 Empirical studies of finite sample properties

Simulation studies were conducted to examine the finite sample properties of $\widehat{\text{var}}_{\text{IJK}}(\widehat{\beta}(\tau))$ and associated tests for $H_0 : \beta(\tau) \leq 0$ vs. $H_1 : \beta(\tau) > 0$ based on

$$\frac{\widehat{\beta}(\tau)}{\sqrt{\widehat{\text{var}}_{\text{IJK}}\big(\widehat{\beta}(\tau)\big)}} = \frac{\widehat{D}(\tau|X=1) - \widehat{D}(\tau|X=0)}{\sqrt{\widehat{\text{var}}_{\text{IJK}}\big(\widehat{D}(\tau|X=1)\big) + \widehat{\text{var}}_{\text{IJK}}\big(\widehat{D}(\tau|X=0)\big)}}\,, \qquad \tau > 0\,. \tag{5.4.17}$$

In the simulations that follow we refer to (5.4.17) as the infinitesimal jackknife (IJK) test statistic. We consider the three-state illness-death model in Figure 5.1(a) for illustration.

5.4.2.1 Simulation setup

We generated data from Markov and mixed-Markov illness-death processes for $n = 100, 250$ and 1000 individuals, following the procedure in Section 2.5 of Cook and Lawless (2018). Individuals are randomized to one of the two treatment groups with probability 0.5, i.e. $X \sim Bernoulli(0.5)$. For the Markov processes, we assume multiplicative transition intensities of the form $\lambda_{kl}(t|X) = \lambda_{kl0}\exp(\gamma_{kl}X), (k,l) \in \{(0,1),(0,2),(1,2)\}$; the intensities are thus time-homogeneous Markov. For the mixed-Markov processes, we assume that $k \longrightarrow l$ transitions occur according to the intensity

$$\begin{aligned}\lambda_{kl}(t|\mathcal{H}(t),U) &= \lim_{\Delta t \downarrow 0} \frac{P(Z(t+\Delta t^-) = l \mid Z(t^-) = k, \mathcal{H}(t), U)}{\Delta t} \\ &= U\lambda_{kl0}\exp(\gamma_{kl}X)\,,\end{aligned} \tag{5.4.18}$$

where $\mathcal{H}(t) = \{Z(u) : 0 < u < t, X\}$ denotes the process history including X, and U is a random effect that follows a Gamma distribution with mean $\mathbb{E}(U) = 1$ and $\text{Var}(U) = \phi$; the process is thus Markov given X and U. We let $G_U(u) = P(U \leq u)$ denote the cumulative distribution function of U, with associated density function $dG_U(u) = g_U(u)du$. The resulting "marginal" intensity (averaged over U)

$$\begin{aligned}\lambda_{kl}(t|\mathcal{H}(t)) &= \mathbb{E}_U\big(\lambda_{kl}(t|\mathcal{H}(t),U) \mid \mathcal{H}(t)\big) = \lambda_{kl0}\exp(\gamma_{kl}X)\mathbb{E}_U(U|\mathcal{H}(t^-)) \\ &= \frac{\lambda_{kl0}\exp(\gamma_{kl}X)}{\mathbb{E}_U(P(Z(u), 0 < u < t|X,U))}\int_0^\infty uP(Z(u), 0 < u < t|X,u)dG_U(u)\end{aligned} \tag{5.4.19}$$

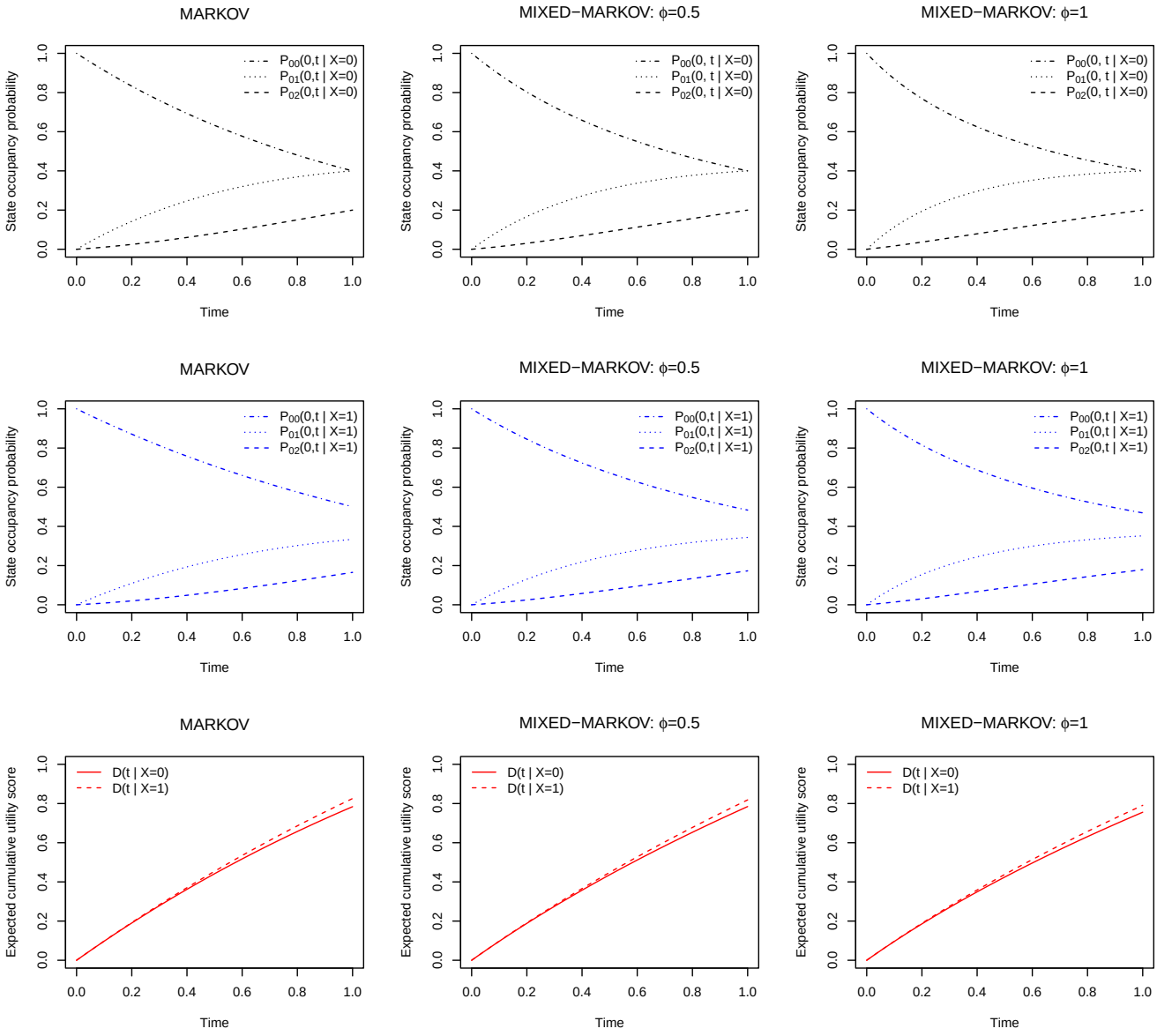


Figure 5.6: Plots of $P_{0k}(0,t|X=0)$, $P_{0k}(0,t|X=1)$ and $D(t|X)=\mathbb{E}(U(t)|X)$ over time when the true illness-death process is either Markov (left panels) or mixed-Markov (middle and right panels) with $\exp(\gamma_{01})=0.75, \exp(\gamma_{02})=0.8$ and $\exp(\gamma_{12})=1$, $k=0,1,2$ and $X=0,1$; $\tau=1, \pi_T=0.6, \pi_T^1=0.9, \pi_D=0.2$, $\phi=0.5, 1.0$, $u_0=1, u_1=0.5, u_2=0$.

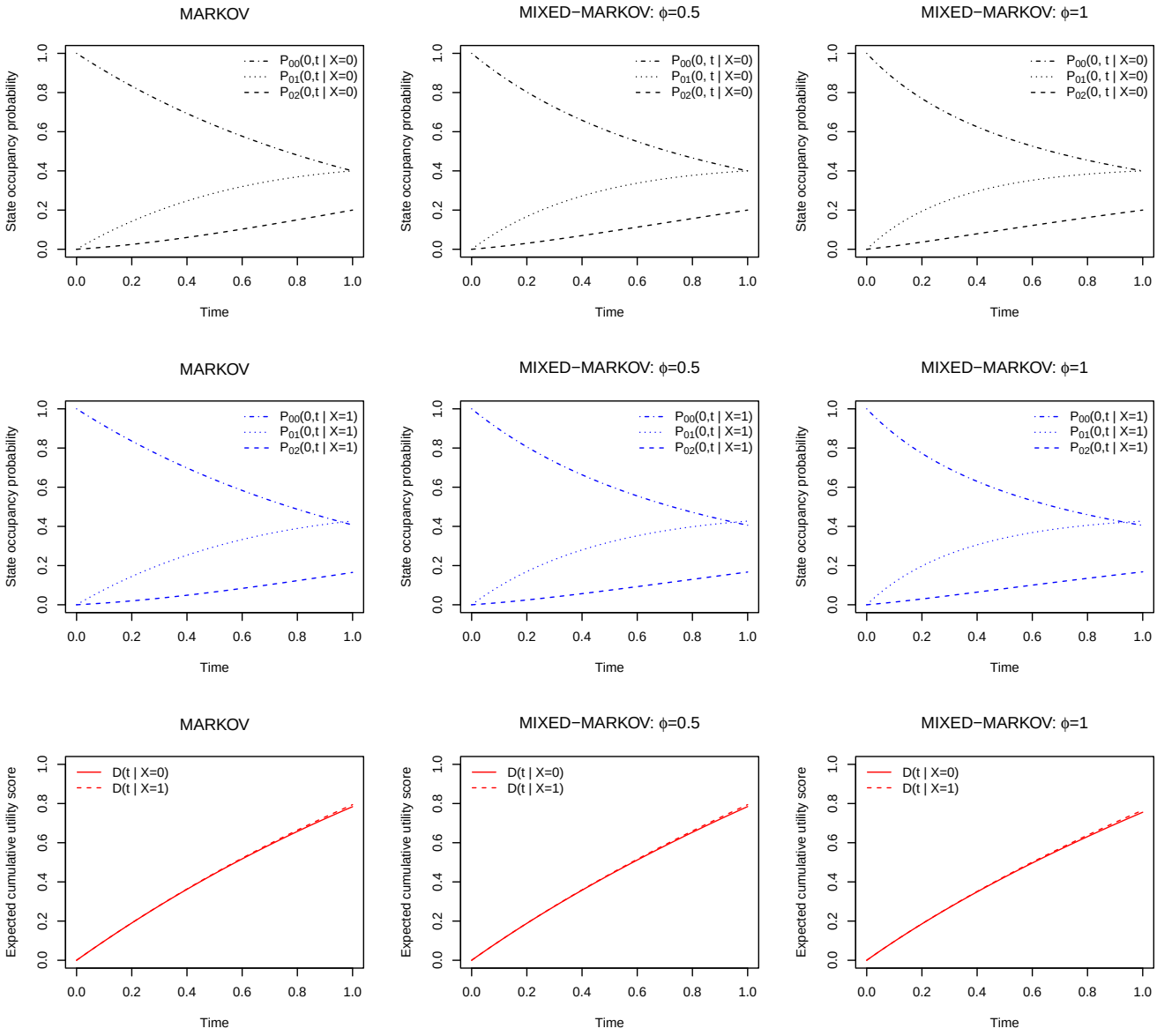


Figure 5.7: Plots of $P_{0k}(0,t|X=0), P_{0k}(0,t|X=1)$ and $D(t|X) = \mathbb{E}(U(t)|X)$ over time when the true illness-death process is either Markov (left panels) or mixed-Markov (middle and right panels) with $\exp(\gamma_{01}) = 1, \exp(\gamma_{02}) = 0.8$ and $\exp(\gamma_{12}) = 0.8$, $k = 0, 1, 2$ and $X = 0, 1$; $\tau = 1, \pi_T = 0.6, \pi_T^1 = 0.9, \pi_D = 0.2$, $\phi = 0.5, 1.0$, $u_0 = 1, u_1 = 0.5, u_2 = 0$.

does however depend on the process history and so the marginal process characterized by (5.4.18) is not Markov.

Recall from Chapter 3 that, in the context of an illness-death model, $T = \inf\{t > 0 : Z(t) \neq 0\}$ represents the exit time of state 0. We choose the administrative censoring time to be $\tau = 1$ and set $\lambda_{010}, \lambda_{020}$ and λ_{120} to satisfy the following constraints: (1) $\pi_T = P(T \leq \tau | X = 0) = 0.6$, so 40% of individuals in the control group are alive and event-free (in state 0) at time τ, (2) $\pi_T^1 = P(T_{01} < T_{02} | T \leq \tau, X = 0) = 0.9$, so 54% of individuals with $X = 0$ are expected to experience the non-fatal event of interest (i.e. pass through state 1) and 6% die event-free (i.e. $0 \longrightarrow 2$ transition), and (3) $\pi_D = P(Z(\tau) = 2 | X = 0) = 0.2$, so 14% of individuals in the control group are expected to die post-event (i.e. $1 \longrightarrow 2$ transition) by time τ. We note that conditions (1)-(3) imply that $P(Z(\tau) = 1 | X = 0) = 0.40$, subject to $\pi_D > \pi_T(1 - \pi_T^1)$ and $\pi_D < \pi_T$; if $\pi_D = \pi_T(1 - \pi_T^1)$, then no $1 \longrightarrow 2$ transitions are observed. If $\pi_D = \pi_T$, then $P(Z(\tau) = 1 | X = 0) = 0$ and no one is expected to survive. For mixed-Markov processes, we set $\phi = 0.5, 1.0$ and determine $\lambda_{010}, \lambda_{020}$ and λ_{120} so that (1) $\pi_T = \mathbb{E}_U(P(T \leq \tau | X = 0, U)) = 0.6$, (2) $\pi_T^1 = \mathbb{E}_U(P(T_{01} < T_{02} | T \leq \tau, X = 0, U)) = 0.9$, and (3) $\pi_D = \mathbb{E}_U(P(Z(\tau) = 2 | X = 0, U)) = 0.2$. We consider 12 different treatment effect scenarios, ranging from no effect of treatment on any transition intensity (i.e. $\exp(\gamma_{01}) = \exp(\gamma_{02}) = \exp(\gamma_{12}) = 1$) to treatment effects on all three intensities (e.g. $\exp(\gamma_{01}) < 1, \exp(\gamma_{02}) < 1, \exp(\gamma_{12}) < 1$). Without loss of generality, we set the utility scores for states $k = 0, 1$ and 2 to $u_0 = 1, u_1 = 0.5, u_2 = 0$. Figures 5.6 and 5.7 contain plots of $P_{0k}(0, t | X = 0)$ (top row), $P_{0k}(0, t | X = 1)$ (middle row) and $D(t|X), X = 0, 1$ (bottom row) over $[0, \tau]$ under the Markov and mixed-Markov models for two particular treatment effect scenarios, $k = 0, 1, 2$. Note that in these plots the extent to which the mixed-Markov models deviate from the Markov assumption is not reflected; here the purpose was rather to calibrate the models so that the state occupancy probabilities by time τ are the same. To accommodate early withdrawal from study we set the censoring intensity to $\lambda_r(t|X) = \lambda_r \exp(\gamma_r X)$, which corresponds to generating the right-censoring time $C_r|X$ from an exponential distribution with rate $\lambda_r e^{\gamma_r X}$; we solve for λ_r to ensure $\pi_r = P(C_r \leq \tau | X = 0) = 0.2$ or 0.4, and set $\exp(\gamma_r) = 1$ or 0.85.

For each data-generating process and parameter setting, we considered estimation of $\beta(\tau)$ by nonparametric Aalen-Johansen methods and conducted the test (5.4.17) with the infinitesimal jackknife procedure of Section 5.4.1.2 adopted for variance estimation of $\widehat{\beta}(\tau)$; the `survfit(...,`

`influence` $=$ `TRUE`) function provides the AJ estimates of $\widehat{P}_{0k}(0, t|X)$ as well as the influence matrices $\widehat{\mathbb{J}}_k = \left[\widehat{I}_{ik}(t_{(s)})\right]$ for each treatment group (Therneau, 2024), based on which $\widehat{\text{var}}_{\text{IJK}}(\widehat{D}(\tau)|X)$ can be calculated using (5.4.13). We report on the difference between the average treatment effect estimates and the true $\beta(\tau)$ (BIAS), the empirical standard error (ESE) of estimates, the average standard error (ASE), the empirical coverage probability (CP) for $\beta(\tau)$ of the nominal one-sided 95% confidence intervals, and the empirical rejection rate (RR) for the right-tail test (5.4.17) for $H_0 : \beta(\tau) \leq 0$ of nominal size $\omega = 0.05$. The last performance measure was calculated as the proportion of $n_{sim} = 10000$ simulations in which the IJK test rejected H_0 at the nominal $\omega = 0.05$ level. With this many simulation if the test rejects the null hypothesis at the nominal $\omega = 5\%$ level, the empirical rejection rate is expected to lie in $[0.0457, 0.0543]$ 95% of the time.

5.4.2.2 Simulation results

Simulation results for the Markov models are shown in Table 5.1 and Table 5.2 for $n_{sim} = 10000$ simulation runs; the former table corresponds to the scenarios where $\pi_r = 0.2, 0.4$ and $\exp(\gamma_r) = 1$, and the latter pertains to the scenarios where $\pi_r = 0.2, 0.4$ and $\exp(\gamma_r) = 0.85$.

The empirical bias of the AJ-based estimator for $\beta(\tau)$ is very small in all scenarios considered; this is not unexpected. The simulations in this section were primarily conducted to evaluate the performance of the IJK standard errors in reflecting the sampling variability in $\widehat{\beta}(\tau)$. The average standard errors agree well with the empirical standard errors when the sample size is either $n = 250$ or $n = 1000$. The empirical coverage probability is in close agreement with the nominal 95% level in such settings. Note that for some scenarios the average standard error is smaller than the empirical standard error when the sample size is $n = 100$; in these settings the empirical coverage probability is slightly outside the acceptable range.
The empirical type I error rates are compatible with the nominal 5% level in most settings; there is a slight inflation of the type I error rate in settings where the sample size is small and the censoring rate is high (e.g. when $\pi_r = 0.4$ and $n = 100$ in Table 5.1). The IJK test (5.4.17) has robust power for detecting treatment effects on $D(\tau)$ across a broad range of alternatives. Naturally, the power increases as the magnitude of the treatment effect $\beta(\tau)$ increases and as the sample size n increases. For fixed values of $\beta(\tau)$ and n, the power decreases with increasing censoring rate.

		MARKOV PROCESSES														
		$n = 100$					$n = 250$					$n = 1000$				
$[e^{\gamma_{01}}, e^{\gamma_{02}}, e^{\gamma_{12}}]$	$\beta(\tau)$	BIAS	ESE	ASE	RR	CP	BIAS	ESE	ASE	RR	CP	BIAS	ESE	ASE	RR	CP
							$\pi_r = 0.20$ and $\exp(\gamma_r) = 1$									
1,1,1	0.0000	0.0006	0.0504	0.0504	0.0500	0.9500	0.0002	0.0320	0.0320	0.0527	0.9473	-0.0002	0.0162	0.0160	0.0517	0.9483
1,1,0.8	0.0047	0.0006	0.0499	0.0498	0.0604	0.9506	0.0001	0.0316	0.0316	0.0693	0.9474	-0.0002	0.0160	0.0158	0.0911	0.9492
1,0.8,1	0.0058	0.0005	0.0504	0.0500	0.0668	0.9487	0.0003	0.0317	0.0317	0.0756	0.9496	-0.0002	0.0161	0.0159	0.1016	0.9479
1,0.8,0.8	0.0105	0.0004	0.0498	0.0494	0.0799	0.9485	0.0003	0.0313	0.0313	0.0997	0.9493	-0.0002	0.0159	0.0157	0.1632	0.9484
0.9,1,1	0.0135	0.0005	0.0510	0.0502	0.0912	0.9468	0.0003	0.0318	0.0319	0.1166	0.9476	-0.0002	0.0161	0.0160	0.2098	0.9508
0.75, 1.1, 1.2	0.0285	0.0004	0.0511	0.0505	0.1477	0.9456	0.0001	0.0323	0.0321	0.2336	0.9486	-0.0001	0.0162	0.0161	0.5492	0.9493
0.75,1,1.2	0.0316	0.0004	0.0510	0.0502	0.1618	0.9440	0.0000	0.0323	0.0319	0.2630	0.9469	-0.0001	0.0162	0.0160	0.6323	0.9478
0.75,0.9,1.2	0.0348	0.0002	0.0512	0.0500	0.1837	0.9405	0.0000	0.0318	0.0317	0.2937	0.9486	-0.0001	0.0161	0.0159	0.7006	0.9470
0.75,1,1	0.0351	0.0004	0.0505	0.0497	0.1798	0.9437	0.0000	0.0320	0.0316	0.3007	0.9476	-0.0001	0.016	0.0158	0.7146	0.9486
0.75,1,0.8	0.0388	0.0003	0.0500	0.0492	0.2011	0.9442	0.0000	0.0317	0.0313	0.3458	0.9477	-0.0001	0.0158	0.0157	0.7963	0.9493
0.75,0.9,0.9	0.0401	0.0002	0.0504	0.0492	0.2177	0.9417	0.0000	0.0313	0.0312	0.3613	0.9494	-0.0001	0.0158	0.0157	0.8199	0.9470
0.75,0.8,1	0.0415	0.0002	0.0502	0.0492	0.2194	0.9438	0.0000	0.0312	0.0312	0.3777	0.9489	-0.0002	0.0156	0.0157	0.8384	0.9514
							$\pi_r = 0.40$ and $\exp(\gamma_r) = 1$									
1,1,1	0.0000	0.0007	0.0529	0.0524	0.0569	0.9431	0.0001	0.0332	0.0332	0.0523	0.9477	-0.0002	0.0168	0.0166	0.0495	0.9505
1,1,0.8	0.0047	0.0005	0.0523	0.0518	0.0681	0.9433	0.0001	0.0329	0.0328	0.0683	0.9474	-0.0002	0.0166	0.0164	0.0865	0.9501
1,0.8,1	0.0058	0.0006	0.0523	0.0521	0.0662	0.9454	0.0001	0.0331	0.0329	0.0710	0.9475	-0.0001	0.0166	0.0165	0.0994	0.9497
1,0.8,0.8	0.0105	0.0004	0.0517	0.0514	0.0794	0.9461	0.0001	0.0327	0.0325	0.0909	0.9484	-0.0002	0.0164	0.0163	0.1599	0.9497
0.9,1,1	0.0135	0.0006	0.0527	0.0522	0.0866	0.9465	0.0002	0.0333	0.0331	0.1126	0.9494	-0.0001	0.0167	0.0166	0.2044	0.9504
0.75, 1.1, 1.2	0.0285	0.0003	0.0533	0.0525	0.1428	0.9432	0.0002	0.0334	0.0333	0.2180	0.9458	-0.0001	0.0168	0.0167	0.5195	0.9484
0.75,1,1.2	0.0316	0.0005	0.0536	0.0522	0.1629	0.9421	0.0000	0.0334	0.0331	0.2505	0.9465	-0.0001	0.0166	0.0166	0.6014	0.9485
0.75,0.9,1.2	0.0348	0.0007	0.0531	0.0520	0.1799	0.9423	0.0000	0.0333	0.0329	0.2794	0.9454	-0.0001	0.0166	0.0165	0.6785	0.9471
0.75,1,1	0.0351	0.0004	0.0531	0.0517	0.1782	0.9410	0.0000	0.0331	0.0328	0.2848	0.9469	-0.0001	0.0164	0.0164	0.6894	0.9478
0.75,1,0.8	0.0388	0.0004	0.0525	0.0512	0.1999	0.9410	0.0000	0.0327	0.0324	0.3275	0.9456	-0.0001	0.0162	0.0163	0.7716	0.9481
0.75,0.9,0.9	0.0401	0.0006	0.0523	0.0512	0.2085	0.9435	0.0000	0.0328	0.0324	0.3487	0.9462	-0.0001	0.0164	0.0163	0.7939	0.9479
0.75,0.8,1	0.0415	0.0005	0.0527	0.0512	0.2164	0.9442	0.0000	0.0328	0.0324	0.3576	0.9471	-0.0002	0.0162	0.0163	0.8161	0.9495

Table 5.1: Empirical properties of the utility-based estimator $\widehat{\beta}(\tau)$ and associated tests of $H_0 : \beta(\tau) \leq 0$ vs. $H_1 : \beta(\tau) > 0$ when the illness-death process is Markov with intensities $\lambda_{kl}(t|X) = \lambda_{kl0} \exp(\gamma_{kl} X)$; $\tau = 1, \pi_T = 0.6, \pi_T^1 = 0.9, \pi_D = 0.2, \pi_r = P(C_r \leq \tau | X = 0) = 0.2$ and $0.4, \exp(\gamma_r) = 1, u_0 = 1, u_1 = 0.5, u_2 = 0, n_{sim} = 10000$ simulation runs.

		MARKOV PROCESSES														
		$n = 100$					$n = 250$					$n = 1000$				
$[e^{\gamma_{01}}, e^{\gamma_{02}}, e^{\gamma_{12}}]$	$\beta(\tau)$	BIAS	ESE	ASE	RR	CP	BIAS	ESE	ASE	RR	CP	BIAS	ESE	ASE	RR	CP
							$\pi_r = 0.20$ and $\exp(\gamma_r) = 0.85$									
1,1,1	0.0000	0.0004	0.0508	0.0503	0.0544	0.9456	0.0000	0.0320	0.0319	0.0498	0.9502	-0.0002	0.0161	0.016	0.0495	0.9505
1,1,0.8	0.0047	0.0004	0.0502	0.0497	0.0654	0.9458	0.0000	0.0317	0.0315	0.0689	0.9507	-0.0002	0.0159	0.0158	0.0903	0.9499
1,0.8,1	0.0058	0.0003	0.0504	0.0499	0.0672	0.9467	0.0002	0.0317	0.0317	0.0739	0.9477	-0.0002	0.016	0.0159	0.1014	0.9505
1,0.8,0.8	0.0105	0.0002	0.0499	0.0493	0.0789	0.9463	0.0002	0.0313	0.0313	0.0984	0.9479	-0.0002	0.0158	0.0157	0.1602	0.9510
0.9,1,1	0.0135	0.0005	0.0502	0.0501	0.0881	0.9481	0.0003	0.0320	0.0318	0.1171	0.9473	-0.0001	0.0160	0.0159	0.2127	0.9485
0.75, 1.1, 1.2	0.0285	0.0004	0.0509	0.0503	0.1485	0.9422	0.0001	0.0326	0.0320	0.2311	0.9481	-0.0001	0.0162	0.0160	0.5492	0.9486
0.75,1,1.2	0.0316	0.0004	0.0513	0.0501	0.1688	0.9436	-0.0001	0.0320	0.0318	0.2624	0.9481	-0.0001	0.0162	0.016	0.6303	0.9470
0.75,0.9,1.2	0.0348	0.0002	0.0504	0.0499	0.1797	0.9455	0.0000	0.0317	0.0317	0.2979	0.9477	-0.0001	0.016	0.0159	0.7039	0.9493
0.75,1,1	0.0351	0.0004	0.0508	0.0496	0.1875	0.9438	-0.0001	0.0317	0.0315	0.3045	0.9488	-0.0001	0.016	0.0158	0.7120	0.9473
0.75,1,0.8	0.0388	0.0003	0.0502	0.0491	0.2060	0.9448	-0.0001	0.0314	0.0312	0.3506	0.9491	-0.0001	0.0159	0.0156	0.7962	0.9482
0.75,0.9,0.9	0.0401	0.0002	0.0497	0.0491	0.2099	0.9463	0.0000	0.0312	0.0312	0.3620	0.9486	-0.0001	0.0157	0.0156	0.8179	0.9499
0.75,0.8,1	0.0415	0.0003	0.0504	0.0491	0.2271	0.9441	-0.0001	0.0314	0.0312	0.3803	0.9496	-0.0002	0.0157	0.0156	0.8385	0.9478
							$\pi_r = 0.40$ and $\exp(\gamma_r) = 0.85$									
1,1,1	0.0000	0.0007	0.0524	0.0521	0.0541	0.9459	0.0001	0.0336	0.033	0.0531	0.9469	-0.0003	0.0168	0.0166	0.0519	0.9481
1,1,0.8	0.0047	0.0006	0.0518	0.0515	0.0643	0.9462	0.0001	0.0332	0.0326	0.0712	0.9474	-0.0003	0.0166	0.0164	0.0913	0.9484
1,0.8,1	0.0058	0.0006	0.0524	0.0518	0.0703	0.9440	0.0001	0.0328	0.0328	0.0703	0.9500	-0.0002	0.0165	0.0164	0.0956	0.9491
1,0.8,0.8	0.0105	0.0004	0.0519	0.0512	0.0809	0.9450	0.0001	0.0324	0.0324	0.0916	0.9509	-0.0002	0.0163	0.0162	0.1581	0.9503
0.9,1,1	0.0135	0.0008	0.0526	0.0519	0.0905	0.9432	0.0002	0.0330	0.0329	0.1059	0.9484	-0.0001	0.0166	0.0165	0.1996	0.9484
0.75, 1.1, 1.2	0.0285	0.0003	0.0533	0.0521	0.1453	0.9412	0.0001	0.0332	0.0331	0.2160	0.9477	-0.0001	0.0168	0.0166	0.5261	0.9481
0.75,1,1.2	0.0316	0.0005	0.0531	0.0519	0.1589	0.9423	-0.0001	0.0333	0.033	0.2464	0.9467	-0.0001	0.0167	0.0165	0.6054	0.9468
0.75,0.9,1.2	0.0348	0.0006	0.0525	0.0517	0.1770	0.9443	0.0000	0.0332	0.0328	0.2844	0.9468	0.0000	0.0165	0.0164	0.6811	0.9484
0.75,1,1	0.0351	0.0004	0.0526	0.0514	0.1770	0.9424	-0.0001	0.0329	0.0326	0.2838	0.9455	-0.0001	0.0166	0.0164	0.6936	0.9485
0.75,1,0.8	0.0388	0.0004	0.0521	0.0509	0.1982	0.9418	-0.0001	0.0326	0.0323	0.3310	0.9468	-0.0001	0.0164	0.0162	0.7704	0.9486
0.75,0.9,0.9	0.0401	0.0005	0.0517	0.0509	0.2072	0.9436	-0.0001	0.0327	0.0323	0.3455	0.9466	0.0000	0.0162	0.0162	0.8005	0.9491
0.75,0.8,1	0.0415	0.0006	0.0519	0.0510	0.2133	0.9439	0.0000	0.0324	0.0323	0.3595	0.9481	-0.0001	0.0163	0.0162	0.8159	0.9501

Table 5.2: Empirical properties of the utility-based estimator $\widehat{\beta}(\tau)$ and associated tests of $H_0 : \beta(\tau) \leq 0$ vs. $H_1 : \beta(\tau) > 0$ when the illness-death process is Markov with intensities $\lambda_{kl}(t|X) = \lambda_{kl0} \exp(\gamma_{kl} X)$; $\tau = 1, \pi_T = 0.6, \pi_T^1 = 0.9, \pi_D = 0.2, \pi_r = P(C_r \leq \tau | X = 0) = 0.2$ and $0.4, \exp(\gamma_r) = 0.85, u_0 = 1, u_1 = 0.5, u_2 = 0, n_{sim} = 10000$ simulation runs.

		MIXED-MARKOV PROCESSES														
		$n = 100$					$n = 250$					$n = 1000$				
$[e^{\gamma_{01}}, e^{\gamma_{02}}, e^{\gamma_{12}}]$	$\beta(\tau)$	BIAS	ESE	ASE	RR	CP	BIAS	ESE	ASE	RR	CP	BIAS	ESE	ASE	RR	CP
$\phi = 0.5:\ \pi_r = 0.20$ and $\exp(\gamma_r) = 1$																
1,1,1	0.0000	-0.0004	0.0529	0.0524	0.0544	0.9456	0.0007	0.0336	0.0333	0.0562	0.9438	0.0001	0.0167	0.0167	0.0521	0.9479
1,1,0.8	0.0049	-0.0004	0.0523	0.0518	0.0639	0.9466	0.0007	0.0332	0.0329	0.0745	0.9432	0.0001	0.0165	0.0165	0.0924	0.9484
1,0.8,1	0.0057	-0.0006	0.0523	0.0521	0.0617	0.9509	0.0007	0.0332	0.0330	0.0747	0.9445	0.0001	0.0165	0.0165	0.0963	0.9498
1,0.8,0.8	0.0106	-0.0007	0.0517	0.0514	0.0749	0.9511	0.0007	0.0328	0.0326	0.0987	0.9454	0.0001	0.0163	0.0163	0.1601	0.9490
0.9,1,1	0.0122	0.0002	0.0528	0.0522	0.0820	0.9480	-0.0002	0.0334	0.0332	0.1033	0.9489	0.0002	0.0167	0.0166	0.1828	0.9474
0.75,1.1,1.2	0.0251	-0.0001	0.0531	0.0526	0.1280	0.9462	0.0001	0.0336	0.0334	0.1879	0.9454	0.0001	0.0169	0.0168	0.4451	0.9466
0.75,1,1.2	0.0283	-0.0001	0.0529	0.0524	0.1383	0.9469	0.0006	0.0333	0.0333	0.2217	0.9475	0.0000	0.0166	0.0167	0.5219	0.9495
0.75,0.9,1.2	0.0315	-0.0004	0.0529	0.0522	0.1560	0.9466	0.0006	0.0333	0.0331	0.2564	0.9480	0.0001	0.0165	0.0166	0.6023	0.9518
0.75,1,1	0.0320	-0.0001	0.0524	0.0519	0.1556	0.9476	0.0006	0.0330	0.0330	0.2597	0.9479	0.0000	0.0165	0.0165	0.6142	0.9496
0.75,1,0.8	0.0360	-0.0001	0.0519	0.0514	0.1768	0.9481	0.0006	0.0327	0.0326	0.3040	0.9470	0.0000	0.0163	0.0163	0.7112	0.9498
0.75,0.9,0.9	0.0372	-0.0004	0.0521	0.0514	0.1805	0.9461	0.0006	0.0328	0.0326	0.3205	0.9486	0.0000	0.0163	0.0164	0.7386	0.9522
0.75,0.8,1	0.0385	-0.0003	0.0520	0.0514	0.1886	0.9483	0.0006	0.0329	0.0326	0.3334	0.9443	0.0000	0.0162	0.0164	0.7677	0.9487
$\phi = 0.5:\ \pi_r = 0.40$ and $\exp(\gamma_r) = 1$																
1,1,1	0.0000	-0.0008	0.0556	0.0543	0.0533	0.9467	-0.0001	0.0346	0.0344	0.0496	0.9504	0.0003	0.0172	0.0173	0.0510	0.9490
1,1,0.8	0.0049	-0.0008	0.0550	0.0537	0.0624	0.9478	-0.0001	0.0342	0.0340	0.0673	0.9502	0.0003	0.0170	0.0171	0.0916	0.9495
1,0.8,1	0.0057	-0.0006	0.0547	0.0540	0.0671	0.9458	-0.0003	0.0343	0.0342	0.0693	0.9515	0.0002	0.0173	0.0171	0.0998	0.9473
1,0.8,0.8	0.0106	-0.0007	0.0540	0.0533	0.0788	0.9468	-0.0003	0.0339	0.0337	0.0913	0.9522	0.0002	0.0170	0.0169	0.1580	0.9477
0.9,1,1	0.0122	-0.0004	0.0549	0.0542	0.0827	0.9468	-0.0003	0.0347	0.0343	0.1000	0.9502	0.0003	0.0171	0.0172	0.1758	0.9488
0.75,1.1,1.2	0.0251	-0.0004	0.0545	0.0546	0.1198	0.9475	-0.0001	0.0348	0.0346	0.1854	0.9478	0.0001	0.0171	0.0174	0.4208	0.9520
0.75,1,1.2	0.0283	-0.0004	0.0548	0.0543	0.1352	0.9453	-0.0001	0.0348	0.0345	0.2107	0.9465	0.0001	0.0171	0.0173	0.5029	0.9501
0.75,0.9,1.2	0.0315	-0.0004	0.0548	0.0541	0.1485	0.9430	-0.0002	0.0345	0.0343	0.2365	0.9481	0.0001	0.0169	0.0172	0.5764	0.9490
0.75,1,1	0.0320	-0.0004	0.0543	0.0538	0.1503	0.9454	-0.0002	0.0345	0.0341	0.2474	0.9474	0.0001	0.0169	0.0171	0.5951	0.9502
0.75,1,0.8	0.0360	-0.0005	0.0537	0.0533	0.1708	0.9458	-0.0002	0.0341	0.0338	0.2851	0.9474	0.0001	0.0167	0.0169	0.6896	0.9498
0.75,0.9,0.9	0.0372	-0.0005	0.0540	0.0533	0.1773	0.9443	-0.0002	0.0339	0.0338	0.2955	0.9491	0.0001	0.0167	0.0169	0.7148	0.9496
0.75,0.8,1	0.0385	-0.0005	0.0543	0.0534	0.1840	0.9445	-0.0003	0.0338	0.0338	0.3029	0.9509	0.0001	0.0167	0.0169	0.7422	0.9501

Table 5.3: Empirical properties of the utility-based estimator $\widehat{\beta}(\tau)$ and associated tests of $H_0 : \beta(\tau) \leq 0$ vs. $H_1 : \beta(\tau) > 0$ when the illness-death process is mixed-Markov with conditional intensities $\lambda_{kl}(t|X,U) = U\lambda_{kl0}\exp(\gamma_{kl}X)$ for $\phi = 0.5$; $\tau = 1, \pi_T = 0.6, \pi_T^1 = 0.9, \pi_D = 0.2, \pi_r = P(C_r \leq \tau|X = 0) = 0.2$ and $0.4, \exp(\gamma_r) = 1, u_0 = 1, u_1 = 0.5, u_2 = 0, n_{sim} = 10000$ simulation runs.

		MIXED-MARKOV PROCESSES														
		$n = 100$					$n = 250$					$n = 1000$				
$[e^{\gamma_{01}}, e^{\gamma_{02}}, e^{\gamma_{12}}]$	$\beta(\tau)$	BIAS	ESE	ASE	RR	CP	BIAS	ESE	ASE	RR	CP	BIAS	ESE	ASE	RR	CP
$\phi = 1.0: \ \pi_r = 0.20$ and $\exp(\gamma_r) = 1$																
1,1,1	0.0000	-0.0007	0.0552	0.0543	0.0517	0.9483	0.0004	0.0348	0.0345	0.0502	0.9498	0.0002	0.0173	0.0173	0.0521	0.9479
1,1,0.8	0.0051	-0.0007	0.0546	0.0537	0.0618	0.9484	0.0005	0.0344	0.0341	0.0676	0.9506	0.0003	0.0171	0.0171	0.0938	0.9476
1,0.8,1	0.0056	-0.0007	0.0548	0.0540	0.0680	0.9438	0.0005	0.0344	0.0342	0.0686	0.9515	0.0002	0.0171	0.0171	0.0950	0.9485
1,0.8,0.8	0.0107	-0.0007	0.0541	0.0533	0.0797	0.9453	0.0005	0.0340	0.0338	0.0929	0.9502	0.0002	0.0169	0.0169	0.1582	0.9483
0.9,1,1	0.0108	0.0008	0.0547	0.0542	0.0806	0.9452	0.0001	0.0345	0.0345	0.0946	0.9486	0.0001	0.0172	0.0173	0.1573	0.9508
0.75,1.1,1.2	0.0216	0.0004	0.0548	0.0547	0.1096	0.9500	0.0001	0.0349	0.0348	0.1559	0.9473	0.0002	0.0174	0.0174	0.3477	0.9477
0.75,1,1.2	0.0248	0.0002	0.0555	0.0546	0.1285	0.9429	0.0004	0.0350	0.0346	0.1889	0.9447	0.0003	0.0173	0.0174	0.4215	0.9468
0.75,0.9,1.2	0.0280	0.0000	0.0555	0.0543	0.1367	0.9446	0.0005	0.0349	0.0345	0.2129	0.9441	0.0003	0.0173	0.0173	0.4949	0.9462
0.75,1,1	0.0287	0.0002	0.0550	0.0541	0.1435	0.9434	0.0004	0.0347	0.0343	0.2175	0.9448	0.0003	0.0171	0.0172	0.5195	0.9468
0.75,1,0.8	0.0330	0.0002	0.0544	0.0535	0.1599	0.9430	0.0005	0.0344	0.0340	0.2591	0.9459	0.0004	0.0170	0.0170	0.6292	0.9474
0.75,0.9,0.9	0.0341	0.0000	0.0547	0.0535	0.1653	0.9456	0.0006	0.0344	0.0340	0.2724	0.9452	0.0004	0.0171	0.0170	0.6476	0.9467
0.75,0.8,1	0.0352	0.0001	0.0544	0.0536	0.1714	0.9448	0.0007	0.0346	0.0340	0.2821	0.9440	0.0004	0.0170	0.0170	0.6757	0.9469
$\phi = 1.0: \ \pi_r = 0.40$ and $\exp(\gamma_r) = 1$																
1,1,1	0.0000	0.0002	0.0571	0.0562	0.0552	0.9448	-0.0007	0.0358	0.0356	0.0494	0.9506	0.0001	0.0179	0.0178	0.0509	0.9491
1,1,0.8	0.0051	0.0001	0.0565	0.0555	0.0641	0.9457	-0.0007	0.0354	0.0352	0.0666	0.9497	0.0002	0.0177	0.0176	0.0915	0.9477
1,0.8,1	0.0056	0.0000	0.0566	0.0558	0.0629	0.9484	-0.0006	0.0356	0.0353	0.0693	0.9497	0.0000	0.0178	0.0177	0.0980	0.9483
1,0.8,0.8	0.0107	-0.0001	0.0559	0.0551	0.0755	0.9484	-0.0005	0.0352	0.0349	0.0912	0.9490	0.0001	0.0176	0.0175	0.1568	0.9474
0.9,1,1	0.0108	0.0000	0.0579	0.0560	0.0816	0.9429	-0.0004	0.0360	0.0356	0.0906	0.9502	0.0004	0.0179	0.0178	0.1587	0.9473
0.75,1.1,1.2	0.0216	0.0003	0.0581	0.0566	0.1144	0.9412	-0.0005	0.0362	0.0360	0.1500	0.9474	0.0003	0.0180	0.0180	0.3361	0.9484
0.75,1,1.2	0.0248	0.0003	0.0573	0.0563	0.1201	0.9454	-0.0004	0.0359	0.0358	0.1737	0.9495	0.0004	0.0179	0.0179	0.4031	0.9471
0.75,0.9,1.2	0.0280	0.0003	0.0571	0.0561	0.1343	0.9446	-0.0004	0.0358	0.0356	0.1955	0.9502	0.0004	0.0178	0.0178	0.4778	0.9465
0.75,1,1	0.0287	0.0003	0.0568	0.0558	0.1361	0.9441	-0.0004	0.0356	0.0355	0.2044	0.9501	0.0004	0.0178	0.0178	0.4983	0.9468
0.75,1,0.8	0.0330	0.0003	0.0562	0.0552	0.1541	0.9444	-0.0004	0.0353	0.0351	0.2464	0.9502	0.0005	0.0176	0.0176	0.6018	0.9469
0.75,0.9,0.9	0.0341	0.0003	0.0562	0.0553	0.1627	0.9447	-0.0004	0.0353	0.0351	0.2545	0.9501	0.0005	0.0175	0.0176	0.6248	0.9458
0.75,0.8,1	0.0352	0.0003	0.0565	0.0553	0.1690	0.9448	-0.0004	0.0353	0.0351	0.2616	0.9486	0.0005	0.0175	0.0176	0.6444	0.9446

Table 5.4: Empirical properties of the utility-based estimator $\widehat{\beta}(\tau)$ and associated tests of $H_0 : \beta(\tau) \leq 0$ vs. $H_1 : \beta(\tau) > 0$ when the illness-death process is mixed-Markov with conditional intensities $\lambda_{kl}(t|X, U) = U\lambda_{kl0}\exp(\gamma_{kl}X)$ for $\phi = 1.0$; $\tau = 1, \pi_T = 0.6, \pi_T^1 = 0.9, \pi_D = 0.2, \pi_r = P(C_r \leq \tau|X = 0) = 0.2$ and $0.4, \exp(\gamma_r) = 1, u_0 = 1, u_1 = 0.5, u_2 = 0, n_{sim} = 10000$ simulation runs.

Simulation results for the mixed-Markov processes are summarized in Table 5.3 for $\phi = 0.5$ and in Table 5.4 for $\phi = 1$, respectively. Similar results can be found here; the IJK procedure performs well in most settings. The average standard error is smaller than the empirical standard error in settings where $n = 100$ and $\phi = 1$; see Table 5.4.

The simulation studies confirm that the IJK procedure provides valid inference in the utility-based framework (1.2.3.2) for moderate- to large-sized clinical trials. Additional results for scenarios where $\pi_T = 0.6, \pi_T^1 = 0.7$ and $\pi_D = 0.4$ are given in Tables D.3-D.6 of Appendix D.5.

5.5 Bleeding in chemotherapy-induced thrombocytopenia

Here we consider analysis of data from the multi-center randomized PREPAReS trial in thrombocytopenic patients with hemato-oncology diseases discussed in Section 1.3.5 (Ypma et al., 2016; van der Meer et al., 2018). This non-inferiority trial was conducted in the Netherlands, Norway and Canada to assess whether the Mirasol pathogen-reduced transfusion strategy (arm B: PR-plasma PC) is not worse than the transfusion strategy involving standard plasma stored platelets (arm A: plasma PC), measured on the proportion of WHO grade ≥ 2 bleeding complications occurring over $[0, 42]$. A utility-based analysis is appealing in this context as it not only incorporates the timing of bleeding complications and occurrences of other relevant outcomes (e.g. platelet recovery, death), but also accounts for the differing bleeding severities. We reanalyze the PREPAReS trial based on the four-state model of Figure 1.4, and evaluate non-inferiority of the pathogen-reduced transfusion strategy in terms of the following measure of burden of bleeding:

$$D(t) = \sum_{k=0}^{3} u_k \int_0^t P_{0k}(0, s)ds, \qquad 0 \leq t \leq 42 \ ,$$

with $u_k \in [0, 1]$ the utility score assigned to state k and $D(t)$ taking values on $(0, t]$, $k = 0, 1, 2, 3$. Recall from Section 1.3.5 that the four states are 0: no bleeds or WHO grade 1 bleeds, 1: WHO grade 2 bleeds, 2: WHO grade 3 or 4 bleeds, and 3: death. Utilities are defined such that higher values of u_k correspond to less (bleeding) burden and, consequently, higher values of $D(t)$ to higher efficacy of the transfusion type. Naturally, by the definition of the 4-point WHO bleeding scale, we assume $1 = u_0 > u_1 > u_2 > u_3 = 0$ in which case $D(t) \in (0, t)$. Note that $D(t)$ takes

its maximum value t only if an individual stays in state 0 on all days over $(0, t]$, i.e. $D(t) = t$ for $u_0 = 1$. We consider a range of utility scores for illustration.

Because pathogen inactivation reduces the risk of bacteria and transfusion-transmitted infections, some loss of efficacy on bleeding in arm B is tolerated. If $X = 1$ stands for arm B and $X = 0$ for arm A, we aim to test for non-inferiority using the hypotheses

$$\begin{aligned} H_0 &: D(\tau|X=0) - D(\tau|X=1) \geq \delta \qquad \text{(B is inferior to A by at least } \delta\text{)} \\ H_1 &: D(\tau|X=0) - D(\tau|X=1) < \delta \qquad \text{(B is inferior to A by less than } \delta \text{ or superior to A)} , \end{aligned}$$

or equivalently expressed as

$$\begin{aligned} H_0 &: D(\tau|X=1) - D(\tau|X=0) \leq -\delta \\ H_1 &: D(\tau|X=1) - D(\tau|X=0) > -\delta , \end{aligned}$$

where δ is a pre-determined non-inferiority margin and $\tau > 0$ a pre-determined time point; we set $\tau = 42$. The null hypothesis H_0 is rejected if the lower limit of the one-sided $(1-\omega)\%$ confidence interval (CI) for $\beta(\tau) = D(\tau|X=1) - D(\tau|X=0)$, given by

$$\left[\widehat{\beta}(\tau) - z_{1-\omega} \sqrt{\widehat{\text{var}}\big(\widehat{\beta}(\tau)\big)},\ 42 \right] ,$$

is greater than $-\delta$. The Food and Drug Administration (FDA) often allows for a 20% relative effect to define margins in non-inferiority trials (FDA, 2016). We follow their guidelines and use a 20% relative decrease in the mean cumulative utility $D(t)$ to define δ. If $D(42|X=0) = 40$ is the anticipated cumulative mean utility score at $\tau = 42$ days in arm A, the non-inferiority margin can be taken as $\delta = 8 = 40\text{-}32$ (i.e. $32 = 0.8 \cdot 40$); hence, a beneficial effect of the pathogen-reduced transfusion strategy on the risk of infections is represented by $\beta(\tau) \geq -8$.

The four-state model of Figure 1.4 was fitted to the PREPAReS data collected from the Canadian study sites only. The Canadian subtrial is comprised of 127 individuals who were to be followed for 42 days, providing justification for the choice of $\tau = 42$. For the analyses here we excluded 16 individuals with grade ≥ 2 bleeding at the time of randomized treatment assignment (c.f. Ypma et al., 2016, Figure 3), leaving $n = 52$ individuals in arm A and $n = 59$ in arm B respectively. Patients who recovered from thrombocytopenia before the administrative end of study were assumed to occupy state 0 (i.e. no bleeding complications) for the remaining follow-up. Patients

u_2	u_1	A: plasma PC		B: PR-plasma PC				
		$\widehat{D}(\tau)$	SE$^\star\big(\widehat{D}(\tau)\big)$	$\widehat{D}(\tau)$	SE$^\star\big(\widehat{D}(\tau)\big)$	$\widehat{\beta}(\tau)$	SE$^\star\big(\widehat{\beta}(\tau)\big)$	95% CI
0.1	0.9	39.787	0.993	39.189	1.236	-0.598	1.586	[-3.206,42]
	0.7	39.265	1.002	38.696	1.228	-0.569	1.585	[-3.177,42]
	0.5	38.743	1.023	38.203	1.227	-0.540	1.598	[-3.168,42]
	0.2	37.959	1.072	37.463	1.239	-0.496	1.638	[-3.191,42]
0.3	0.9	39.802	0.986	39.315	1.190	-0.487	1.545	[-3.029,42]
	0.7	39.280	0.996	38.822	1.182	-0.458	1.545	[-3.000,42]
	0.5	38.758	1.016	38.329	1.181	-0.429	1.558	[-2.992,42]
	0.4	38.766	1.013	38.392	1.162	-0.374	1.541	[-2.909,42]

Table 5.5: Nonparametric Aalen-Johansen-based estimates of $D(\tau|X=0), D(\tau|X=1)$ and $\beta(\tau) = D(\tau|X=1) - D(\tau|X=0)$ for different utility scores $u_1 \in [0.9, 0.2]$ and $u_2 = 0.1, 0.3$, along with SEs obtained by the infinitesimal jackknife* procedure (Therneau, 2024); $\tau = 42, 1 > u_1 > u_2 > 0, u_0 = 1, u_3 = 0$.

who withdraw prematurely from the study for any reason, such as non-compliance or adverse transfusion reactions, were censored; a thorough discussion of intercurrent events is beyond the present scope of this analysis.

Figure 5.8 contains plots of the estimated cumulative mean utility $\widehat{D}(t)$ in each transfusion group for different utility values u_1 and u_2. It is apparent that, for a given utility u_2, the mean cumulative utility decreases as u_1 decreases. Overall, there is no evidence of a difference in the cumulative mean utility from use of pathogen-reduced platelets compared to standard platelets for any combination of u_1 and u_2. By setting $\tau = 42$ and using the utilities $u_0 = 1, u_1 = 0.7, u_2 = 0.3$ and $u_3 = 0$, we obtain $\widehat{D}(42|X=1) = 38.822$ (IJK SE=1.182) and $\widehat{D}(42|X=0) = 39.280$ (IJK SE=0.996). This in turn yields $\widehat{\beta}(42) = -0.458$ (SE=1.545, 95% CI: $[-3, 42]$); the infinitesimal jackknife standard error for $\widehat{\beta}(42)$ was calculated based on the `R` output provided by `survfit(..., influence = TRUE)` (Therneau, 2024); see also Appendix D.4. Since $-\delta < -3$ for $\delta = 8$, the null hypothesis H_0 is rejected at the $\omega = 5\%$ significance level and so non-inferiority of the pathogen inactivated platelets compared to standard platelets can be claimed. The left panel of Figure 5.9 depicts the treatment effect $\widehat{\beta}(t)$ over time, the lower 95% pointwise confidence band, and the non-inferiority margin. Results for alternative choices of u_1 and u_2 are summarized in Table 5.5.

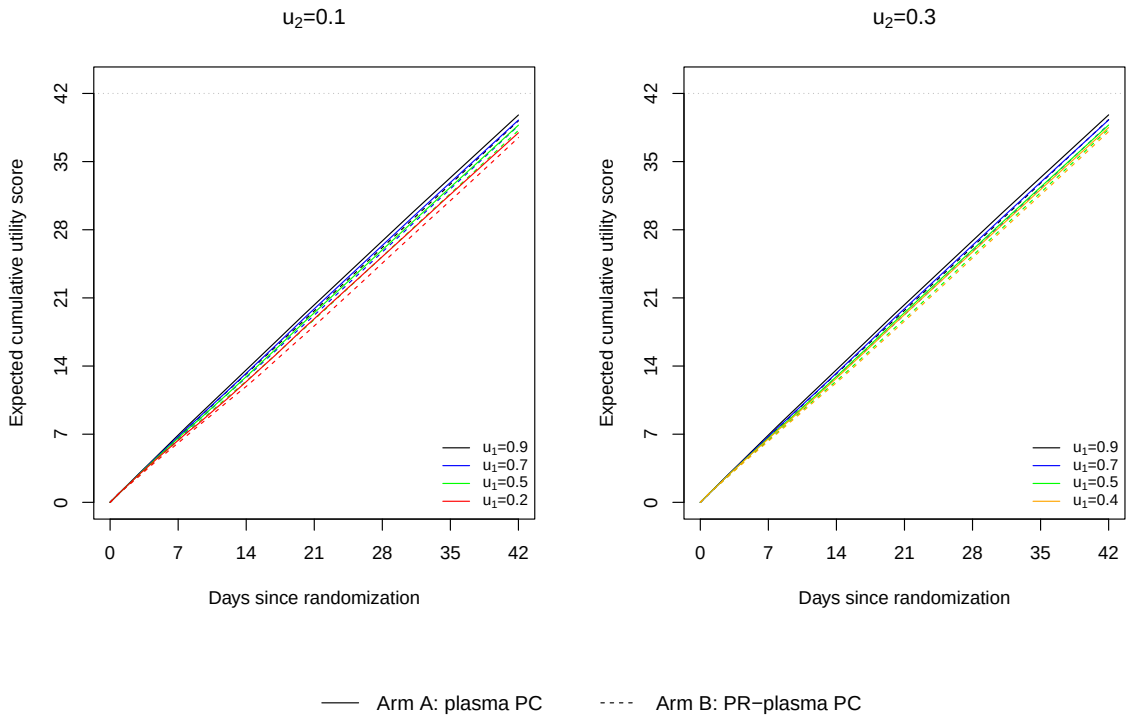


Figure 5.8: Nonparametric estimates of the expected cumulative utility score $D(t)$ for different utility scores $u_1 \in [0.9, 0.2]$ and $u_2 = 0.1$ (left panel), 0.3 (right panel); $1 > u_1 > u_2 > 0$, $u_0 = 1, u_3 = 0$.

The marginal utility-based estimand $\beta(\tau)$ synthesizes evidence across multiple bleeding outcomes and death. We next provide secondary analyses to enhance a deeper understanding of $\widehat{\beta}(42)$, as recommended per Chapter 2 for any marginal treatment effect. Nonparametric Aalen-Johansen estimates of the state occupancy probabilities $P_{0k}(0, t|X)$ serve as the basis for calculating $\widehat{D}(t|X)$ and are illustrated in Figure 5.10. Figure 5.11 depicts the nonparametric Nelson-Aalen estimates of the cumulative transition intensities $\lambda_{kl}(t|X), k \neq l$. Nelson-Aalen as well as Aalen-Johansen estimation is implemented in standard `R` packages such as `survival`, `mvna` and `etm` (Therneau, 2024; Allignol et al., 2008, 2011); we used the former package here.

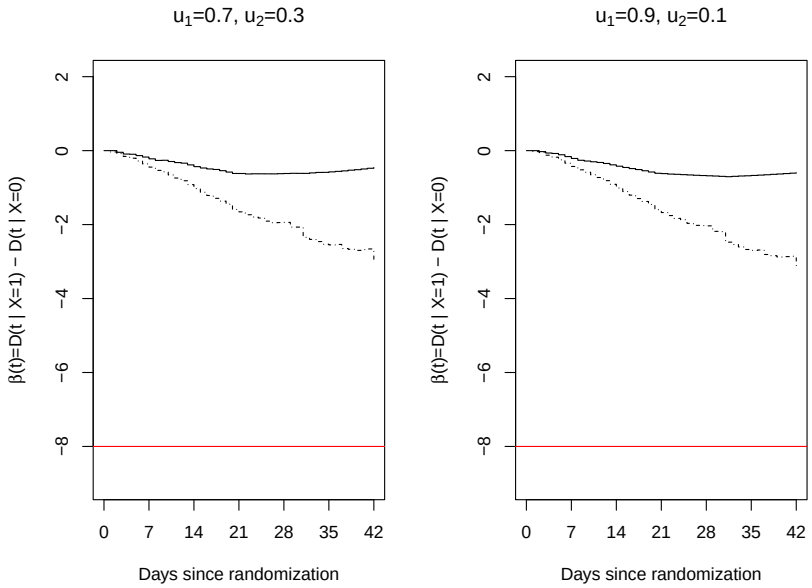


Figure 5.9: Plots of $\widehat{\beta}(t) = \widehat{D}(t|X = 1) - \widehat{D}(t|X = 0)$ for $u_1 = 0.7, u_2 = 0.3$ (left panel) and $u_1 = 0.9, u_2 = 0.1$ (right panel); the dashed black line represents the lower 95% pointwise confidence band of $\widehat{\beta}(t)$, the red line represents the non-inferiority margin of -8.

Patients in both arms are at the highest risk of transitioning from state 0 to state 1 within the first ten days of follow-up; see black curves in the left panel of Figure 5.11 and Figure D.2 in Appendix D.6. The latter figure displays the Nelson-Aalen estimates for transitions out of state 0 for a finer y-axis scale. This explains to some extent both the initial decrease in the proportion of patients with no or relatively minor bleeding complications (i.e. in state 0) over $[0, 10]$ and the increasing trend in the proportion of patients with grade 2 bleeds (i.e. in state 1) over that time period; see black and red curves in the upper panels of Figure 5.10. According to the middle panel of Figure 5.11, patients in both arms are at a fairly high and constant risk of transitioning from state 1 to state 0. Beyond 10 days there is thus a clear trend of an increasing proportion of patients in state 0 and a decreasing proportion of patients in state 1, with estimates of $\widehat{P}_{00}(0, 42|X = 0) = 0.901, \widehat{P}_{01}(0, 42|X = 0) = 0$ and $\widehat{P}_{00}(0, 42|X = 1) = 0.929, \widehat{P}_{01}(0, 42|X = 0) = 0$ respectively. Only a few patients in each group experienced grade 3 or 4 bleeding; transitions from grade 2 to grade 3 or 4 bleeding are most likely, as one would expect. Approximately 10% of patients in

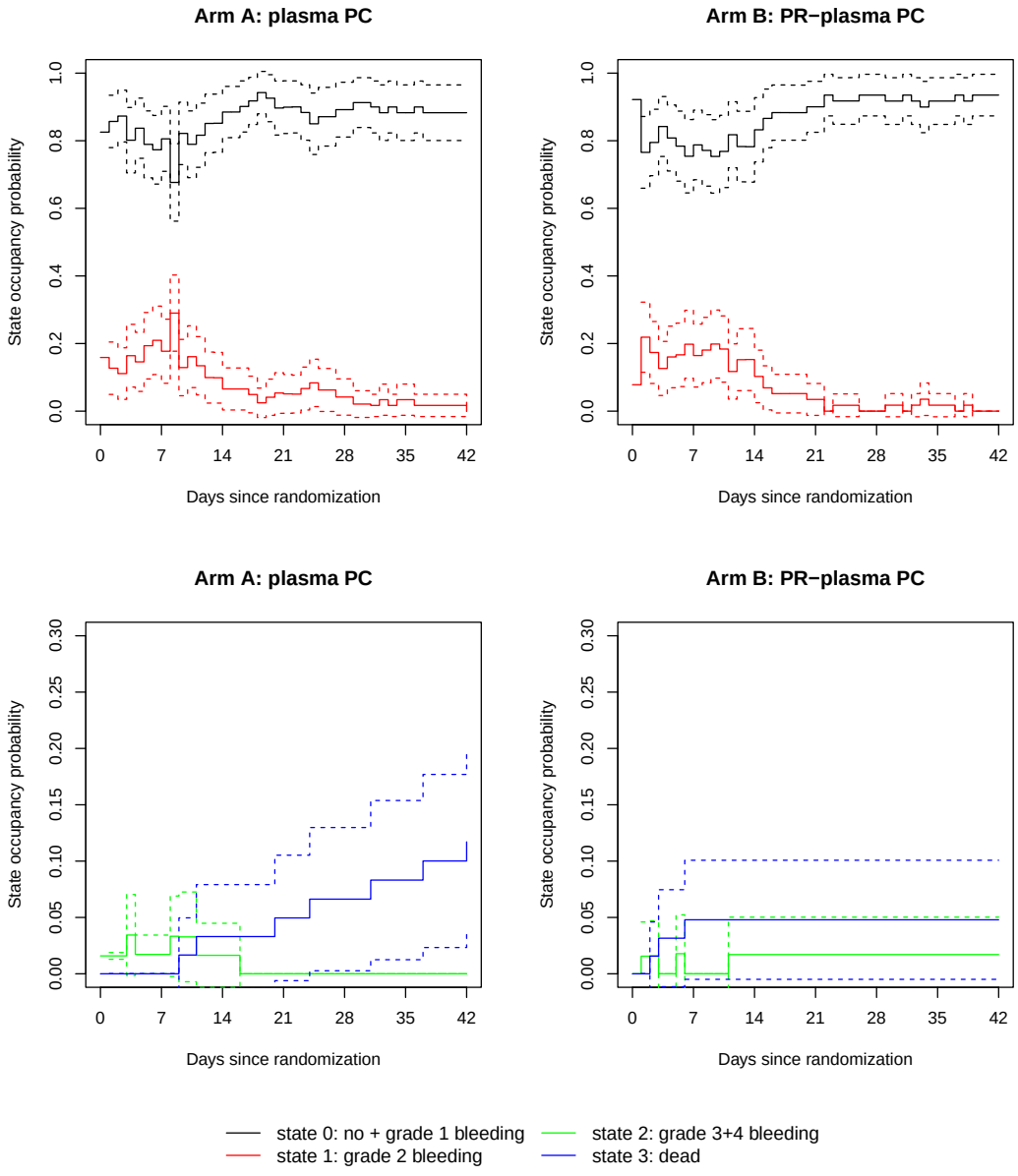


Figure 5.10: Nonparametric Aalen-Johansen estimates of state occupancy probabilities $P_{0k}(0,t)$ and pointwise 95% confidence intervals by transfusion type; top row: states 0 and 1 occupancy probabilities, bottom row: states 2 and 3 occupancy probabilities; CIs were calculated based on the IJK standard errors for $\widehat{P}_{0k}(0,t)$.

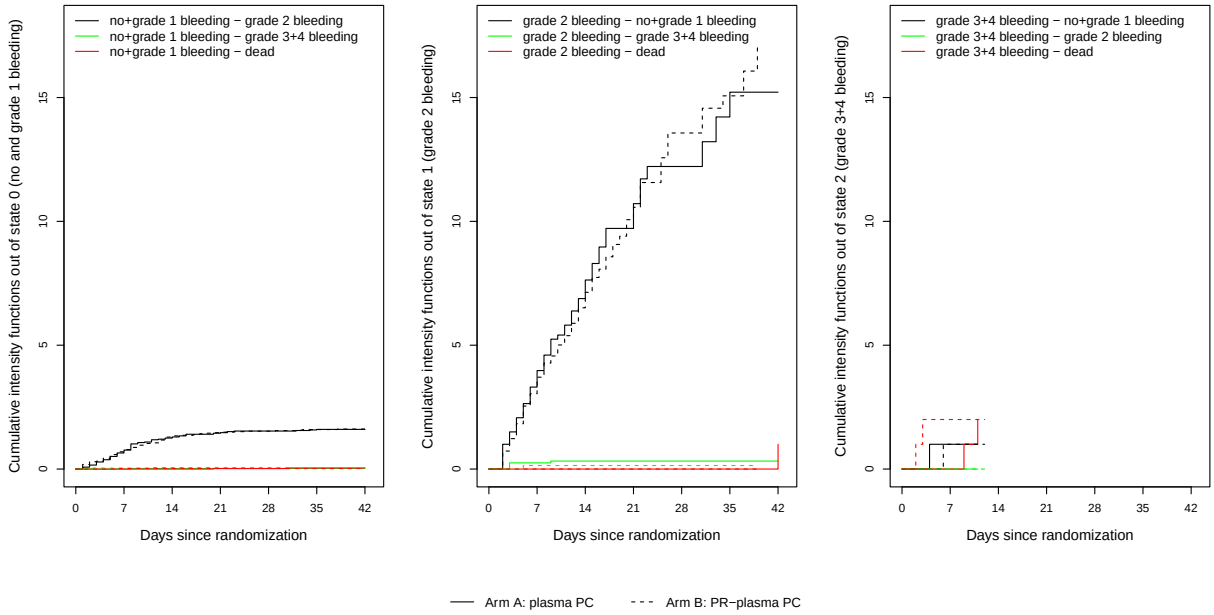


Figure 5.11: Nonparametric Nelson-Aalen estimates of the cumulative intensity functions for transitions out of states 0 (left panel), 1 (middle panel) and 2 (right panel).

group A died over the 42-day period, as compared to 5% in group B; see blue curves in the lower panels of Figure 5.10. Patients in group B, however, tended to die earlier than patients in group A. These findings are also reflected in Figure 5.12 and Table 5.6.

Figure 5.12 contains stacked probability plots of the nonparametric Aalen-Johansen estimates of $P_{0k}(0,t|X)$, along with estimates of associated restricted mean times summarized in Table 5.6. We briefly explain the nature of Figure 5.12: the dotted, dashed-dotted and dashed lines represent $\widehat{P}_{02}(0,t|X)$, $\sum_{k=1}^{2}\widehat{P}_{0k}(0,t|X)$ and $\sum_{k=0}^{2}\widehat{P}_{0k}(0,t|X)$, respectively; the red, yellow, purple and green areas represent $\tau-\int_0^\tau\sum_{k=0}^{2}\widehat{P}_{0k}(0,t|X)dt$, $\int_0^\tau\widehat{P}_{00}(0,t|X)dt$, $\int_0^\tau\widehat{P}_{01}(0,t|X)dt$ and $\int_0^\tau\widehat{P}_{02}(0,t|X)dt$, respectively.

For both transfusion strategies, patients are free of bleeding or have minor bleeding complications for most of the 42 days (i.e. are in state 0), which explains why $\widehat{D}(42|X=1)$ and $\widehat{D}(42|X=0)$ are close to 42 days when u_0 is set to 1. It is obvious that both $\widehat{D}(42|X=1)$ and $\widehat{D}(42|X=0)$ are closest to 42 when u_1 and u_2 are set to the highest utility values considered, i.e. $u_1=0.9$ and

$u_2 = 0.3$; see 5th row in Table 5.5. Patients in group B spend less time in states 0 (no or grade 1 bleeding) and 1 (grade 2 bleeding), and more time in states 2 (grade 3+4 bleeding) and 3 (death). The differences in restricted mean times between transfusion types are however rather minor, with discrepancies in estimates related to states 1 and 2 (i.e. $\int_0^\tau \widehat{P}_{0k}(0, t|X)dt, k = 1, 2$) most relevant for the primary utility-based analysis. It can be seen from Table 5.5 that the estimated treatment effect $\widehat{\beta}(42)$ appears greatest when the utilities are set to $u_1 = 0.9$ and $u_2 = 0.1$, and smallest when $u_1 = 0.4$ and $u_2 = 0.3$.

5.6 Discussion

The utility-based framework of Sections 1.2.3.2, 5.3.3 and 5.4 has great appeal in type (I) clinical trial settings involving multiple types of events. Specification and incorporation of utilities into the evaluation of treatments both enables the handling of complex disease processes and makes explicit the relative importance to an individual of specific events and time spent in specific states. As mentioned in Section 2.2.4, utilities are often adopted implicitly. For example, progression-free survival and other composite time-to-event responses equally weight the component outcomes in terms of importance, and relative utilities of different sample paths are implied when computing win ratios (Oakes, 2016). We emphasized that no method is free of problems and argue that specification of utilities is analogous to the need to rank sample paths or disease states. While some may view utilities as subjective and therefore undesirable for assessing the effects of experimental treatments in clinical trials, we note that they are used to great effect in quality of life and health economic analyses (Wenger et al., 1984; Torrance, 1986; Glasziou et al., 1990; Revicki and Ehreth, 1997; Revicki et al., 2006; Wong et al., 2023b,a).

There is a long history of methods for the study and specification of health utilities (Llewellyn-Thomas et al., 1984). Direct methods for eliciting utility values include the basic rating scale system, standard gamble (SG) and time trade-off (TTO) (Torrance, 1986, 1987; Feeny and Torrance, 1989; Bennett and Torrance, 1996). The *basic rating scale* procedure consists of a line, with one end representing full health $(= 0)$ and the other one representing death $(= 1)$. Individuals are asked to place the remaining disease states on the line such that the intervals between the placements reflect the differences in their preferences. In the *standard gamble* method, individuals

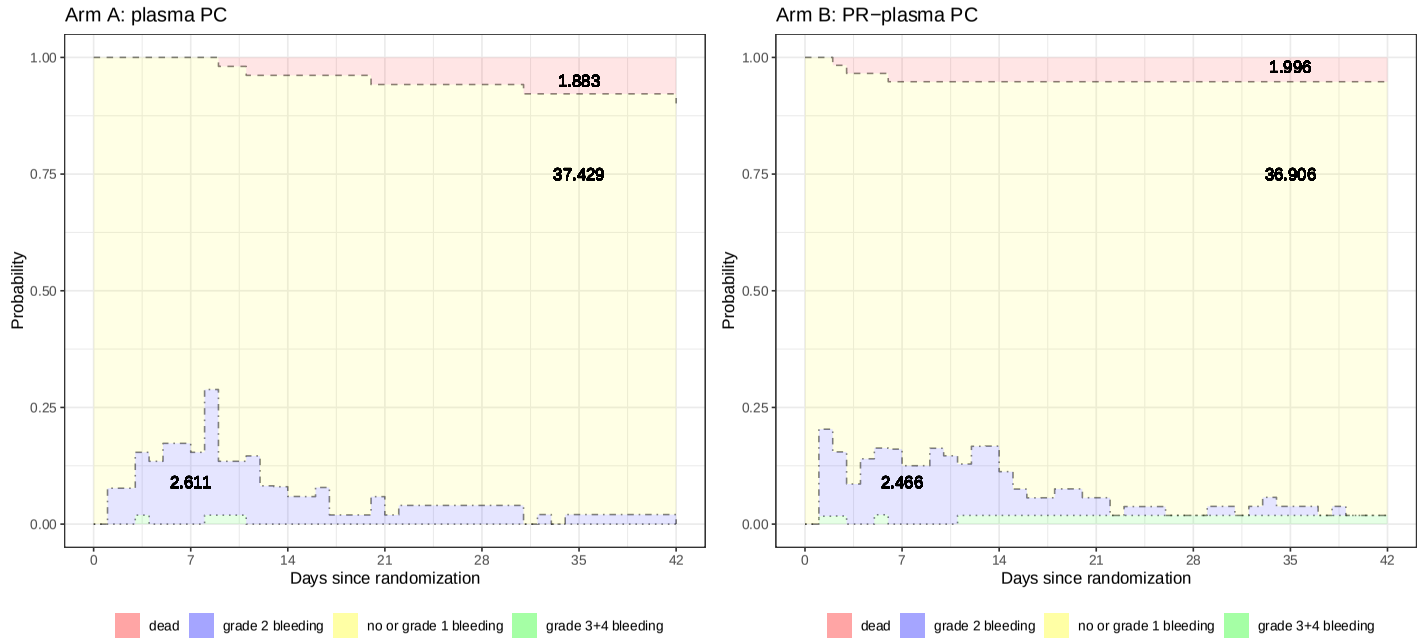


Figure 5.12: Stacked plots of the nonparametric Aalen-Johansen estimates of the state occupancy probabilities for no or grade 1 bleeding (state 0), grade 2 bleeding (state 1), grade 3+4 bleeding (state 2) and death (state 3) by transfusion type.

Transfusion type	State 0 (no bleeding, grade 1 bleeding)		State 1 (grade 2 bleeding)		State 2 (grade 3+4 bleeding)		State 3 (death)	
	MEAN	SE*	MEAN	SE*	MEAN	SE*	MEAN	SE*
A: plasma PC	37.429	1.119	2.611	0.526	0.077	0.046	1.883	0.962
B: PR-plasma PC	36.906	1.281	2.466	0.462	0.632	0.592	1.996	1.122

Table 5.6: Nonparametric estimates of the restricted mean times spent in states 0, 1, 2 and 3 over $[0, 42]$ by transfusion type, see also Figure 5.12 for illustration; *SEs are obtained by infinitesimal jackknife (Therneau, 2024).

are given two alternatives. Alternative 1 is a treatment with two possible outcomes: full health with probability p and death with probability $1 - p$. Alternative 2 represents an outcome that is less favorable than full health but more favorable than death. The probability p is varied until the individual is indifferent between the two alternatives, at which point the utility value for the intermediate disease outcome is p. The *time trade-off* method is similar to the standard gamble approach; here an individual is also asked to choose between two alternatives. Alternative 1 represents living with the current disease condition for the next 10 years, while alternative 2 offers a shorter but healthier life. Individuals are asked how many years in their current disease state they would be willing to "trade-off" in order to regain full health. Selection of such a trade-off clearly depends on an individual's perspective and socio-demographic characteristics. A concern is that individuals may have difficulties in understanding the concept of the task, particularly those in vulnerable populations such as the elderly and children (Xie et al., 2020). With these direct methods, it is also debatable whether utilities should be estimated by patients, clinicians (e.g. nurses, physicians), or unaffected individuals of the general population (Versteegh and Brouwer, 2016; Karimi et al., 2017).
Indirect methods include multi-attribute utility systems such as the EuroQol (EQ)-5D questionnaire (Kind, 1996), the short form 6D (SF-6D) questionnaire (Brazier et al., 2002) and the health utility index (HUI) (Torrance et al., 1995; Feeny et al., 1995; Horsman et al., 2003).

We considered utility values bounded by 0 and 1. Negative utilities have seen increasing use in recent years and correspond to states that are considered worse than death. Taylor et al. (2017) discussed the potential benefit of expressing utilities in terms of changes in health status. They argue in particular that a given state of health could be valued differently depending on what state of health an individual starts in.
Overall, utility scoring systems remain an active area of research (Torrance et al., 2002; Bleichrodt, 2002; Parkin and Devlin, 2006; Nord et al., 2009; Richardson et al., 2015; Taylor et al., 2017; Rodríguez-Míguez et al., 2019; Xie et al., 2020; Dewitt and Torrance, 2020; Ma et al., 2024; Hirniak et al., 2024). It would be desirable for any specified utility scores in a given disease area to be universally accepted. Regardless of how the utilities are determined, however, sensitivity analyses can be conducted to assess how much particular utility values may have influenced the study findings; see for instance Glasziou et al. (1990) and Revicki et al. (2006). Threshold

analyses where utility values for certain health states are varied between 0 and 1 are most common. For the utility-based analysis of Section 5.5, the thresholds are the values (u_1, u_2) for which the effect $\widehat{\beta}(\tau)$ is 0.

We note that while we focused on progressive disease processes for most of this chapter, the utility-based approach can be easily adopted to reversible processes; see for instance the disease process in the platelet transfusion trial of Section 5.5. Further study of utility-based estimands in the setting of state- and time-dependent utility scores is needed, including variance estimation. For the design of clinical trials we recommend a simulation-based approach to estimate power for tests based on (5.4.17), with sample size chosen to ensure adequate power. Historical data can be used along with model assumptions to estimate parameter values to guide such simulations. Murray and Cole (2000) discussed sample size calculations for the Q-TWiST measure. Use of adaptive sample size calculations based on internal pilot studies is also an area of future research (Mütze and Friede, 2021).

Chapter 6

Conclusions and future work

6.1 Overview and contributions of thesis

The randomized controlled trial is considered the most rigorous design for assessing causal effects of experimental treatments in clinical research. Despite the benefits of randomization and the careful conduct of such trials, specification of estimands for the assessment of treatment effects is not straightforward when the disease processes under study are "complex". This is particularly true for cancer (Hortobagyi et al., 1996; Rittmeyer et al., 2017) and cardiovascular (CV) trials (O'Connor et al., 2009; Marso et al., 2016; Solomon et al., 2019) where disease processes typically involve multiple types of potentially recurring events and competing terminal events. The occurrence of intercurrent events that affect or even preclude observation of a targeted primary outcome are a further complication for causal interpretation of treatment effects (ICH E9 Steering Committee, 2019).

Chapter 2 was motivated by the need to provide guidance on how best to design, analyze and interpret randomized clinical trials for complex disease processes. We emphasized the value of a comprehensive multistate model for conceptualizing the event generating process and providing a powerful framework for the analysis and interpretation of treatment effects. Using this multistate model for the disease process and occurrences of intercurrent events estimands may be defined, estimated and used for inference.

Randomized trials allow causal inferences about the effects of treatment. We discussed different views of causality (Aalen, 2012; Dawid, 2000; Dawid and Musio, 2022), and recommend inferences concerning marginal estimands that do not condition on any post-randomization events and are related to outcomes that are observable in the real world context of the trial. For processes involving intercurrent events this view employs analyses within the intention-to-treat (ITT) framework (Montori and Guyatt, 2001); in association with our recommendations, we proposed three guiding principles for defining estimands that are essential to any trial setting.
Generalized linear models (1.2.6) (GLM) for marginal features of a process allow for one-dimensional estimands, which can be used as a basis for trial design. One of the guiding principles emphasizes the importance of assessing the adequacy of such models and their assumptions to ensure validity of conclusions about a treatment. Marginal features are typically complicated functions of the intensities of the true process (see e.g. (3.2.1) and (4.2.2)), and so misspecification of models (1.2.6) may arise due to these features not being of GLM form. This was demonstrated in Chapters 3 and 4 in the context of trials focusing on non-fatal events. We considered in particular marginal estimands for type (II) trial settings where primary interest lies in the prevention of non-fatal events but mortality complicates the analysis and interpretation of data, and type (I) trial settings where interest lies primarily in survival but other relevant disease-related events may also be of interest.

Chapter 3 was concerned with estimands based on cumulative incidence function (CIF) regression in semi-competing risks processes. Important insights into the interpretability of estimands under model misspecification and robustness of tests were gained by studying the limiting values of estimators of treatment effect in the complementary loglog transformed CIF model (1.2.8) (Fine and Gray, 1999; Scheike et al., 2008) as functions of parameters of an intensity-based data-generating process. We recommend the use of models (1.2.8) in trials where the mortality rate is low and treatment has a fairly modest effect on mortality; in that case $\beta^\star$ is interpretable in the context of the full competing risks process and related hypothesis tests show satisfactory performance. Analyses based on models (1.2.8) alone do however not provide a sufficient basis for comprehensive assessments of treatment effects, and should be accompanied by supportive secondary analyses.

Chapter 4 was concerned with recurrent and terminal event processes where analyses based on non-fatal (or non-fatal and disease-related fatal) events create a competing risks problem. We

recommend the Ghosh and Lin (2002) model in type (II) trial settings only in addition to supportive secondary analyses of treatment effects to ensure a full understanding of response to treatment. The Mao and Lin (2016) model is not recommended for primary analyses of randomized trials due to its composite nature.

Chapter 5 was motivated by the need to provide guidance on strategies for evaluating treatment effects in trials (I) involving multiple events of varying disease severity. We critically discussed various methods based on utility functions and ranked outcomes (Pocock et al., 2012; Mao, 2023), and recommend causal inferences concerning marginal utility-based estimands (1.2.9). They provide a more nuanced assessment of treatment effects by assigning a utility score to each time unit spent in a certain disease state and accumulating this over the trial duration $[0, \tau]$ (Gelber et al., 1989; Cook et al., 2003). Such utility-based analyses yield treatment effect measures with a metric but require specification of utility scores. We stress however the need for "decisions" in each ranking or weighting approach; the win ratio makes assumptions about the ranking of individuals' sample paths (Pocock et al., 2012; Oakes, 2016), whereas the restricted mean time in favor of treatment (RMT-IF) relies on assumptions based on the favorability of event types (Mao, 2023). We developed an infinitesimal jackknife variance estimator for the expected cumulative utility score (1.2.4) to provide robust methods for causal inference of randomized trials.

In the following sections we briefly outline several topics for future research.

6.2 Future research on cumulative incidence function (CIF) regression

6.2.1 CIF regression and intermittent observation

Trials in advanced diseases often aim to study treatment effects on the cumulative probability of a particular non-fatal event, while accommodating the competing risk of death. As shown in Chapter 3, complementary log-log transformed models for the cumulative incidence of non-fatal events offer a powerful framework for addressing this. The estimation method of Fine and Gray (1999) requires the non-fatal event times to be subject only to right-censoring. Direct

binomial (DB) regression is an alternative estimation procedure and relies on binary indicators of whether the non-fatal event has been observed prior to times $a_m, m = 1, \ldots, M$ (Scheike et al., 2008); see Section 3.3.1.2. In the palliative trial of individuals with cancer metastatic to bone, however, patients are assessed intermittently at scheduled study visits for the development of new skeletal metastases (Hortobagyi et al., 1996). When such a non-fatal event process is subject to an intermittent observation scheme, the non-fatal event status is only known at the assessment times. This is further complicated by the fact that the non-fatal event status cannot be determined for individuals who were in state 0 at their last assessment before death or loss to follow-up; thus whether some individuals experienced the non-fatal event or not before death or censoring will be unknown for many individuals. This makes both procedures unusable for estimating β and $\alpha(t)$ in the cloglog model (3.2.3). The fatal event process (i.e. death) is typically subject to right-censoring, resulting in so-called dual observation schemes (Boruvka and Cook, 2016; Eaton and Zabor, 2022).

The development of cumulative incidence function regression methodology is needed to facilitate analysis of illness-death processes under dual observation schemes. We will next briefly outline the statistical challenges involved in this work. Consider a two-arm randomized clinical trial in which individuals experience illness-death processes (see Figure 1.1(b)) and interest lies in fitting models for $F_1(t|X)$. We let T_k denote the entry time to state $k = 1, 2$, and define the indicator $N_1(t) = \mathbb{1}(T_1 \leq t)$ to be used for binomial estimation.

Suppose the non-fatal event time T_1 is unknown but the status is observed at intermittent assessments at times $0 = a_0 < a_1 < \ldots < a_M = \tau$; often these will be pre-scheduled but they may also be random. We focus here on the former. If an individual is in state 0 at the last assessment a_M, $N_1(a_m)$ is known to be equal to 0 for all $m = 0, \ldots, M$. Individuals with $T_1 \in (a_{m-1}, a_m]$ and $a_m < T_2 < a_M$ (i.e. $0 \to 1 \to 2$ transition) have $N_1(a_{m'}) = 0$ for $m' \leq m - 1$ and $N_1(a_{m'}) = 1$ for $m' \geq m$. The same applies to individuals who were lost to follow-up at a random time $C_r < a_M$, following a non-fatal event in $(a_{m-1}, a_m]$. If an individual dies between two successive assessments (i.e. $T_2 \in (a_{m-1}, a_m]$) and was observed to occupy state 0 at a_{m-1} (i.e. $Z(a_{m-1}) = 0, N_1(a_{m-1}) = 0$), the indicator $N_1(a_m)$ is unknown; that is, we do not know if state 1 was entered or not. We will first assess the bias in DB estimation when $N_1(a_m)$ is set to 0 whenever $T_2 \in (a_{m-1}, a_m]$. This seems relevant since "simple" analysis strategies are typically wanted in clinical trials. The magnitude of the bias is expected to depend on the length of time

between successive assessments. Strategies for mitigating this bias include inverse weighting and a full model specification (Cook and Lawless, 2020). The former involves censoring individuals artificially at their last assessment a_{m-1} if $T_2 \in (a_{m-1}, a_m]$ and using inverse weighting techniques to adjust for this. The latter involves constructing the likelihood under all possible sample paths in the spirit of Cook and Lawless (2020), and requires models for $F_2(t|X)$ and $\lambda_{12}(t|X)$ in addition to $F_1(t|X)$.

6.2.2 Multinomial logistic CIF regression

Chapter 3 was concerned with illness-death processes where analyses based on the first event create a competing risks problem with two causes of failure. There are many competing risk settings with two causes of failure where interest lies in the effects of treatment on the cumulative probabilities of both failure types. Fyles et al. (2004), for instance, reported on a breast cancer trial designed to assess the effects of tamoxifen plus breast irradiation compared to tamoxifen alone on the prevention of local and distant relapses. Tu et al. (2009) considered a study of patients hospitalized with acute myocardial infarction, where interest lies in the simultaneous inference on the cumulative incidences of cardiovascular (CV) death and non-CV death. The standard implementation of the Fine and Gray (1999) approach involves fitting two separate models, one for each cumulative incidence function $F_k(t|X), k = 1, 2$. Such an analysis yields estimates $\widehat{F}_k(t|X)$ which do not satisfy the constraint $\sum_{k=1}^{2} \widehat{F}_k(t|X) \leq 1$ (Austin et al., 2021; Bonneville et al., 2024).

Gerds et al. (2012) proposed a multinomial logistic regression model for the specification of covariate effects on both cumulative incidence functions $F_1(t|X)$ and $F_2(t|X)$, as briefly discussed in Section 3.4.1. In the case of a binary treatment $X = 0, 1$, their model takes the form

$$\log\left(\frac{F_k(t|X)}{S(t|X)}\right) = A_k(t) + \beta_k X \ , \tag{6.2.1}$$

with the three implied expressions

$$F_1(t|X) = \frac{\exp(A_1(t) + \beta_1 X)}{1 + \exp(A_1(t) + \beta_1 X) + \exp(A_2(t) + \beta_2 X)}$$
$$F_2(t|X) = \frac{\exp(A_2(t) + \beta_2 X)}{1 + \exp(A_1(t) + \beta_1 X) + \exp(A_2(t) + \beta_2 X)}$$

$$S(t|X) = \frac{1}{1 + \exp(A_1(t) + \beta_1 X) + \exp(A_2(t) + \beta_2 X)}$$

summing up to 1 at any time t and for each $X = 0, 1$. The marginal estimands are of the form

$$\exp(\beta_k) = \frac{F_k(t|X=1)/S(t|X=1)}{F_k(t|X=0)/S(t|X=0)}\,, \qquad k = 1, 2\,.$$

The baseline function

$$A_k(t) = \log\left(\frac{F_k(t|X=0)}{S(t|X=0)}\right)$$

must fulfill the following properties:

1. $A_k(0) \downarrow -\infty$ so that $\exp(A_k(0)) = 0$ and $F_k(0|X) = 0, S(0|X) = 1$ for $X = 0, 1$
2. $a_k(t) = \dfrac{\partial A_k(t)}{\partial t} \geq 0$
3. $a_1(t) \geq \dfrac{a_2(t)(1 + \exp(A_2(t) + \beta_2 X))}{\exp(A_2(t) + \beta_2 X)}$ and $a_2(t) \geq \dfrac{a_1(t)(1 + \exp(A_1(t) + \beta_1 X))}{\exp(A_1(t) + \beta_1 X)}$.

The second property ensures that $A_k(t)$ is a monotonically increasing function; the third property follows from the fact that $f_k(t|X) = \partial F_k(t|X_i)/\partial t \geq 0$ and ensures monotonicity of $F_1(t|X)$ and $F_2(t|X)$. The likelihood function for "direct" inference on the cumulative incidence functions (Jeong and Fine, 2007) can be derived as

$$\mathcal{L} \propto \prod_{i=1}^{n} f_1\big(\widetilde{t}_i|x_i\big)^{\mathbb{1}(Z_i(\widetilde{t}_i)=1)} f_2\big(\widetilde{t}_i|x_i\big)^{\mathbb{1}(Z_i(\widetilde{t}_i)=2)} \big(1 - F_1\big(\widetilde{t}_i|X_i\big) - F_2\big(\widetilde{t}_i|X_i\big)\big)^{\mathbb{1}(Z_i(\widetilde{t}_i)=0)}\,,$$

where $T_i = \min(T_{1i}, T_{2i})$ is the exit time from state 0, C_i the censoring time and $\widetilde{T}_i = \min(T_i, C_i)$, $i = 1, 2, ..., n$.

A number of issues regarding model (6.2.1) have not been adequately addressed. A key issue is how to specify methods of estimation that accommodate property 3.

Extensions of the Gerds et al. (2012) modeling approach to the case of $K \geq 2$ competing events will also be explored.

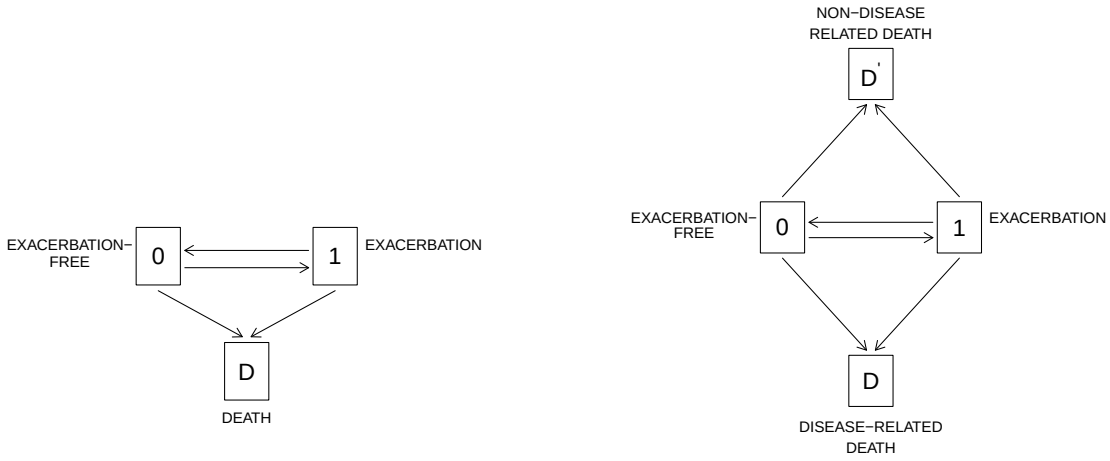
6.3 Future research on recurrent and terminal events

6.3.1 Mean function regression and intermittent observation

When the non-fatal recurrent event process is under intermittent observation and the failure time is subject to right-censoring, we again have a dual observation scheme (Boruvka and Cook, 2016). It is important to investigate how the marginal Ghosh and Lin (2002) approach can be adapted to this setting. Similar challenges arise here as in cumulative incidence function regression. Following the notation of Chapter 4, we let $\{N(t), 0 < t\}$ denote the counting process for recurrent non-fatal events, and let $T = \inf\{t > 0 : Z(t) \in \{D_1, D_2, \ldots, D_K\}\}$ be the failure time in the setting of Figure 1.1(e). The pseudo-Poisson procedure requires the number of non-fatal events experienced over $[0, a_m]$ (i.e. $N(a_m)$) to be exactly known at a_m. If the non-fatal event process is only intermittently observed and an individual dies between two successive assessment times (i.e. $T \in [a_{m-1}, a_m)$), $N(a_m)$ is unknown; it is unclear whether $N(a_m) = N(a_{m-1})$ or $N(a_m) > N(a_{m-1})$. It will be important to examine the bias in pseudo-Poisson estimation when it is assumed that $N(a_m) = N(a_{m-1})$ whenever $T \in (a_{m-1}, a_m]$. Strategies for mitigating this bias include inverse weighting and a full model specification, as discussed before in Section 6.2.1.

6.3.2 Estimands for disease processes involving recurrent episodes and fatal events

The marginal Ghosh and Lin (2002) model, in combination with extensive secondary analyses of survival, has been recommended in this thesis for type (II) trial settings involving recurrent and terminal events. Intensity-based Cox regression is useful for assessing treatment effects on fatal and non-fatal events, and for assessing possible dependence of the fatal event intensity on the non-fatal event. Furberg et al. (2023) considered a slightly different strategy, and combined the estimands of treatment effect estimators from mean function regression and Cox regression into a bivariate summary measure. As briefly outlined in Section 4.6, mean function methods for non-fatal events are not sufficiently informative about response to treatment in disease process settings where the non-fatal events are associated with a duration.



(a) All-cause death.

(b) Death from two different causes.

Figure 6.1: Multistate diagrams for processes involving recurrent episodes and terminal events.

This future research builds on the work of Lee and Cook (2019) and Andersen et al. (2019), and explores strategies for estimand specification in disease processes involving recurrent episodes and fatal events. There are many disease processes in which the non-fatal events are not instantaneous but instead indicate the onset of symptomatic periods. Examples include recurrent hospitalizations in cardiovascular diseases (O'Connor et al., 2009; McMurray et al., 2014; Rogers et al., 2014) or psychiatric disorders (Kessing et al., 2004). A motivating context here is a multi-center randomized clinical trial designed to assess the effect of daily administration of ciprofloxacin versus standard antibiotic care on the occurrence and resolution of acute exacerbations in chronic bronchitis (Grossman et al., 1998); see Figure 6.1(a) where individuals move from state 0 to state 1 upon the occurrence of a new bronchial exacerbation, and vice versa upon resolution (Cook and Lawless, 2007, Sections 1.5.3 and 6.7.2). Some individuals tend to experience frequent exacerbations of shorter duration, whereas others tend to have less frequent exacerbations but of longer duration.

With such processes, we recommend to evaluate treatment efficacy based on comparisons of (i) the expected number of exacerbations over $[0, t]$ (i.e. number of entries to state 1)

$$\mu(t) = \mathbb{E}(N_1(t)) = \int_0^t P_{00}(0, u)\mathbb{E}(dN_1(u)|Z(u) = 0) \ ,$$

(ii) the expected cumulative time spent with symptoms of acute exacerbations (i.e. in state 1)

$$\widetilde{\mu}(t) = \mathbb{E}\left(\int_0^t \mathbb{1}(Z(u) = 1)du\right) = \int_0^t P_{01}(0, u)du \ ,$$

and (iii) the distribution function $F(t) = P_{0D}(0, t)$ for death. While the Ghosh and Lin (2002) approach yields a scalar model-based estimand for feature (i) and the cloglog model (2.2.6) for feature (iii), direct regression modeling for the marginal feature (ii) is less developed and we consider this here. We propose in particular a Ghosh-Lin type model for the mean function $\widetilde{\mu}(t)$ of the form

$$\widetilde{\mu}(t|X) = \widetilde{\mu}_0(t) \exp(\widetilde{\beta}(t)X) \ , \tag{6.3.1}$$

where $\widetilde{\mu}_0(t)$ is an unspecified mean function and

$$\widetilde{\beta}(t) = \log\left(\frac{\widetilde{\mu}(t|X = 1)}{\widetilde{\mu}(t|X = 0)}\right)$$

is a function of time. If $\widetilde{\beta}(t) = \widetilde{\beta}$, we obtain a scalar estimand that allows for causal inferences according to the estimand principles of Section 2.2.2. In a similar spirit to the investigations of Section 4.4.1, we evaluate model (6.3.1) in terms of adequacy, estimability and interpretability.

Extensions to disease settings as those in Figure 6.1(b) are also of interest.

6.4 Future research on utility-based estimands

Further work will be conducted on approaches based on ranking through utility functions.

The problem described in Section 6.3.2 may also be addressed by a utility-based approach. Figure 6.2 displays an alternative multistate representation of the reversible illness-death process in Figure

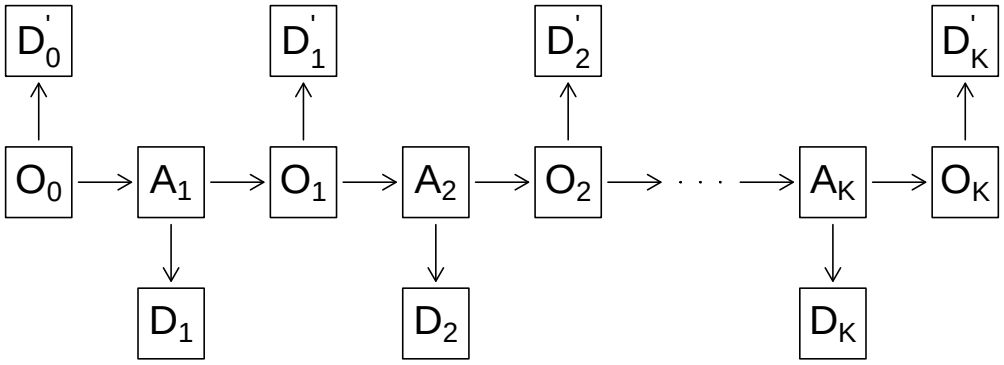


Figure 6.2: Multistate diagram for recurrent hospitalizations and death.

6.1(a); here we consider the problem of Section 6.3.2 in the context of recurrent hospitalizations in cardiovascular (CV) diseases (O'Connor et al., 2009) or psychiatric disorders (Andersen et al., 2019), and assume that the process starts in the initial state O_0 at time 0. In a CV trial, for instance, state O_0 may represent the condition of being out of hospital at the time of treatment assignment, with states A_k and O_k entered upon admission or discharge from the kth hospital stay, $k = 1, \ldots, K$. The absorbing state D_k is entered upon death by individuals who died during the kth hospital stay, whereas state D'_k is entered upon death by individuals who died out of hospital, following k hospital admissions. We let $S_k = \inf\{t > 0 : Z(t) = A_k\}$, $\widetilde{S}_k = \inf\{t > 0 : Z(t) = O_k\}$ and $T = \inf\{t > 0 : Z(t) \in \{D_1, \ldots, D_K, D'_0, D'_1, \ldots, D'_K\}\}$ be the kth admission time to hospital, the kth discharge time from hospital and the time to death, respectively. If $u_{D_k} = u_{D'_k} = 0$, the expected cumulative utility score (1.2.4) reduces to

$$D(\tau) = \sum_{k=0}^{K} u_{O_k} \int_0^\tau P(Z(u) = O_k)du \; + \; \sum_{k=1}^{K} u_{A_k} \int_0^\tau P(Z(u) = A_k)du \; , \tag{6.4.1}$$

where units of time spent at the kth hospital stay could be assigned a higher utility than time units spent at the $(k+1)^{th}$ stay; that is, $u_{A_k} > u_{A_{k+1}}$. Utility scores for O_k could likewise decrease with each additional hospital stay.

The utility-based estimand (6.4.1) can be graphically supplemented by stacked probability plots similar to those in Figure 5.12. The restricted mean time alive over $[0, \tau]$, i.e.

$$\int_0^\tau P(T > t)dt = \int_0^\tau \underbrace{P(Z(t) \in \{A_1, \ldots, A_K\})dt}_{\text{in hospital}} + \int_0^\tau \underbrace{P(Z(t) \in \{O_0, O_1, \ldots, O_K\})dt}_{\text{out of hospital}}$$

can be decomposed into the restricted mean times alive in and out of hospital. Since

$$P(S_1 \leq t) = \sum_{k=1}^{K} \big(P(Z(t) = A_k) + P(Z(t) = O_k) + P(Z(t) = D_k) + P(Z(t) = D'_k)\big)$$

and

$$P(\widetilde{S}_1 \leq t) = \sum_{k=1}^{K} \big(P(Z(t) = O_k) + P(Z(t) = D'_k)\big) + \sum_{k=2}^{K} \big(P(Z(t) = A_k) + P(Z(t) = D_k)\big) \ ,$$

the area between the $P(S_1 \leq t)$ and $P(\widetilde{S}_1 \leq t)$ curves, i.e.

$$\int_0^\tau \big(P(S_1 \leq t) - P(\widetilde{S}_1 \leq t)\big)dt = \int_0^\tau \big(P(Z(t) = A_1) + P(Z(t) = D_1)\big)dt \ ,$$

represents the restricted mean time in hospital at the kth visit.

Cumulative utilities can accommodate intercurrent events, besides addressing disease burden and quality of life. This is facilitated by including states that deal with disease severity and effects of treatment in a model. In an oncology trial involving alternative chemotherapies (Ludwig Breast Cancer Study Group, 1988), we could for example split a relapse-free state into two states: one while an individual is undergoing chemotherapy and one for when toxic effects of the therapy have subsided. The first of the two states would have a lower utility score than the second. In a CV trial, the severity and after-effects of strokes or myocardial infarctions could likewise be represented with multiple states, as opposed to simply recognizing the occurrences of events.

Another topic concerns how to assign utility scores both to units of time spent in states and to "point" events that have an instantaneous impact.

References

Aalen, O. (1980). A model for nonparametric regression analysis of counting processes. In Klonecki, W., Kozek, A., and Rosiński, J., editors, *Mathematical Statistics and Probability Theory*, pages 1–25, New York, NY. Springer New York.

Aalen, O., Borgan, Ø., and Gjessing, H. (2008). *Survival and Event History Analysis: A Process Point of View*. Springer Science + Business Media, New York, NY.

Aalen, O. O. (2012). Armitage lecture 2010: Understanding treatment effects: the value of integrating longitudinal data and survival analysis. *Statistics in Medicine*, 31(18):1903–1917.

Aalen, O. O., Borgan, , and Fekjær, H. (2001). Covariate adjustment of event histories estimated from Markov chains: the additive approach. *Biometrics*, 57(4):993–1001.

Aalen, O. O., Cook, R. J., and Røysland, K. (2015). Does Cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Analysis*, 21(4):579–593.

Allignol, A., Beyersmann, J., and Schumacher, M. (2008). mvna: An R package for the Nelson-Aalen estimator in multistate models. *R News*, 8(2):48–50.

Allignol, A., Schumacher, M., and Beyersmann, J. (2011). Empirical transition matrix of multi-state models: the etm package. *Journal of Statistical Software*, 38(4):1–15.

Andersen, P., Angst, J., and Ravn, H. (2019). Modeling marginal features in studies of recurrent events in the presence of a terminal event. *Lifetime Data Analysis*, 25(4):681–695.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York, NY.

Andersen, P. K. and Keiding, N. (2012). Interpretability and importance of functionals in competing risks and multistate models. *Statistics in Medicine*, 31(11-12):1074–1088.

Andersen, P. K. and Ravn, H. (2023). *Models for Multi-State Survival Data: Rates, Risks, and Pseudo-Values*. Chapman and Hall/CRC.

Anker, S. D. and McMurray, J. J. V. (2012). Time to move on from 'time-to-first': should all events be included in the analysis of clinical trials? *European Heart Journal*, 33(22):2764–2765.

Anker, S. D., Schroeder, S., Atar, D., Bax, J. J., Ceconi, C., Cowie, M. R., Crisp, A., Dominjon, F., Ford, I., Ghofrani, H.-A., Gropper, S., Hindricks, G., Hlatky, M. A., Holcomb, R., Honarpour, N., Jukema, J. W., Kim, A. M., Kunz, M., Lefkowitz, M., Le Floch, C., Landmesser, U., McDonagh, T. A., McMurray, J. J., Merkely, B., Packer, M., Prasad, K., Revkin, J., Rosano, G. M., Somaratne, R., Stough, W. G., Voors, A. A., and Ruschitzka, F. (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.

Austin, P. C., Steyerberg, E. W., and Putter, H. (2021). Fine-Gray subdistribution hazard models to simultaneously estimate the absolute risk of different event types: cumulative total failure probability may exceed 1. *Statistics in Medicine*, 40(19):4200–4212.

Bacchetti, P., Boylan, R. D., Terrault, N. A., Monto, A., and Berenguer, M. (2010). Non-Markov multistate modeling using time-varying covariates, with application to progression of liver fibrosis due to Hepatitis C following liver transplant. *The International Journal of Biostatistics*, 6(1):1–16.

Barnett, H., Taylor, D., Eliasziw, M., Fox, A., Ferguson, G., Haynes, R., Rankin, R., Clagett, G., Hachiniski, V., Sackett, D., and Thorpe, K. (1998). Benefit of carotid endarterectomy in patients with symptomatic moderate or severe stenosis. *New England Journal of Medicine*, 339(20):1415–1425.

Bateman, R. J., Smith, J., Donohue, M. C., Delmar, P., Abbas, R., Salloway, S., Wojtowicz, J., Blennow, K., Bittner, T., Black, S. E., Klein, G., Boada, M., Grimmer, T., Tamaoka, A., Perry, R. J., Turner, R. S., Watson, D., Woodward, M., Thanasopoulou, A., Lane, C., Baudler, M., Fox, N. C., Cummings, J. L., Fontoura, P., and Doody, R. S. (2023). Two phase 3 trials of

gantenerumab in early Alzheimer's disease. *New England Journal of Medicine*, 389(20):1862–1876.

Beal, E. W., Dalmacy, D., Paro, A., Hyer, J. M., Cloyd, J., Dillhoff, M., Ejaz, A., and Pawlik, T. M. (2022). Comparing minimally invasive and open pancreaticoduodenectomy for the treatment of pancreatic cancer: a win ratio analysis. *Journal of Gastrointestinal Surgery: Official Journal of the Society for Surgery of the Alimentary Tract*, 26(8):1697–1704.

Bebu, I. and Lachin, J. M. (2015). Large sample inference for a win ratio analysis of a composite outcome based on prioritized components. *Biostatistics*, 17(1):178–187.

Bennett, K. J. and Torrance, G. W. (1996). *Measuring health state preferences and utilities: rating scale, time trade-off, and standard gamble techniques.* In: Spilker B, ed. Quality of life and pharmacoeconomics in clinical trials. 2nd edn. Philadelphia, PA: Lippincott-Raven.

Beyersmann, J., Allignol, A., and Schumacher, M. (2012). *Competing Risks and Multistate Models in R*. Springer-Verlage, New York, NY.

Blanche, P., Holt, A., and Scheike, T. (2023). On logistic regression with right censored data, with or without competing risks, and its use for estimating treatment effects. *Lifetime Data Analysis*, 29(2):441–482.

Bleichrodt, H. (2002). A new explanation for the difference between time trade-off utilities and standard gamble utilities. *Health Economics*, 11(5):447–456.

Boher, J.-M., Filleron, T., Giorgi, R., Kramar, A., and Cook, R. J. (2017). Goodness-of-fit test for monotone proportional subdistribution hazards assumptions based on weighted residuals. *Statistics in Medicine*, 36(2):362–377.

Bonneville, E. F., de Wreede, L. C., and Putter, H. (2024). Why you should avoid using multiple Fine–Gray models: insights from (attempts at) simulating proportional subdistribution hazards data. *Journal of the Royal Statistical Society Series A: Statistics in Society*.

Booth, C. M. and Eisenhauer, E. A. (2012). Progression-free survival: meaningful or simply measurable? *Journal of Clinical Oncology*, 30(10):1030–1033.

Bornkamp, B., Rufibach, K., Lin, J., Liu, Y., Mehrotra, D. V., Roychoudhury, S., Schmidli, H., Shentu, Y., and Wolbers, M. (2021). Principal stratum strategy: Potential role in drug development. *Pharmaceutical Statistics*, 20(4):737–751.

Boruvka, A. and Cook, R. J. (2016). Sieve estimation in a Markov illness-death process under dual censoring. *Biostatistics*, 17(2):350–363.

Brazier, J., Roberts, J., and Deverill, M. (2002). The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics*, 21(2):271–292.

Broglio, K. R. and Berry, D. A. (2009). Detecting an Overall Survival Benefit that is derived from Progression-Free Survival. *JNCI: Journal of the National Cancer Institute*, 101(23):1642–1649.

Buja, A., Brown, L., Berk, R., George, E., Pitkin, E., Traskin, M., Zhang, K., and Zhao, L. (2019). Models as approximations I: consequences illustrated with linear regression. *Statistical Science*, 34(2):523–544.

Buyse, M. (2010). Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Statistics in Medicine*, 29(30):3245–3257.

Casey, M., Degtyarev, E., Lechuga, M. J., Aimone, P., Ravaud, A., Motzer, R. J., Liu, F., Stalbovskaya, V., Tang, R., Butler, E., Sailer, O., Halabi, S., and George, D. (2021). Estimand framework: Are we asking the right questions? A case study in the solid tumor setting. *Pharmaceutical Statistics*, 20(2):324–334.

Chiou, S. H., Xu, G., Yan, J., and Huang, C.-Y. (2023). Regression modeling for recurrent events possibly with an informative terminal event using R package reReg. *Journal of Statistical Software*, 105(5):1–34.

Claggett, B., Pocock, S., L.J. Wei, L., Pfeffer, M. A., McMurray, J. J., and Solomon, S. D. (2018). Comparison of time-to-first event and recurrent event methods in randomized clinical trials. *Circulation*, 138(6):570–577.

Coleman, R. (2006). Clinical features of metastatic bone disease and risk of skeletal morbidity. *Clinical Cancer Research*, 12(20):6243–6249.

Coleman, R. E. (2001). Metastatic bone disease: clinical features, pathophysiology and treatment strategies. *Cancer Treatment Reviews*, 27(3):165–176.

Comment, L., Mealli, F., Haneuse, S., and Zigler, C. (2019). Survivor average causal effects for continuous time: a principal stratification approach to causal inference with semicompeting risks. *arXiv, https://arxiv.org/abs/1902.09304*.

Cook, R. and Lawless, J. (1997). Marginal analysis of recurrent events and a terminating event. *Statistics in Medicine*, 16(8):911–924.

Cook, R., Lawless, J., Lakhal-Chaieb, L., and Lee, K.-A. (2009). Robust estimation of mean functions and treatment effects for recurrent events under event-dependent censoring and termination: application to skeletal complications in cancer metastatic to bone. *Journal of the American Statistical Association*, 104(485):60–75.

Cook, R. J. and Lawless, J. F. (2007). *Statistical Analysis of Recurrent Events*. Springer-Verlage, New York, NY.

Cook, R. J. and Lawless, J. F. (2014). Statistical issues in modeling chronic disease in cohort studies. *Statistics in Biosciences*, 6(1):127–161.

Cook, R. J. and Lawless, J. F. (2018). *Multistate Models for the Analysis of Life History Data*. CRC Press, New York, NY.

Cook, R. J. and Lawless, J. F. (2020). Failure time studies with intermittent observation and losses to follow-up. *Scandinavian Journal of Statistics*, 47(4):1035–1063.

Cook, R. J., Lawless, J. F., and Lee, K.-A. (2003). Cumulative processes related to event histories. *SORT*, 27(1):13–30.

Cook, R. J. and Lawless, J. L. (2022). Life history analysis: a review and some current issues. *Canadian Journal of Statistics*, 50(4):1270–1298.

Cortese, G. and Scheike, T. H. (2022). Efficient estimation of the marginal mean of recurrent events. *Journal of the Royal Statistical Society Series C*, 71(5):1787–1821.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

Cox, D. R. and Miller, H. D. (1965). *The Theory of Stochastic Processes*. Wiley, New York.

Datta, S. and Satten, G. (2002). Estimation of integrated transition hazards and stage occupation probabilities for non-Markov systems under dependent censoring. *Biometrics*, 58(4):792–802.

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(450):407–424.

Dawid, A. P. and Musio, M. (2022). Effects of causes and causes of effects. *Annual Review of Statistics and Its Application*, 9(1):261–287.

Deng, Y., Han, S., and Zhou, X.-H. (2024). Inference for cumulative incidences and treatment effects in randomized controlled trials with time-to-event outcomes under ICH E9 (R1). *arXiv, https://arxiv.org/abs/2401.14684*.

Dewitt, B. and Torrance, G. W. (2020). Incorporating mortality in health utility measures. *Medical Decision Making*, 40(7):862–872.

DiRienzo, A. G. and Lagakos, S. W. (2001). Effects of model misspecification on tests of no randomized treatment effect arising from Cox's proportional hazards model. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(4):745–757.

Dobler, D. and Pauly, M. (2017). Approximate tests for the equality of two cumulative incidence functions of a competing risk. *Statistics*, 51(6):1238–1258.

Dong, G., Hoaglin, D. C., Qiu, J., Matsouaka, R. A., Chang, Y., Wang, J., and Vandemeulebroecke, M. (2020a). The win ratio: on interpretation and handling of ties. *Statistics in Biopharmaceutical Research*, 12(1):99–106.

Dong, G., Huang, B., Chang, Y.-W., Seifu, Y., Song, J., and Hoaglin, D. C. (2020b). The win ratio: impact of censoring and follow-up time and use with nonproportional hazards. *Pharmaceutical Statistics*, 19(3):168–177.

Dong, G., Huang, B., Wang, D., Verbeeck, J., Wang, J., and Hoaglin, D. C. (2021). Adjusting win statistics for dependent censoring. *Pharmaceutical Statistics*, 20(3):440–450.

Dong, G., Mao, L., Huang, B., Gamalo-Siebers, M., Wang, J., Yu, G., and Hoaglin, D. C. (2020c). The inverse-probability-of-censoring weighting (IPCW) adjusted win ratio statistic: an unbiased estimator in the presence of independent censoring. *Journal of Biopharmaceutical Statistics*, 30(5):882–899.

Dormuth, I., Liu, T., Xu, J., Pauly, M., and Ditzhaus, M. (2023). A comparative study to alternatives to the log-rank test. *Contemporary Clinical Trials*, 128(7):107–165.

Du, Q., Wang, Y.-J., Yang, S., Zhao, Y.-Y., and Han, P. (2014). Liraglutide for the treatment of type 2 diabetes mellitus: a meta-analysis of randomized placebo-controlled trials. *Advances in Therapy*, 31(11):1182–1195.

Eaton, A. A. and Zabor, E. C. (2022). Analysis of composite endpoints with component-wise censoring in the presence of differential visit schedules. *Statistics in Medicine*, 41(9):1599–1612.

Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589–599.

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1):36–48.

EMA (2017). Points to consider on multiplicity issues in clinical trials.

Erdmann, A., Beyersmann, J., and Bluhmki, E. (2024). Comparison of nonparametric estimators of the expected number of recurrent events. *Pharmaceutical Statistics*, 23(3):339–369.

Eriksson, F., Li, J., Scheike, T., and Zhang, M. J. (2015). The proportional odds cumulative incidence model for competing risks. *Biometrics*, 71(3):687–695.

Evans, S. R. and Follmann, D. (2016). Using outcomes to analyze patients rather than patients to analyze outcomes: a step toward pragmatism in benefit:risk evaluation. *Statistics in Biopharmaceutical Research*, 8(4):386–393.

FDA (2007). Guidance for industry: clinical trial endpoints for the approval of cancer drugs and biologics. *Washington, DC*.

FDA (2016). Non-inferiority clinical trials to establish effectiveness. *Guidance for Industry*.

Feeny, D., Furlong, W., Boyle, M., and Torrance, G. W. (1995). Multi-attribute health status classification systems. Health Utilities Index. *PharmacoEconomics*, 7(6):490–502.

Feeny, D. H. and Torrance, G. W. (1989). Incorporating utility-based quality-of-life assessment measures in clinical trials: Two examples. *Medical Care*, 27(3):190–204.

Feinstein, A. R. and Horwitz, R. I. (1982). Double standards, scientific methods, and epidemiologic research. *New England Journal of Medicine*, 307(26):1611–1617.

Ferreira, J. P., Jhund, P. S., Duarte, K., Claggett, B. L., Solomon, S. D., Pocock, S., Petrie, M. C., Zannad, F., and McMurray, J. J. V. (2020). Use of the win ratio in cardiovascular trials. *JACC: Heart Failure*, 8(6):441–450.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509.

Finkelstein, D. M. and Schoenfeld, D. A. (1999). Combining mortality and longitudinal measures in clinical trials. *Statistics in Medicine*, 18(11):1341–1354.

Finkelstein, D. M. and Schoenfeld, D. A. (2019). Graphing the win ratio and its components over time. *Statistics in Medicine*, 38(1):53–61.

Fisher, R. A. (1925). *Statistical Methods for Research Workers (11th ed. rev.)*. Oliver and Boyd: Edinburgh.

Franchini, M., Frattini, F., Crestani, S., and Bonfanti, C. (2013). Bleeding complications in patients with hematologic malignancies. *Seminars in Thrombosis and Hemostasis*, 39(1):94–100.

Freemantle, N. nd Calvert, M., Wood, J., Eastaugh, J., and Griffin, C. (2003). Composite outcomes in randomized trials greater precision but with greater uncertainty? *JAMA*, 289(19):2554–2559.

Freidlin, B. and Korn, E. L. (2005). Testing treatment effects in the presence of competing risks. *Statistics in Medicine*, 24(11):1703–1712.

Friedman, L., Furberg, C., and DeMets, D. (2015). *Fundamentals of Clinical Trials*. Springer New York, 5th edition.

Furberg, J., Andersen, P. K., Korn, S., Overgaard, M., and Ravn, H. (2023). Bivariate pseudo-observations for recurrent event analysis with terminal events. *Lifetime Data Analysis*, 29(2):256–287.

Furberg, J. F., Andersen, P. K., Scheike, T., and Ravn, H. (2024). Simulation-based sample size calculations of marginal proportional means models for recurrent events with competing risks. *Pharmaceutical Statistics*, 0(0):1–22.

Furberg, J. K., Rasmussen, S., Andersen, P. K., and 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.

Fyles, A. W., McCready, D. R., Manchul, L. A., Trudeau, M. E., Merante, P., Pintilie, M., Weir, L. M., and Olivotto, I. A. (2004). Tamoxifen with or without breast irradiation in women 50 years of age or older with early breast cancer. *New England Journal of Medicine*, 351(10):963–970.

Gelber, R. D., Cole, B. F., Gelber, S., and Goldhirsch, A. (1995). Comparing treatments using quality-adjusted survival: the Q-TWiST method. *The American Statistician*, 49(2):161–169.

Gelber, R. D., Gelman, R. S., and Goldhirsch, A. (1989). A quality-of-life-oriented endpoint for comparing therapies. *Biometrics*, 45(3):781–795.

Genetti, L., Cortese, G., Ravn, H., and Scheike, T. (2024). Efficient estimation of the marginal mean of recurrent events in randomized clinical trials. *Manuscript*, 0(0):0–0.

Gerds, T. A., Scheike, T. H., and Andersen, P. K. (2012). Absolute risk regression for competing risks: interpretation, link functions, and prediction. *Statistics in Medicine*, 31(29):3921–3930.

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

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

Giordano, R., Stephenson, W., Liu, R., Jordan, M. I., and Broderick, T. (2020). A Swiss army infinitesimal jackknife.

Glasziou, P. P., Simes, R. J., and Gelber, R. D. (1990). Quality adjusted survival analysis. *Statistics in Medicine*, 9:1259–1276.

Glidden, D. (2002). Robust inference for event probabilities with non-Markov event data. *Biometrics*, 58(2):361–368.

Grambauer, N., Schumacher, M., and Beyersmann, J. (2010). Proportional subdistribution hazards modeling offers a summary analysis, even if misspecified. *Statistics in Medicine*, 29(7-8):875–884.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526.

Gray, R. J. (1988). A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk. *The Annals of Statistics*, 16(3):1141–1154.

Grossman, R., Mukherjee, J., Vaughan, D., Eastwood, C., Cook, R. J., LaForge, J., and Lampron, N. (1998). A 1-year community-based health economic study of ciprofloxacin vs usual antibiotic treatment in acute exacerbations of chronic bronchitis: The Canadian ciprofloxacin health economic study group. *Chest*, 113(1):131–141.

Halekoh, U., Højsgaard, S., and Yan, J. (2006). The R package geepack for generalized estimating equations. *Journal of Statistical Software*, 15(2):1–11.

Hauser, S., Bar-Or, A., Comi, G., Giovannoni, G., Hartung, H.-P., Hemmer, B., Lublin, F., Montalban, X., Rammohan, K. W., Selmaj, K., Traboulsee, A., Wolinsky, J. S., Arnold, D. L., Klingelschmitt, G., Masterman, D., Fontoura, P., Belachew, S., Chin, P., Mairon, N., Garren, H.,

and Kappos, L. (2017). Ocrelizumab versus interferon beta-1a in relapsing multiple sclerosis. *New England Journal of Medicine*, 376(3):221–234.

He, P., Eriksson, F., Scheike, T. H., and Zhang, M.-J. (2016). A proportional hazards regression model for the subdistribution with covariates-adjusted censoring weight for competing risks data. *Scandinavian Journal of Statistics*, 43(1):103–122.

Heddle, N. M., Cook, R. J., Tinmouth, A., Kouroukis, C. T., Hervig, T., Klapper, E., Brandwein, J. M., Szczepiorkowski, Z. M., AuBuchon, J. P., Barty, R. L., Lee, K.-A., and for the SToP Study investigators of the BEST Collaborative (2009). A randomized controlled trial comparing standard- and low-dose strategies for transfusion of platelets (SToP) to patients with thrombocytopenia. *Blood*, 113(7):1564–1573.

Henshall, C., Latimer, N., Sansom, L., and Ward, R. (2016). Treatment switching in cancer trials: issues and proposals. *International Journal of Technology Assessment in Health Care*, 32(3):167–174.

Hernán, M. and Robins, J. (2020). *Causal Inference: What if*. Boca Raton: Chapman & Hill/CRC.

Hernán, M. A. (2010). The hazards of hazard ratios. *Epidemiology*, 21(1):13–15.

Hernán, M. A. and Scharfstein, D. (2018). Cautions as regulators move to end exclusive reliance on intention to treat. *Annals of Internal Medicine*, 168(7):515–516.

Hirniak, S., Edginton, A. N., Iorio, A., Alsabbagh, W., Hajducek, D. M., and Wong, W. W. L. (2024). Health utilities in adults with hemophilia A: a retrospective cohort study. *Haemophilia*, 30(3):733–742.

Holst, K. K., Scheike, T. H., and Hjelmborg, J. B. (2016). The liability threshold model for censored twin data. *Computational Statistics and Data Analysis*, 93:324–335.

Horsman, J., Furlong, W., Feeny, D., and Torrance, G. W. (2003). The Health Utilities Index (hui): concepts, measurement properties and applications. *Health Quality Life Outcomes*, 54(1):1–16.

Hortobagyi, G. N., Theriault, R. L., Porter, L., Blayney, D., Lipton, A., Sinoff, C., Wheeler, H., Simeone, J. F., Seaman, J., and Knight, R. D. (1996). Efficacy of pamidronate in reducing

skeletal complications in patients with breast cancer and lytic bone metastases. *New England Journal of Medicine*, 335(34):1785–1791.

ICH E9 Steering Committee (2019). ICH E9 (R1): Addendum to statistical principles for clinical trials on choosing appropriate estimands and defining sensitivity analyses in clinical trials. *International Conference on Harmonization*.

Jaeckel, L. A. (1972). The infinitesimal jackknife. *Bell Labs Memorandum*, MM 72-1215-11.

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

Jeong, J.-H. and Fine, J. P. (2006). Direct parametric inference for cumulative incidence function. *Journal of the Royal Statistical Society, Series C, Applied Statistics*, 55:187–200.

Jeong, J.-H. and Fine, J. P. (2007). Parametric regression on cumulative incidence function. *Biostatistics*, 8(2):184–196.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, 2nd edition.

Karimi, M., Brazier, J., and Paisley, S. (2017). Are preferences over health states informed? *Health Quality Life Outcomes*, 15(1):1–11.

Keiding, N., Klein, J. P., and Horowitz, M. M. (2001). Multi-state models and outcome prediction in bone marrow transplantation. *Statistics in Medicine*, 20(12):1871–1885.

Kessing, L. V., Hansen, M. G., Andersen, P. K., and Angst, J. (2004). The predictive effect of episodes on the risk of recurrence in depressive and bipolar disorders – a life-long perspective. *Acta Psychiatrica Scandinavica*, 109(5):339–344.

Kind, P. (1996). *The EuroQol instrument: an index of health-related quality of life.* In: Spilker B, ed. Quality of life and pharmacoeconomics in clinical trials. 2nd edn. Philadelphia, PA: Lippincott-Raven.

Klein, J., Van Houwelingen, H., Scheike, T., and Ibrahim, J. (2014). *Handbook of Survival Analysis*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis Techniques for Censored and Truncated Data*.

Klein, P. J. and Andersen, P. K. (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics*, 61(1):223–229.

Kondo, T., Gasparyan, B., Jhund, P., Bengtsson, O., Claggett, B., de Boer, R., Hernandez, A., Inzucchi, S., Kosiborod, M., Køber, L., Lam, C., Langkilde, A., Martinez, F., Petersson, M., Ponikowski, P., Sabatine, M., Shah, S., Sjostrand, M., Wilderang, U., and McMurray, J. (2023). Use of win statistics to analyze outcomes in the DAPA-HF and DELIVER trials. *NEJM Evidence*, 2(11).

Latimer, N., Abrams, K., and Siebert, U. (2019a). Two-stage estimation to adjust for treatment switching in randomized trials: a simulation study investigating the use of inverse probability weighting instead of re-censoring. *BMC Medical Research Methodology*, 19(1):1–19.

Latimer, N., White, I., Abrams, K., and Siebert, U. (2019b). Causal inference for long-term survival in randomized trials with treatment switching: Should re-censoring be applied when estimating counterfactual survival times? *Statistical Methods in Medical Research*, 28(8):2475–2493.

Latouche, A., Boisson, V., Chevret, S., and Porcher, R. (2007). Misspecified regression model for the subdistribution hazard of a competing risk. *Statistics in Medicine*, 26(5):965–974.

Latouche, A., Porcher, R., and Chevret, S. (2004). Sample size formula for proportional hazards modelling of competing risks. *Statistics in Medicine*, 23(21):3263–3274.

Lawless, J. and Cook, R. (2019). A new perspective on loss to follow-up in failure time and life history studies. *Statistics in Medicine*, 38(23):4583–4610.

Lawless, J. F. (2002). *Statistical Models and Methods for Lifetime Data*. Wiley Series in Probability and Statistics.

Lee, J. and Cook, R. J. (2019). On estimands arising from misspecified semiparametric rate-based analysis of recurrent episodic conditions. *Statistics in Medicine*, 38(25):4977–4998.

Li, G. and Yang, Q. (2016). Joint inference for competing risks survival data. *Journal of the American Statistical Association*, 111(515):1289–1300.

Li, J., Scheike, T. H., and Zhang, M. J. (2015). Checking Fine and Gray subdistribution hazards model with cumulative sums of residuals. *Lifetime Data Analysis*, 21(2):197–217.

Lin, D. Y. and Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, 84(408):1074–1078.

Lipkovich, I., Ratitch, B., Qu, Y., Zhang, X., Shan, M., and Mallinckrodt, C. (2022). Using principal stratification in analysis of clinical trials. *Statistics in Medicine*, 41(19):3837–3877.

Llewellyn-Thomas, H. A., Sutherland, H. J., Tibshirani, R., Ciampi, A., Till, J. E., and Boyd, N. F. (1984). Describing health states: methodologic issues in obtaining values for health states. *Medical Care*, 22:543–552.

Ludwig Breast Cancer Study Group (1988). Combination adjuvant chemotherapy for node-positive breast cancer. *New England Journal of Medicine*, 319(11):677–683.

Luo, X., Qiu, J., Bai, S., and Tian, H. (2017). Weighted win loss approach for analyzing prioritized outcomes. *Statistics 2452–2465in Medicine*, 36(15):2452–2465.

Ma, B., Chen, G., Badji, S., and Petrie, D. (2024). Mapping the 12-item World Health Organization disability assessment schedule 2.0 (WHODAS 2.0) onto the assessment of quality of life (AQol)-4D utilities. *Quality Life Research*, 33(2):411–422.

Manitz, J., Kan-Dobrosky, N., Buchner, H., Casadebaig, M.-L., Degtyarev, E., Dey, J., Haddad, V., Jie, F., Martin, E., Mo, M., Rufibach, K., Shentu, Y., Stalbovskaya, V., (Sammi)Tang, R., Yung, G., and Zhou, J. (2022). Estimands for overall survival in clinical trials with treatment switching in oncology. *Pharmaceutical Statistics*, 21(1):150–162.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50(3):163–170.

Mao, L. (2018). On the alternative hypotheses for the win ratio. *Biometrics*, 75(1):347–351.

Mao, L. (2023). On restricted mean time in favor of treatment. *Biometrics*, 79(1):61–72.

Mao, L. and 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. M., and Li, Y. (2022). On recurrent-event win ratio. *Statistical Methods in Medical Research*, 31(6):1120–1134.

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

Mao, L. and Wang, T. (2021). A class of proportional win-fractions regression models for composite outcomes. *Biometrics*, 77(4):1265–1275.

Mao, L. and Wang, T. (2024). Dissecting the restricted mean time in favor of treatment. *Journal of Biopharmaceutical Statistics*, 34(1):111–126.

Marshall, G., Blacklock, J. W. S., Cameron, C., Capon, N. B., Cruickshank, R., Gaddum, J. H., Heaf, F. R. G., Bradford Hill, A., Houghton, L. E., Clifford Hoyle, J., Raistrick, H., Scadding, J. G., Tytler, W. H., Wilson, G. S., and D'Arcy Hart, P. (1948). Streptomycin treatment of pulmonary tuberculosis. A Medical Research Council investigation. *British Medical Journal*, 0(0):0–0.

Marso, S. P., Daniels, G. H., Brown-Frandsen, K., Kristensen, P., Mann, J., Nauck, M. A., Nissen, S. E., Pocock, S., Poulter, N. R., Ravn, L. S., Steinberg, W. M., Stockner, M., Zinman, B., Bergenstal, R. M., and Buse, J. B. (2016). Liraglutide and cardiovascular outcomes in type 2 diabetes. *New England Journal of Medicine*, 375(4):311–322.

Martinussen, T., Vansteelandt, S., and Andersen, P. K. (2020). Subtleties in the interpretation of hazard contrasts. *Lifetime Data Analysis*, 26(4):833–855.

Maurer, M. S., Schwartz, J. H., Gundapaneni, B., Elliott, P. M., Merlini, G., Waddington-Cruz, M., Kristen, A. V., Grogan, M., Witteles, R., Damy, T., Drachman, B. M., Shah, S. J., Hanna, M., Judge, D. P., Barsdorf, A. I., Huber, P., Patterson, T. A., Riley, S., Schumacher, J., Stewart, M., Sultan, M. B., and Rapezzi, C. (2018). Tafamidis treatment for patients with transthyretin amyloid cardiomyopathy. *New England Journal of Medicine*, 379(11):1007–1016.

McMurray, J., Packer, M., Desai, A. S., Gong, J., Lefkowitz, M. P., Rizkala, A. R., Rouleau, J. L., Shi, V. C., Solomon, S. D., Swedberg, K., and Zile, M. R. (2014). Angiotensin–neprilysin inhibition versus enalapril in heart failure. *New England Journal of Medicine*, 371(11):993–1004.

Miller, R. G. (1974). The jackknife–a review. *Biometrika*, 61(1):1–15.

Montalban, X., Hauser, S. L., Kappos, L., Arnold, D. L., Bar-Or, A., Comi, G., de Seze, J., Giovannoni, G., Hartung, H.-P., Hemmer, B., Lublin, F., Rammohan, K. W., Selmaj, K., Traboulsee, A., Sauter, A., Masterman, D., Fontoura, P., Belachew, S., Garren, H., Mairon, N., Chin, P., and Wolinsky, J. S. (2017). Ocrelizumab versus placebo in primary progressive multiple sclerosis. *New England Journal of Medicine*, 376(3):209–220.

Montori, V. and Guyatt, G. (2001). Intention-to-treat principle. *CMAJ*, 165(10):1339–1341.

Moyé, L. A. (2003). *Multiple Analyses in Clinical Trials: Fundamentals for Investigators*. Springer New York, New York, NY.

Murray, S. and Cole, B. (2000). Variance and sample size calculations in quality-of-life-adjusted survival analysis (q-twist). *Biometrics*, 56(1):173–182.

Mütze, T. and Friede, T. (2021). *Sample size re-estimation*. In: Handbook of Statistical Methods for Randomized Controlled Trials, Chapman and Hall/CRC.

Nevo, D. and Gorfine, M. (2021). Causal inference for semi-competing risks data. *Biostatistics*, 23(4):1115–1132.

Nord, E., Daniels, N., and Kamlet, M. (2009). QALYs: Some challenges. *Value in Health*, 12:S10–S15.

North American Symptomatic Carotid Endarterectomy Trial Collaborators (1991). Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis. *New England Journal of Medicine*, 325(7):445–453.

North American Symptomatic Carotid Endarterectomy Trial Steering Committee (1991). North American Symptomatic Carotid Endarterectomy Trial: methods, patient characteristics and progress. *Stroke*, 22(6):711–720.

Oakes, D. (2016). On the win-ratio statistic in clinical trials with multiple types of event. *Biometrika*, 103(3):742–745.

O'Connor, C. M., Whellan, D. J., Lee, K. L., Keteyian, S. J., Cooper, L. S., Ellis, S. J., Leifer, E. S., Kraus, W. E., Kitzman, D. W., Blumenthal, J. A., Rendall, D. S., Miller, N. H., Fleg, J. L., Schulman, K. A., McKelvie, R. S., Zannad, F., and Piña, I. L. (2009). Efficacy and safety of exercise training in patients with chronic heart failure: HF-ACTION randomized controlled trial. *JAMA*, 301(14):1439–1450.

Ogdie, A. and Coates, L. (2017). The changing face of clinical trials in psoriatic arthritis. *Current Rheumatology Reports*, 19(4):21–22.

Overgaard, M., Andersen, P. K., and Parner, E. T. (2023). Pseudo-observations in a multistate setting. *The Stata Journal*, 23(2):491–517.

Parkin, D. and Devlin, N. (2006). Is there a case for using visual analogue scale valuations in cost-utility analysis? *Health Economics*, 15:653–664.

Parner, E. T. and Overgaard, M. (2023). Win-loss parameters for right-censored event data, with application to recurrent events. *Statistics in Medicine*, 42(30):5723–5735.

Parner, E. T. and Overgaard, M. (2024). Estimation of win, loss probabilities and win ratio based on right-censored event data. *Manuscript*, 0(0):0–0.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, 135(2):185–207.

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J., and Smith, P. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *British Journal of Cancer*, 34(6):585–612.

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J., and Smith, P. G. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *British Journal of Cancer*, 35(1):1–39.

Pocock, S. J., Ariti, C. A., Collier, T. J., and 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.

Pocock, S. J. and Collier, T. J. (2019). Statistical appraisal of 6 recent clinical trials in cardiology: JACC state-of-the-art review. *Journal of the American College of Cardiology*, 73(21):2740–2755.

Pocock, S. J., Ferreira, J. P., Collier, T. J., Angermann, C. E., Biegus, J., Collins, S. P., Kosiborod, M., Nassif, M. E., Ponikowski, P., Psotka, M. A., Teerlink, J. R., Tromp, J., Gregson, J., Blatchford, J. P., Zeller, C., and Voors, A. A. (2023). The win ratio method in heart failure trials: lessons learnt from EMPULSE. *European Journal of Heart Failure*, 25(5):632–641.

Powles, T., Durán, I., van der Heijden, M., Loriot, Y., Vogelzang, N., De Giorgi, U., Oudard, S., Retz, M., Castellano, D., Bamias, A., Fléchon, A., Gravis, G., Hussain, S., Takano, T., Leng, N., Kadel, E., Banchereau, R., Hegde, P., Mariathasan, S., Cui, N., Shen, X., Derleth, C., Green, M., and Ravaud, A. (2018). Atezolizumab versus chemotherapy in patients with platinum-treated locally advanced or metastatic urothelial carcinoma (IMvigor211): a multicentre, open-label, phase 3 randomised controlled trial. *Lancet*, 391(10122):748–757.

Poythress, J. C., Lee, M. Y., and Young, J. (2020). Planning and analyzing clinical trials with competing risks: recommendations for choosing appropriate statistical methodology. *Pharmaceutical Statistics*, 19(1):4–21.

Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, 34(4):541–554.

Putter, H., Fiocco, M., and Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389–2430.

Putter, H., Schumacher, M., and van Houwelingen, H. C. (2020). On the relation between the cause-specific hazard and the subdistribution rate for competing risks data: The Fine–Gray model revisited. *Biometrical Journal*, 62(3):790–807.

Qu, Y., Shurzinske, L., and Sethuraman, S. (2021). Defining estimands using a mix of strategies to handle intercurrent events in clinical trials. *Pharmaceutical Statistics*, 20(2):314–323.

Revicki, D. A. and Ehreth, J. L. (1997). Health-related quality-of-life assessment and planning for the pharmaceutical industry. *Clinical Therapeutics*, 19(5):1101–1115.

Revicki, D. A., Feeny, D., Hunt, T. L., and Cole, B. F. (2006). Analyzing oncology clinical trial data using the Q-TWiST method: clinical importance and sources for health state preference data. *Quality of Life Research*, 15(3):411–423.

Richardson, J., Iezzi, A., and Khan, M. (2015). Why do multi-attribute utility instruments produce different utilities: the relative importance of the descriptive systems, scale and "micro-utility" effects. *Quality Life Research*, 24:2045–2053.

Rittmeyer, A., Barlesi, F., Waterkamp, D., Park, K., Ciardiello, F., von Pawel, J., Gadgeel, S., T., H., Kowalski, D., Dols, M., Cortinovis, D., Leach, J., Polikoff, J., Barrios, C., Kabbinavar, F., Frontera, O., De Marinis, F., Turna, H., Lee, J., Ballinger, M., Kowanetz, M., He, P., Chen, D., Sandler, A., Gandara, D., and for the OAK Study Group (2017). Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. *The Lancet*, 389(10066):255–265.

Rodríguez-Míguez, E., Pinto-Prades, J. L., and Mosquera-Nogueira, J. (2019). Eliciting health state utilities using paired-gamble methods: the role of the starting point. *Value in Health*, 22(4):446–452.

Rogers, J. K., Pocock, S. J., McMurray, J. J., Granger, C. B., Michelson, E. L., Östergren, J., Pfeffer, M. A., Solomon, S. D., Swedberg, K., and Yusuf, S. (2014). Analysing recurrent hospitalizations in heart failure: a review of statistical methodology, with application to charm-preserved. *European Journal of Heart Failure*, 16(1):33–40.

Rosenblum, M. and van der Laan, M. J. (2009). Using regression models to analyze randomized trials: asymptotically valid hypothesis tests despite incorrectly specified models. *Biometrics*, 65(3):937–945.

Royston, P. and Parmar, M. (2013). Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Medical Research Methodology*, 13(152).

Royston, P. and Parmar, M. (2016). Augmenting the log-rank test in the design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated. *BMC Med Res Methodol*, 16(16).

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Rubin, D. B. (2006). Causal inference through potential outcomes and principal stratification: application to studies with "censoring" due to death. *Statistical Science*, 21(3):299–309.

Rufibach, K. (2019). Treatment effect quantification for time-to-event endpoints – estimands, analysis strategies, and beyond. *Pharmaceutical Statistics*, 18(2):145–165.

Scharfstein, D. O. (2019). A constructive critique of the draft ICH E9 Addendum. *Clinical Trials*, 16(4):375–380.

Scheike, T. and Zhang, M.-J. (2007). Direct modelling of regression effects for transition probabilities in multistate models. *Scandinavian Journal of Statistics*, 34(1):17–32.

Scheike, T. H., Martinussen, T., and Ozenne, B. (2022). Efficient estimation in the Fine and Gray model. *Journal of the American Statistical Association*, 118(544):2482–2490.

Scheike, T. H. and Zhang, M.-J. (2008). Flexible competing risks regression modeling and goodness-of-fit. *Lifetime Data Analysis*, 14(4):464–483.

Scheike, T. H. and Zhang, M.-J. (2011). Analyzing competing risk data using the R timereg package. *Journal of Statistical Software*, 38(2):1–15.

Scheike, T. H., Zhang, M.-J., and Gerds, T. A. (2008). Predicting cumulative incidence probability by direct binomial regression. *Biometrika*, 95(1):205–220.

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

Shi, H., Cheng, Y., and Jeong, J.-H. (2013). Constrained parametric model for simultaneous inference of two cumulative incidence functions. *Biometrical Journal*, 55(1):82–96.

Snowden, J. M., Reavis, K. M., and Odden, M. C. (2020). Conceiving of questions before delivering analyses: relevant question formulation in reproductive and perinatal epidemiology. *Epidemiology*, 31(5):644–648.

Solomon, S. D., McMurray, J. J., Anand, I. S., Ge, J., Lam, C. S., Maggioni, A. P., Martinez, F., Packer, M., Pfeffer, M. A., Pieske, B., Redfield, M. M., Rouleau, J. L., van Veldhuisen, D. J., Zannad, F., Zile, M. R., Desai, A. S., Claggett, B., Jhund, P. S., Boytsov, S. A., Comin-Colet, J., Cleland, J., Düngen, H.-D., Goncalvesova, E., Katova, T., Kerr Saraiva, J. F., Lelonek, M., Merkely, B., Senni, M., Shah, S. J., Zhou, J., Rizkala, A. R., Gong, J., Shi, V. C., and Lefkowitz, M. P. (2019). Angiotensin–neprilysin inhibition in heart failure with preserved ejection fraction. *New England Journal of Medicine*, 381(17):1609–1620.

Solomon, S. D., Rizkala, A. R., Gong, J., Wang, W., Anand, I. S., Ge, J., Lam, C. S., Maggioni, A. P., Martinez, F., Packer, M., Pfeffer, M. A., Pieske, B., Redfield, M. M., Rouleau, J. L., Van Veldhuisen, D. J., Zannad, F., Zile, M. R., Desai, A. S., Shi, V. C., Lefkowitz, M. P., and McMurray, J. J. (2017). Angiotensin receptor neprilysin inhibition in heart failure with preserved ejection fraction: Rationale and design of the PARAGON-HF Trial. *JACC: Heart Failure*, 5(7):471–482.

Sprott, D. A. and Farewell, V. T. (1993). Randomization in experimental science. *Statistical Papers*, 34(0):89–94.

Stanworth, S. J., Estcourt, L. J., Powter, G., Kahan, B. C., Dyer, C., Choo, L., Bakrania, L., Llewelyn, C., Littlewood, T., Soutar, R., Norfolk, D., Copplestone, A., Smith, N., Kerr, P., Jones, G., Raj, K., Westerman, D. A., Szer, J., Jackson, N., Bardy, P. G., Plews, D., Lyons, S., Bielby, L., Wood, E. M., and Murphy, M. F. (2013). A no-prophylaxis platelet-transfusion strategy for hematologic cancers. *New England Journal of Medicine*, 368(19):1771–1780.

Stensrud, M. J. and Dukes, O. (2022). Translating questions to estimands in randomized clinical trials with intercurrent events. *Statistics in Medicine*, 41(16):3211–3228.

Stensrud, M. J., Young, J. G., Didelez, V., Robins, J. M., and Hernán, M. A. (2020). Separable effects for causal inference in the presence of competing events. *Journal of the American Statistical Association*, 117(537):175–183.

Ster, A. M. C., Cornelius, V., and Cro, S. (2020). Current approaches to handling rescue medication in asthma and eczema randomized controlled trials are inadequate: a systematic review. *Journal of Clinical Epidemiology*, 125:148–157.

Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika*, 73(2):363–369.

Sun, J. and Kalbfleisch, J. D. (1995). Estimation of the mean function of point processes based on panel count data. *Statistica Sinica*, 5(1):279–289.

Sun, J. and Zhao, X. (2013). *Statistical Analysis of Panel Count Data*. Springer.

Taylor, M., Chilton, S., Ronaldson, S., Metcalf, H., and Nielsen, J. S. (2017). Comparing increments in utility of health: An individual-based approach. 20(2):224–229.

Theriault, R. L., Lipton, A., Hortobagyi, G. N., Leff, R., GlückS., Stewart, J. F., Costello, S., Kennedy, I., Simeone, J., Seaman, J. J., and Knight, R. D. (1999). Pamidronate reduces skeletal morbidity in women with advanced breast cancer and lytic bone lesions: a randomized, placebo-controlled trial. *Journal of Clinical Oncology*, 17(3):846–846.

Therneau, T. M. (2024). *A Package for Survival Analysis in R*. R package version 3.6-4.

Tiwari, R., Kulasekera, K. B., and Park, C. (2006). Nonparametric tests for cause-specific hazard rates with censored data for competing risks among several groups. *Journal of Statistical Planning and Inference*, 136(5):1718–1745. Competing Risks: Theory and Applications.

Toenges, G., Mütze, T., and Jahn-Eimermacher, A. (2021). A comparison of semiparametric approaches to evaluate composite endpoints in heart failure trials. *Statistics in Medicine*, 40(26):5702–5724.

Torrance, G. W. (1986). Measurement of health state utilities for economic appraisal: A review. *Journal of Health Economics*, 5(1):1–30.

Torrance, G. W. (1987). Utility approach to measuring health-related quality of life. *Journal of Chronic Diseases*, 40(6):593–600.

Torrance, G. W., Furlong, W., and Feeny, D. (2002). Health utility estimation. *Expert Review of Pharmacoeconomics & Outcomes Research*, 2(2):99–108.

Torrance, G. W., Furlong, W., Feeny, D., and Boyle, M. (1995). Multi-attribute preference functions. health utilities index. *PharmacoEconomics*, 7(6):503–520.

Troendle, J. F., Leifer, E. S., Yang, S., Jeffries, N., Kim, D.-Y., Joo, J., and O'Connor, C. M. (2024). Use of win time for ordered composite endpoints in clinical trials. *Statistics in Medicine*, 43(10):1920–1932.

Tu, J. V., Donovan, L. R., Lee, D. S., Wang, J. T., Austin, P. C., Alter, D. A., and Ko, D. T. (2009). Effectiveness of public report cards for improving the quality of cardiac care: The EFFECT study: A randomized trial. *JAMA*, 302(21):2330–2337.

Vaduganathan, M., Claggett, B. L., Lam, C. S., Pitt, B., Senni, M., Shah, S. J., Voors, A. A., Zannad, F., Desai, A. S., Jhund, P. S., Viswanathan, P., Bomfim Wirtz, A., Schloemer, P., Lay-Flurrie, J., McMurray, J. J., and Solomon, S. D. (2024). Finerenone in patients with heart failure with mildly reduced or preserved ejection fraction: Rationale and design of the FINEARTS-HF trial. *European Journal of Heart Failure*, 26(6):1324–1333.

van der Meer, P. F., Ypma, P. F., van Geloven, N., van Hilten, J. A., van Wordragen-Vlaswinkel, R. J., Eissen, O., Zwaginga, J. J., Trus, M., Beckers, E. A. M., te Boekhorst, P., Tinmouth, A., Lin, Y., Hsia, C., Lee, D., Norris, P. J., Goodrich, R. P., Brand, A., Hervig, T., Heddle, N. M., van der Bom, J. G., and Kerkhoffs, J.-L. H. (2018). Hemostatic efficacy of pathogen-inactivated vs untreated platelets: a randomized controlled trial. *Blood*, 132(2):223–231.

Versteegh, M. M. and Brouwer, W. (2016). Patient and general public preferences for health states: A call to reconsider current guidelines. *Social Science and Medicine*, 165:66–74.

Wang, T., Li, Y., and Qu, Y. (2024). Restricted time win ratio: from estimands to estimation. *Statistics in Biopharmaceutical Research*, 0(0):1–19.

Watkins, C., Huang, X., Latimer, N., Tang, Y., and Wright, E. (2013). Adjusting overall survival for treatment switches: commonly used methods and practical application. *Pharmaceutical Statistics*, 12(6):348–357.

Webert, K. E., Arnold, D. M., Lui, Y., Carruthers, J., Arnold, E., and Heddle, N. M. (2012). A new tool to assess bleeding severity in patients with chemotherapy-induced thrombocytopenia (CME). *Transfusion*, 52(11):2466–2474.

Wei, J., Mütze, T., Jahn-Eimermacher, A., and Roger, J. (2021). Properties of two while-alive estimands for recurrent events and their potential estimators. *Statistics in Biopharmaceutical Research*, 15(2):257–267.

Wenger, N. K., Mattson, M. E., Furberg, C. D., and Elinson, J. (1984). Assessment of quality of life in clinical trials of cardiovascular therapies. *The American Journal of Cardiology*, 54(7):908–913.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

Williamson, P. R., Kolamunnage-Dona, R., and Tudur Smith, C. (2007). The influence of competing-risks setting on the choice of hypothesis test for treatment effect. *Biostatistics*, 8(4):689–694.

Wong, W. L., Haines, A., Wong, J., Hamadeh, A., and Krahn, M. (2023a). A province-by-province cost-effectiveness analysis and budget impact analysis of one-time birth cohort screening of hepatitis C virus (HCV) infection in Canada. *Scientific Reports*, 13(1):0–0.

Wong, W. W. L., Lee, L., Walker, S., Lee, C., Patel, T., Hillier, L. M., Costa, A. P., and Sinha, S. K. (2023b). Cost-utility analysis of a multispecialty interprofessional team dementia care model in Ontario, Canada. 13(4).

Wu, L. and Cook, R. (2012). Misspecification of Cox regression models with composite endpoints. *Statistics in Medicine*, 31(28):3545–3562.

Xie, S., Wu, J., He, X., Chen, G., and Brazier, J. E. (2020). Do discrete choice experiments approaches perform better than time trade-off in eliciting health state utilities? Evidence from SF-6Dv2 in China. *Value in Health*, 23(10):1391–1399.

Xu, Y., Scharfstein, D., Müller, P., and Daniels, M. (2020). A Bayesian nonparametric approach for evaluating the causal effect of treatment in randomized trials with semi-competing risks. *Biostatistics*, 23(1):34–49.

Young, J. G., Stensrud, M. J., Tchetgen Tchetgen, E. J., and Hernán, M. A. (2020). A causal framework for classical statistical estimands in failure-time settings with competing events. *Statistics in Medicine*, 39(8):1199–1236.

Ypma, P. F., van der Meer, P. F., Heddle, N. M., van Hilten, J. A., Stijnen, T., Middelburg, R. A., Hervig, T., van der Bom, J. G., Brand, A., and Kerkhoffs, J.-L. H. (2016). A study protocol for a randomized controlled trial evaluating clinical effects of platelet transfusion products: the pathogen reduction evaluation and predictive analytical rating score (PREPAReS) trial. *British Medical Journal*, 6(1).

Yung, G. and Liu, Y. (2020). Sample size and power for the weighted log-rank test and Kaplan-Meier based tests with allowance for nonproportional hazards. *Biometrics*, 76(3):939–950.

Zhang, D. and Quan, H. (2009). Power and sample size calculation for log-rank test with a time lag in treatment effect. *Statistics in Medicine*, 28(5):864–879.

Zhang, M.-J. and Fine, J. (2008). Summarizing differences in cumulative incidence functions. *Statistics in Medicine*, 27(24):4939–4949.

Zhou, B., Fine, J., and Laird, G. (2013). Goodness-of-fit test for proportional subdistribution hazards model. *Statistics in Medicine*, 32(22):3804–3811.

APPENDICES

Appendix A

Appendix to Chapter 2

A.1 Adequacy of the cloglog transformed CIF model under time-inhomogeneous intensities

Here we assume a competing risks process with Weibull intensities $\lambda_{0k}(t|X) = \lambda_k(t)\exp(\gamma_k X) = \kappa_k\lambda_k(\lambda_k t)^{\kappa_k-1}\exp(\gamma_k X)$, $k = 1, 2$, and investigate how time-inhomogeneous intensities may influence the adequacy of the cloglog model (3.2.3). If $\kappa_1 = \kappa_2 = 1$, then $\lambda_{0k}(t|X) = \lambda_k\exp(\gamma_k X)$ and we retrieve the time-homogeneous setting of Section 2.3.2. For given values of $\kappa_1, \kappa_2 > 0$, we chose λ_1 and λ_2 such that $P(T \leq 1|X = 0) = 0.6$ and $P(T_1 < T_2|T \leq 1, X = 0) = 0.4, 0.6$ or 0.8. For illustration, Figure A.1 contains plots of $\lambda_{01}(t|X = 0), \lambda_{02}(t|X = 0)$ and $F_1(t|X = 0)$ when $\kappa_1 = 0.8, 1, 1.2, 1.4$ and $\kappa_2 = 1$. When $\kappa_1 > 1$ (or < 1), the $0-1$ intensity is increasing (or decreasing) over time; see left panel of Figure A.1.

We next assessed the adequacy of the cloglog model (3.2.3) by calculating $g(F_1(t|X = 1)) - g(F_1(t|X = 0))$ based on the true values of $F_1(t|X)$ for $t \in (0, 1)$ and $g(u) = \log(-\log(1-u))$; results are displayed in Figure A.2. Recall from Section 2.3.2 that departures from a horizontal line are indicative of model misspecification. For a fixed value of $\exp(\gamma_2)$, the transformation model improves with decreasing κ_1 and is, thus, a better approximation when $\kappa_1 = 1$ as compared to when $\kappa_1 > 1$; the approximation is even better when the intensities in the true process are decreasing over time (i.e. $\kappa_1 < 1$).

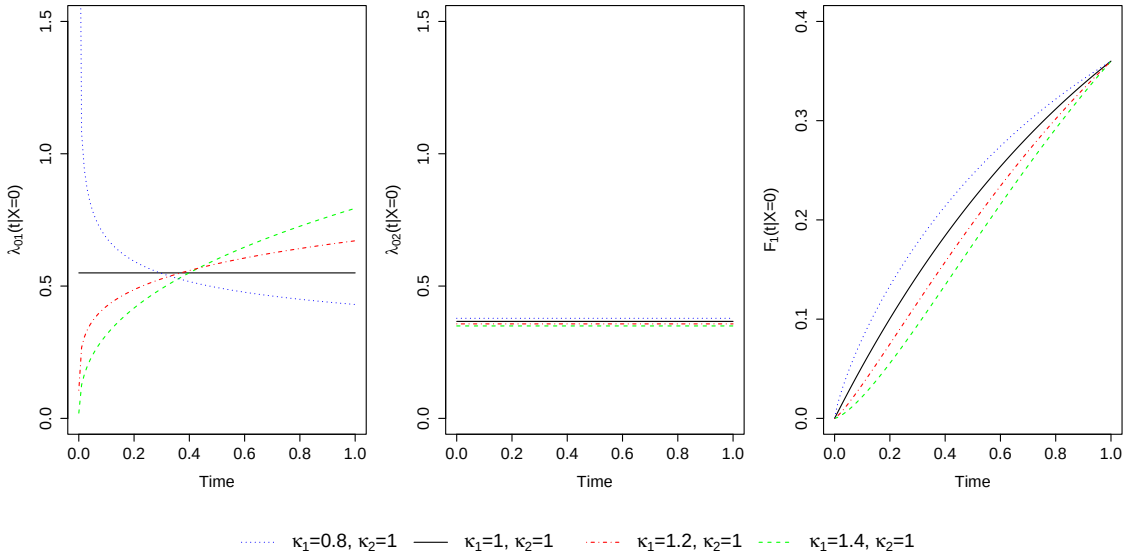


Figure A.1: Plots of $\lambda_{01}(t|X=0)$, $\lambda_{02}(t|X=0)$ and $F_1(t|X=0)$ for different values of κ_1 when the true competing risks process has baseline intensities of Weibull form $\lambda_k(t) = \kappa_k \lambda_k (\lambda_k t)^{\kappa_k - 1}$; $\kappa_2 = 1, \tau = 1, P(T \leq 1|X=0) = 0.6, P(T_1 < T_2 | T \leq 1, X = 0) = 0.6$.

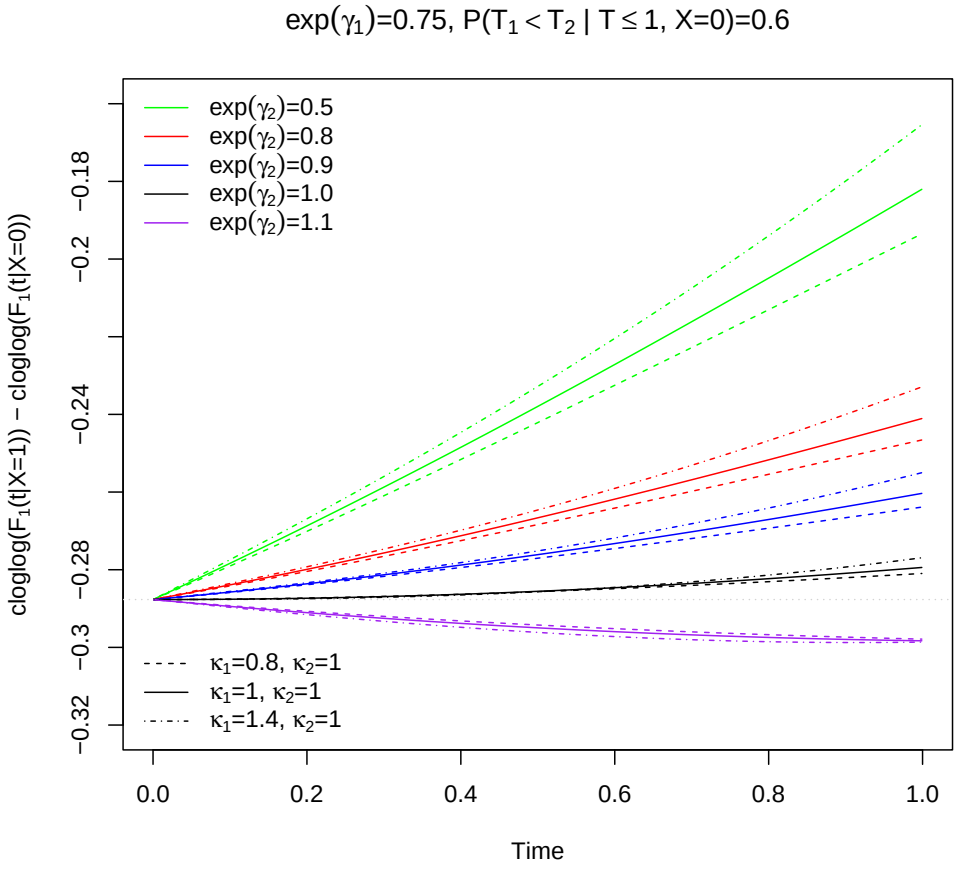


Figure A.2: Plots of $g(F_1(t|X=1)) - g(F_1(t|X=0))$ for $t \in (0,1)$ and different values of κ_1 and $\exp(\gamma_2)$ under the cloglog transformation model; $g(u) = \log(-\log(1-u)), \kappa_2 = 1, \tau = 1, P(T \leq 1|X=0) = 0.6, \gamma_1 = \log(0.75) = -0.288$.

Appendix B

Appendix to Chapter 3

B.1 Estimation of the censoring distribution and large sample results

Let $V = (W, X) = ((W_1, W_2, \ldots, W_{p-1}), X)'$ be a $p \times 1$ vector of observed baseline covariates that may be associated with the random censoring distribution so that $G(t|V) = P(C_r > t|V)$. We define $N_r(t) = \mathbb{1}(C_r \leq t)$, and let $\{N_r(t), t > 0\}$ denote the associated counting process for random loss to follow-up. If $\{Z(u), u > 0\} \perp\!\!\!\perp C_r \mid V$, then the censoring intensity is given by

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta N_r(t) = 1 \mid C_r \geq t, V)}{\Delta t} = \lim_{\Delta t \downarrow 0} \frac{P(\Delta N_r(t) = 1 \mid C_r \geq t, T \geq t, V)}{\Delta t} = \lambda_r(t|V) \ ,$$

where $\Delta N_r(t) = N_r(t + \Delta t^-) - N_r(t^-)$ and $\lim_{\Delta t \downarrow 0} \Delta N_r(t) = dN_r(t)$. We further let the cumulative censoring intensity denote by $\Lambda_r(t|V) = \int_0^t d\Lambda_r(u|V)$ for $d\Lambda_r(u|V) = \lambda_r(u|V)du$. We consider now a sample of n independent individuals indexed by $i = 1, 2, \ldots, n$, and derive the limiting behavior of Kaplan-Meier (KM) and Cox estimators of the censoring survivor function under different true censoring models.

B.1.1 Limiting behavior of the KM estimator under different true censoring mechanisms

1. If C_r does truly not depend on covariates V, the censoring survivor function $G(t) = P(C_r > t)$ can be estimated by the Kaplan-Meier method where

$$\widehat{G}(t) = \prod_{u<t}\big(1 - d\widehat{\Lambda}_r(u)\big) = \prod_{u<t}\left(1 - \frac{\sum_{i=1}^n Y_i(u)dN_{ri}(u)}{\sum_{i=1}^n Y_i(u)}\right), \tag{B.1.1}$$

with $C_i = \min(C_{ri}, \tau), Y_i(t) = \mathbb{1}(\min(T_i, C_i) \geq t)$, and the product taken over all unique random censoring times $u < t$. The Nelson-Aalen type estimate $d\widehat{\Lambda}_r(u)$ in (B.1.1) is obtained as the solution to $\sum_{i=1}^n Y_i(u)\big(dN_{ri}(u) - d\Lambda_r(u)\big) = 0$, and is consistent for the true $d\Lambda_r(u) = \mathbb{E}(dN_r(u)|C_r \geq u)$ when the censoring process is Markov. For more general processes $d\widehat{\Lambda}_r(u)$ is consistent for the marginal censoring rate $d\Lambda_r^\star(u) = \mathbb{E}(dN_r(u)|C_r \geq u)$, the solution to $\mathbb{E}\big(\sum_{i=1}^n Y_i(u)(dN_{ri}(u) - d\Lambda_r(u))\big) = 0$ (Cook and Lawless, 2018). Note that $d\Lambda_r^\star(u) = d\Lambda_r(u)$ for Markov processes. As a consequence,

$$\widehat{G}(t) \longrightarrow G(t) = P(C_r > t; d\Lambda_r(\cdot)) = \prod_{u<t}\big(1 - d\Lambda_r(u)\big), \text{ or} \tag{B.1.2}$$

$$\widehat{G}(t) \longrightarrow G^\star(t) = P(C_r > t; d\Lambda_r^\star(\cdot)) = \prod_{u<t}\big(1 - d\Lambda_r^\star(u)\big). \tag{B.1.3}$$

2. If the true censoring intensity $\lambda_r(t|V_i)$ depends on $V_i = (W_i', X_i)'$ but KM estimation stratified by X_i only is used for estimation of the censoring survivor function, then

$$\widehat{G}(t|X_i) \longrightarrow G^\star(t|X_i) = P(C_r > t|X_i; d\Lambda_r^\star(\cdot|X)), \tag{B.1.4}$$

where

$$d\Lambda_r^\star(t|X) = \frac{\mathbb{E}_W\big(\lambda_r(t|X,W)G(t|X,W)S(t|X,W)\big)}{\mathbb{E}_W\big(G(t|X,W)S(t|X,W)\big)},$$

can be found as the solution to $\mathbb{E}\big(Y(u)\big(dN_r(u) - d\Lambda_r(u)\big)|\ X\big) = 0$ for $X = 0, 1$; with expectations taken with respect to the true censoring, competing risks and covariate processes.

B.1.2 Limiting behavior of the Cox estimator under misspecification

1. The conditional censoring survivor function $G(t|V_i) = P(C_{ri} > t|V_i)$ can be estimated by

$$\widehat{G}(t|V_i) = \prod_{u<t}\Big(1 - d\widehat{\Lambda}_r(u|V_i)\Big) ,$$

where the censoring intensity may be modeled based on the Cox proportional hazards specification. If $S^{(l)}(u,\gamma_r) = \sum_{i=1}^n Y_i(u)V_i^{\otimes l}\exp(\gamma_r' V_i)$ and $s^{(l)}(u,\gamma_r) = \mathbb{E}(S^{(l)}(u,\gamma_r))$ for $l = 0, 1$, the solution to

$$U(\gamma_r) = \sum_{i=1}^n \int_0^\infty Y_i(u)\Big(V_i - \frac{S^{(1)}(u,\gamma_r)}{S^{(0)}(u,\gamma_r)}\Big)dN_{ri}(u) = 0$$

gives the Cox partial likelihood estimator $\widehat{\gamma}_r$. The Breslow-type estimator of the cumulative baseline hazard is $\widehat{\Lambda}_{r0}(t) = \int_0^t d\widehat{\Lambda}_{r0}(u)$, where

$$d\widehat{\Lambda}_{r0}(u) = \sum_{i=1}^n Y_i(u)dN_{ri}(u) \Big/ \sum_{i=1}^n Y_i(u)\exp(\widehat{\gamma}_r' V_i) .$$

Under conditionally independent censoring and a correctly specified censoring model, $\widehat{\gamma}_r$ is consistent for the true γ_r and $d\widehat{\Lambda}_{r0}(t)$ for the true $d\Lambda_{r0}(t)$, in which case

$$\widehat{G}(t|V_i) \longrightarrow G(t|V_i) = P(C_{ri} > t \mid V_i; d\Lambda_{r0}(\cdot), \gamma_r) .$$

2. More generally, under some type of model misspecification, $\widehat{\gamma}_r$ is consistent for the limiting value $\gamma_r^\star$, the solution to

$$\mathbb{E}\big(U(\gamma_r)\big) = \int_0^\infty \left\{\mathbb{E}\Big(\sum_{i=1}^n Y_i(u)V_i dN_{ri}(u)\Big) - \frac{s^{(1)}(u,\gamma_r)}{s^{(0)}(u,\gamma_r)}\mathbb{E}\Big(\sum_{i=1}^n Y_i(u)dN_{ri}(u)\Big)\right\} = 0 ,$$

where expectations are taken with respect to the true censoring, covariate and competing risks processes (Struthers and Kalbfleisch, 1986; Lin and Wei, 1989). If the probability limit of $\widehat{\Lambda}_{r0}(u)$ is denoted by $d\Lambda_{r0}^\star(u)$, we find that

$$\widehat{G}(t|V_i) \longrightarrow G^\star(t|V_i) = P(C_{ri} > t \mid V_i; d\Lambda_{r0}^\star(\cdot), \gamma_r^\star) . \tag{B.1.5}$$

B.2 Limiting values of FG and DB treatment effect estimators under intensity-based processes

B.2.1 Supplementary results for the numerical investigation of Section 3.3.2

In this section we report on additional results about the estimands arising from Fine-Gray (FG) and direct binomial (DB) regression when the true process is governed by time-homogeneous intensities of proportional form as in Section 3.3.2. Figure B.1 illustrates the dependency of $\exp(\beta^\star_{FG})$ and $\exp(\beta^\star_{DB})$ on $\exp(\gamma_2)$ for different values of $\exp(\gamma_1)$ and $P(T_1 < T_2 | T \leq 1, X = 0)$. Figure B.2 depicts the limiting values $\exp(\beta^\star_{FG})$ and $\exp(\beta^\star_{DB})$ as a function of $P(T_1 < T_2 | T \leq 1, X = 0)$ for different values of $\exp(\gamma_1)$ and $\exp(\gamma_2)$; a discussion of results can be found in Section 3.3.2.

The plots in Section 3.3.2 and Figures B.1-B.2 give insights into how the limiting values $\beta^\star_{FG}$ and $\beta^\star_{DB}$ of the FG and DB estimators vary with the parameters of the intensity-based process, and in particular how they compare with $\exp(\gamma_1)$, the treatment effect on the main $0-1$ intensity. Another question of interest concerns the effect of model misspecification on the estimation of $F_1(t|X)$. Specifically, we are interested in how $F_1(t|X)$ (c.f. (3.2.1)) compares with $F^\star_{1,FG}(t|X)$ and $F^\star_{1,DB}(s_r|X)$ as given by (3.3.9) and (3.3.16), respectively. Plots of $F_1(t|X)$ versus $F^\star_{1,\cdot}(t|X)$ are depicted in Figure B.3 for $\exp(\gamma_1) = 1$ and in Figure B.4 for $\exp(\gamma_1) = 0.75$ when $P(T_1 < T_2 | T \leq 1, X = 0) = 0.6$ (upper panels) and 0.2 (lower panels). We find that $F^\star_{1,\cdot}(t|X)$ under either estimation procedure is closer to the truth $F_1(t|X)$ for values of $\exp(\gamma_2)$ close to 1 and as $P(T_1 < T_2 | T \leq 1, X = 0)$ increases.

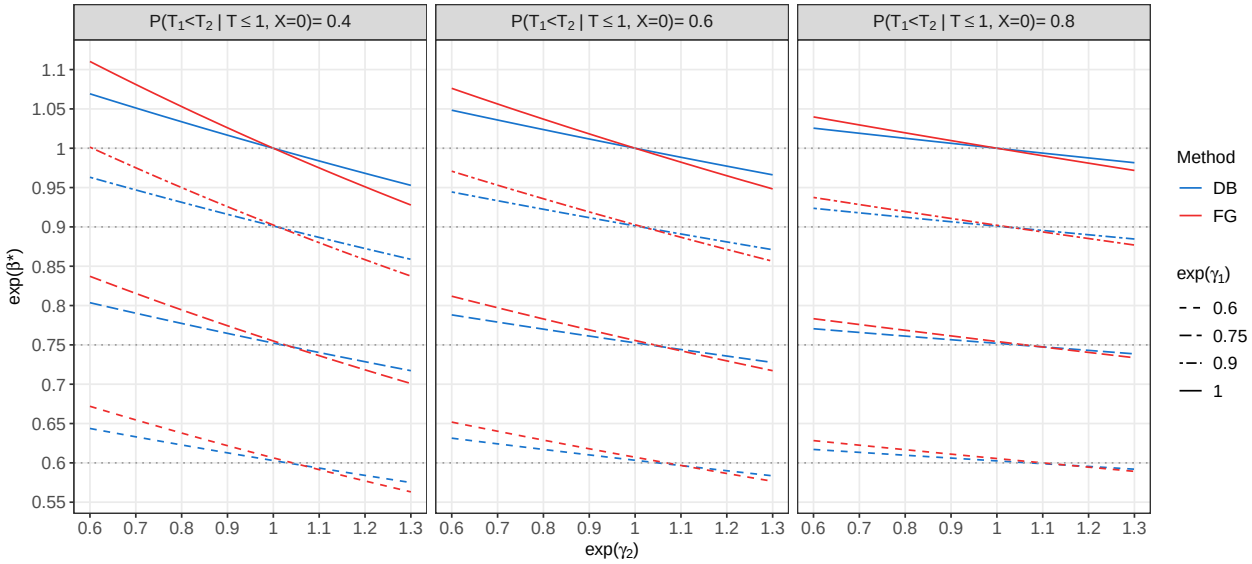


Figure B.1: Limiting values $\exp(\beta^{\star}_{FG})$ and $\exp(\beta^{\star}_{DB})$ as a function of $\exp(\gamma_2)$ for different values of $\exp(\gamma_1)$ and $P(T_1 < T_2 | T \leq 1, X = 0)$ when the true process has intensities $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X)$; $\tau = 1, P(T \leq 1|X = 0) = 0.6$, DB estimation is based on $R = 6$ equi-spaced time points in $(0, \tau), G^{\star} = G$.

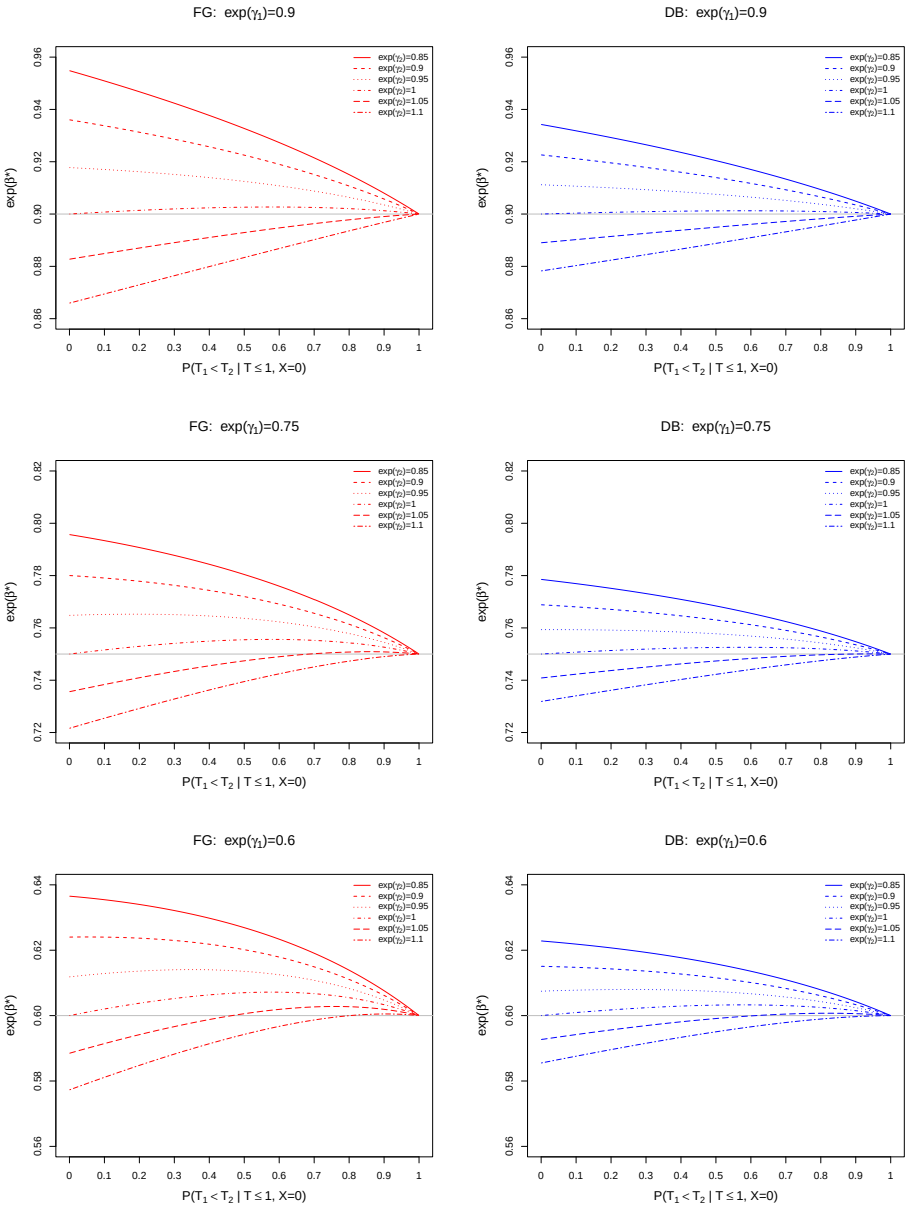


Figure B.2: Limiting values $\exp(\beta^{\star}_{FG})$ and $\exp(\beta^{\star}_{DB})$ as a function of $P(T_1 < T_2 | T \leq 1, X = 0)$ for different values of $\exp(\gamma_2)$ and $\exp(\gamma_1)$ when the true process has intensities $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X)$; $\tau = 1, P(T \leq 1|X = 0) = 0.6$, DB estimation is based on $R = 6$ equi-spaced time points in $(0, \tau), G^{\star} = G$.

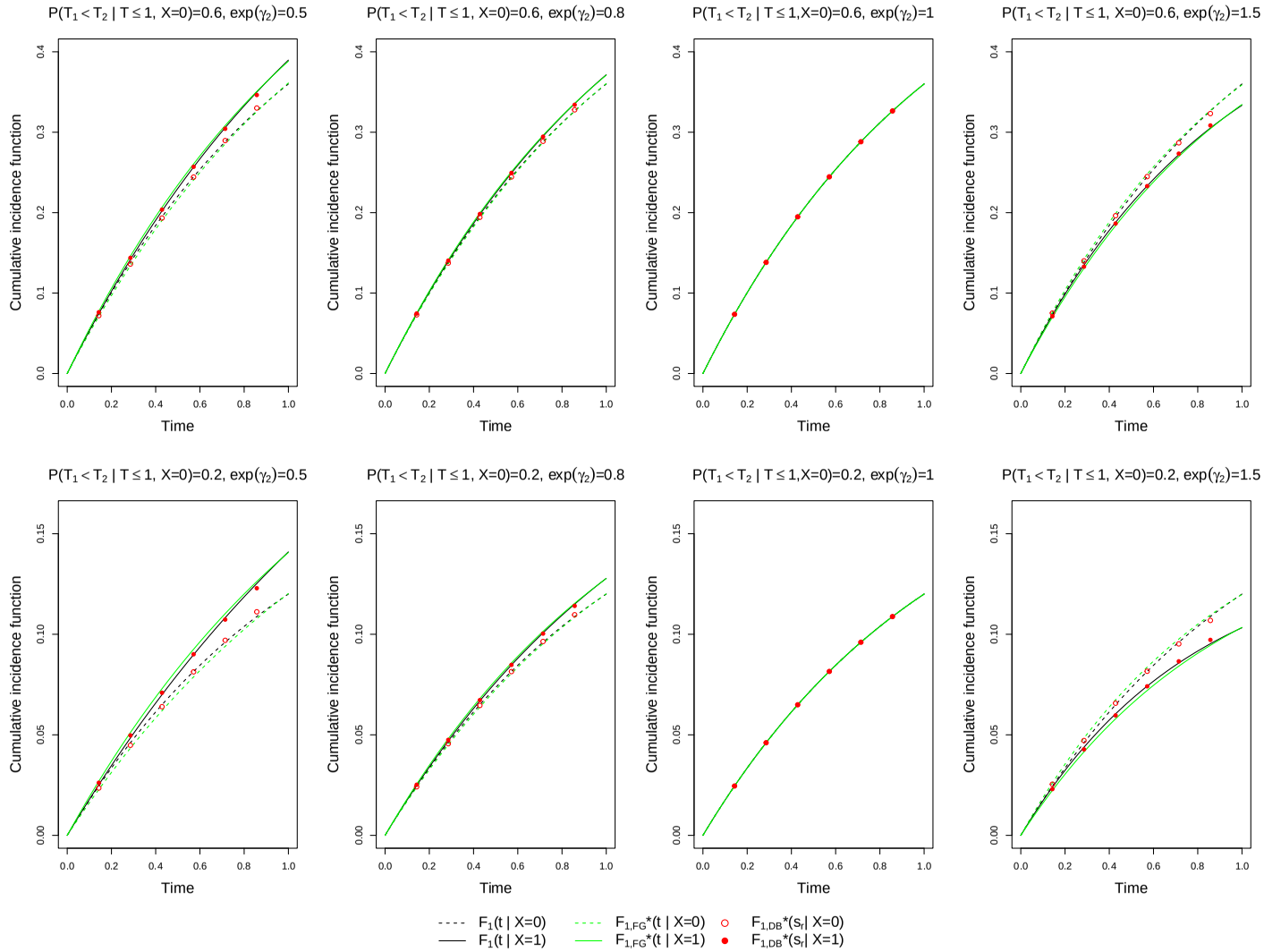


Figure B.3: Comparisons of $F_1(t|X) = \int_0^t \lambda_{01}(u|X)S(u^-|X)du, F^\star_{1,FG}(t|X) = h(\alpha^\star_{FG}(t) + \beta^\star_{FG}X)$ and $F^\star_{1,DB}(s_r|X) = h(\alpha^\star_{rDB} + \beta^\star_{DB}X), X = 0, 1$ when the true process has intensities $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X)$ for $\exp(\gamma_1) = 1, \exp(\gamma_2) = 0.5, 0.8, 1, 1.5$ and $P(T_1 < T_2|T \leq 1, X = 0) = 0.6$ (top row) and 0.2 (bottom row); $\tau = 1, P(T \leq 1|X = 0) = 0.6$, DB procedure is based on $R = 6$ equi-spaced time points in $(0, \tau)$, $G^\star = G$.

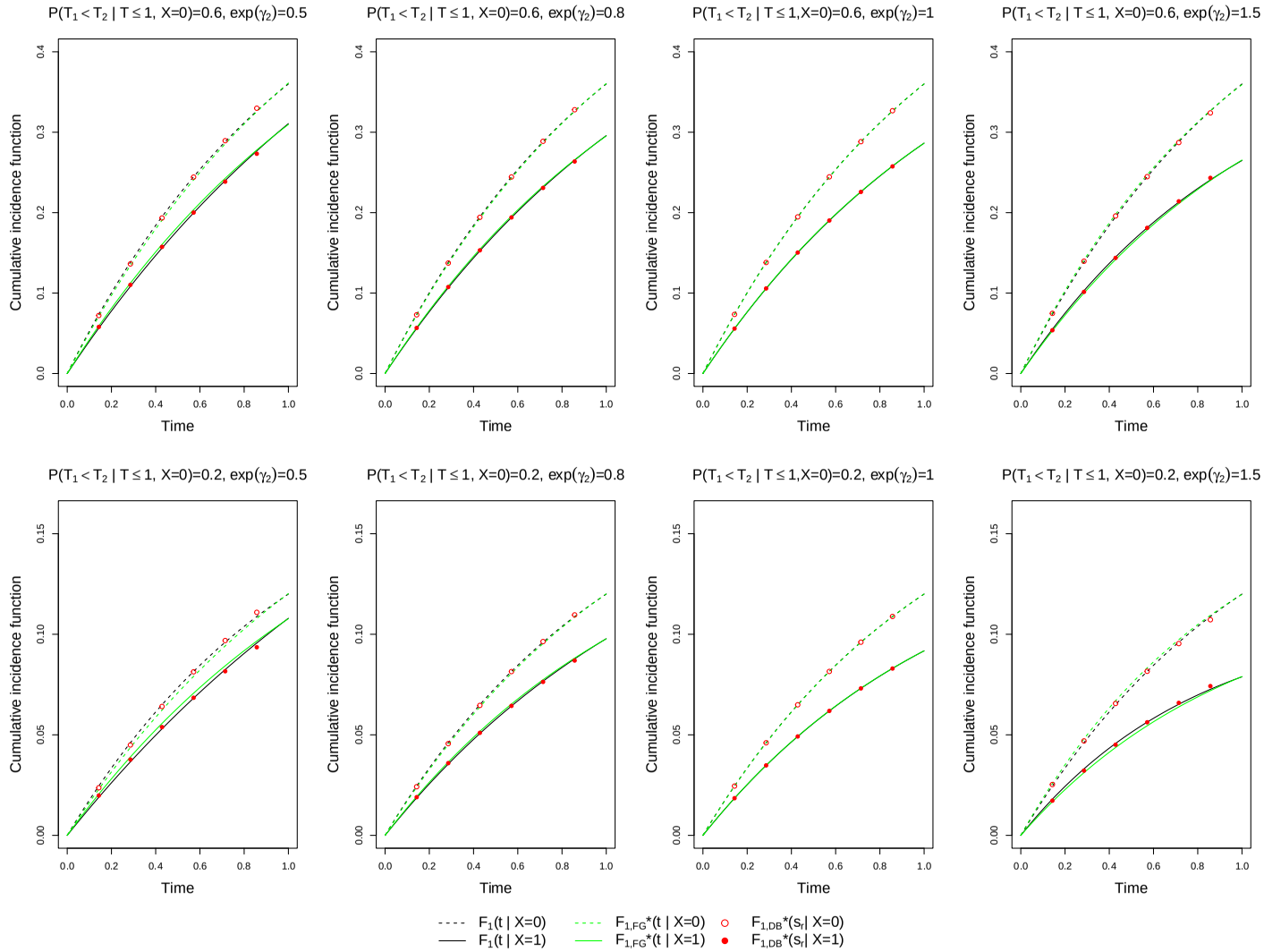
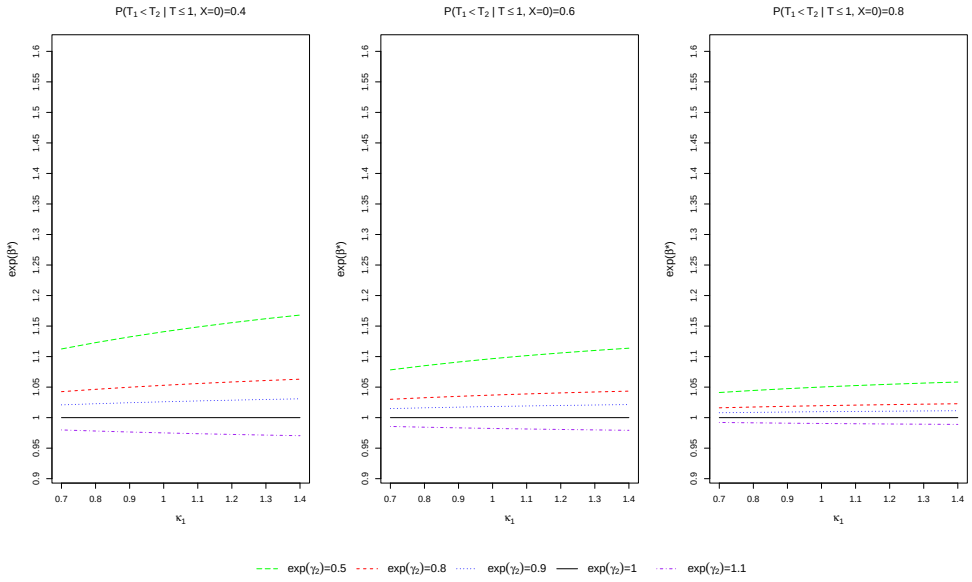


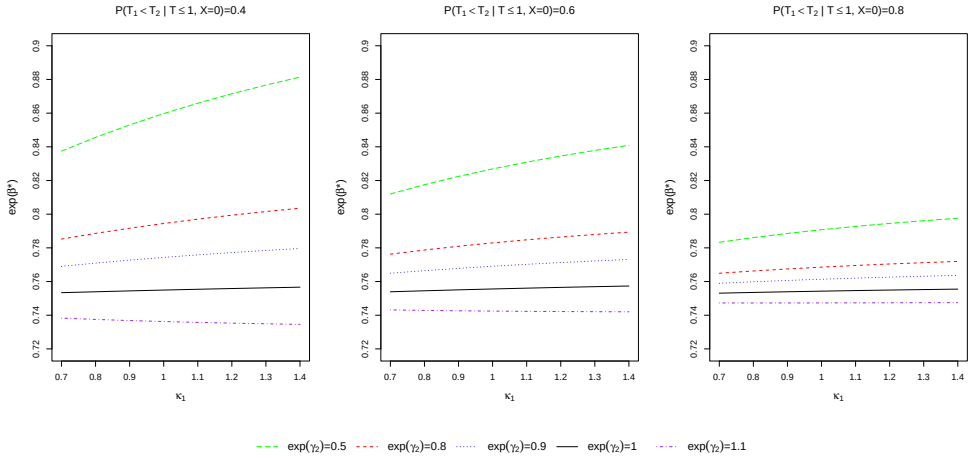
Figure B.4: Comparisons of $F_1(t|X) = \int_0^t \lambda_{01}(u|X)S(u^-|X)du, F^\star_{1,FG}(t|X) = h(\alpha^\star_{FG}(t) + \beta^\star_{FG}X)$ and $F^\star_{1,DB}(s_r|X) = h(\alpha^\star_{rDB} + \beta^\star_{DB}X), X = 0, 1$ when the true process has intensities $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X)$ for $\exp(\gamma_1) = 0.75, \exp(\gamma_2) = 0.5, 0.8, 1, 1.5$ and $P(T_1 < T_2|T \leq 1, X = 0) = 0.6$ (top row) and 0.2 (bottom row); $\tau = 1, P(T \leq 1|X = 0) = 0.6$, DB procedure is based on $R = 6$ equi-spaced time points in $(0, \tau)$, $G^\star = G$.

B.2.2 Limiting values under time-inhomogeneous intensities

Here we assume a competing risks process with Weibull intensities $\lambda_{0k}(t|X) = \lambda_k(t)\exp(\gamma_k X) = \kappa_k\lambda_k(\lambda_k t)^{\kappa_k-1}\exp(\gamma_k X)$, $k = 1, 2$, and investigate how time-inhomogeneous intensities may influence the limiting values of estimators arising from CIF-based FG and DB analyses. Adequacy of the cloglog transformation model (3.2.3) when the true process has time-inhomogeneous Weibull intensities was investigated in Appendix A.1. The limiting values $\beta^\star_{FG}$ were derived by solving (3.3.8) and $\beta^\star_{DB}$ by solving (3.3.14)-(3.3.15), respectively. Figures B.5 and B.6 display plots of $\exp(\beta^\star_{FG})$ and $\exp(\beta^\star_{DB})$ as a function of κ_1 when $\kappa_2 = 1$ and the adopted censoring model is correctly specified; this dependency is illustrated for different values of $\exp(\gamma_1), \exp(\gamma_2)$ and the proportion of type 1 events by the administrative censoring time $\tau = 1$ in the control group, here through $P(T_1 < T_2|T \leq 1, X = 0)$. Figures B.5(a) and B.6(a) show that $\beta^\star_{FG}$ and $\beta^\star_{DB}$ are unaffected by κ_1 if $\exp(\gamma_1) = \exp(\gamma_2) = 1$; see black lines. For all other scenarios, the dependency of $\beta^\star_{FG}$ and $\beta^\star_{DB}$ on κ_1 becomes stronger as the magnitude of γ_2 increases and as $P(T_1 < T_2|T \leq 1, X = 0)$ decreases. Interestingly, the DB approach seems more sensitive to time-varying intensities than the FG approach when $\exp(\gamma_1) = 1$, but less sensitive when $\exp(\gamma_1) = 0.75$.

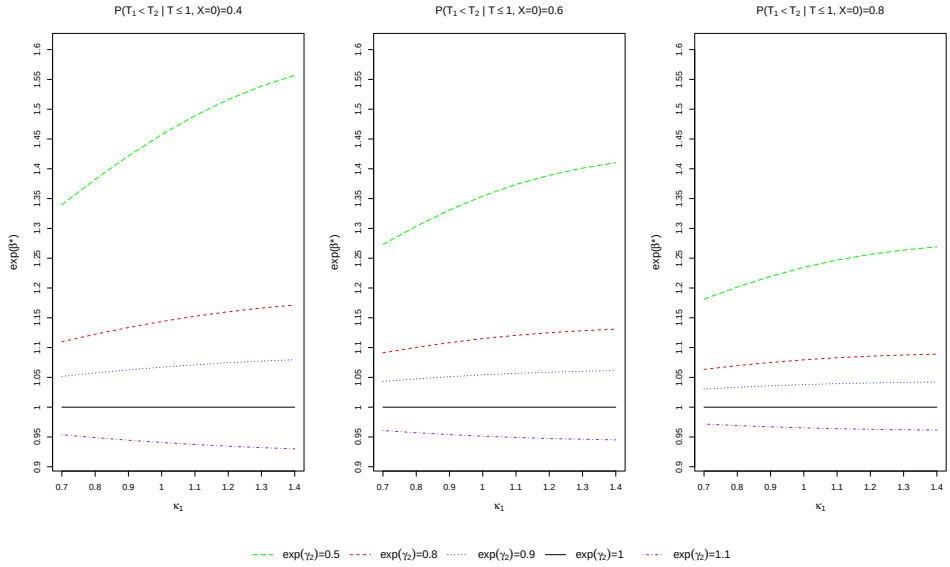


(a) $\exp(\gamma_1) = 1, \kappa_2 = 1$.

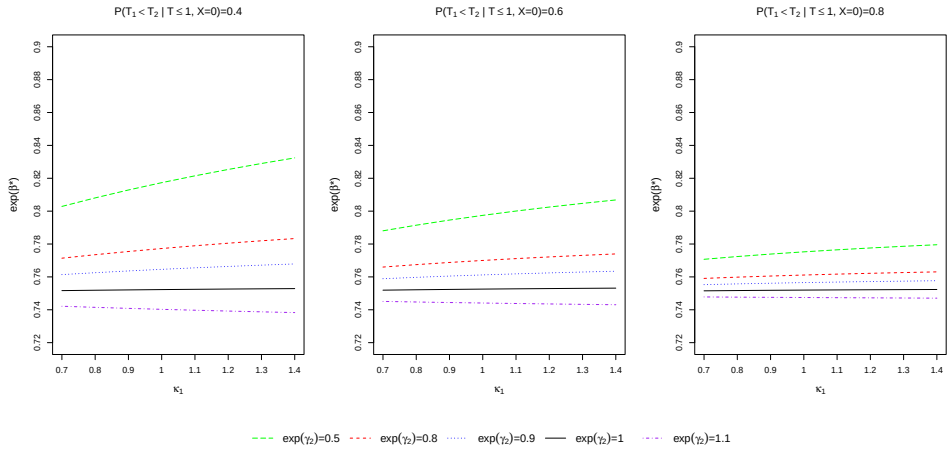


(b) $\exp(\gamma_1) = 0.75, \kappa_2 = 1$.

Figure B.5: Limiting values $\exp(\beta^\star_{FG})$ of the FG estimator as a function of κ_1 for different values of $\exp(\gamma_2)$ and $P(T_1 < T_2 | T \leq 1, X = 0)$ when the true process has Weibull intensities $\lambda_{0k}(t|X) = \kappa_k \lambda_k (\lambda_k t)^{\kappa_k - 1} \exp(\gamma_k X)$ with $\exp(\gamma_1) = 1$ (top row) and $e^{\gamma_1} = 0.75$ (bottom row); $\kappa_2 = 1, \tau = 1, P(T \leq 1|X = 0) = 0.6, G^\star = G$.



(a) $\exp(\gamma_1) = 1, \kappa_2 = 1$.



(b) $\exp(\gamma_1) = 0.75, \kappa_2 = 1$.

Figure B.6: Limiting values $\exp(\beta^\star_{DB})$ of the DB estimator as a function of κ_1 for different values of $\exp(\gamma_2)$ and $P(T_1 < T_2 | T \leq 1, X = 0)$ when the true process has Weibull intensities $\lambda_{0k}(t|X) = \kappa_k \lambda_k (\lambda_k t)^{\kappa_k - 1} \exp(\gamma_k X)$ for $\exp(\gamma_1) = 1$ (top row) and $\exp(\gamma_1) = 0.75$ (bottom row); $\kappa_2 = 1, R = 6, \tau = 1, P(T \leq 1|X = 0) = 0.6, G^\star = G$.

B.3 Proof of the robustness property

We prove the robustness property of Theorem 2 for both estimation methods; see Section B.3.1 for the FG procedure and Section B.3.2 for the DB procedure. Specifically, we show that $\beta^\star_{FG} = 0$ and $\beta^\star_{DB} = 0$ under $H_0^{\lambda_1\lambda_2}$ and H_0, but $\beta^\star_{FG} \neq 0$ and $\beta^\star_{DB} \neq 0$ under $H_0^{\lambda_1}$.

B.3.1 FG-based Wald test

We assume $G^\star = G$, and let $s^{(l,\star)}(t,\beta) = s^{(l)}(t,\beta) = \mathbb{E}_X\big((1 - F_1(t|X))X^l \exp(\beta X)\big)$ and $s^{(l,\star)}(t) = s^{(l)}(t) = \mathbb{E}_X\big(f_1(t|X)X^l\big)$ for $l = 0, 1$. Consider then (3.3.8) labelled as $\mathbb{E}(U^{FG})$ for the remainder of this section:

$$\begin{aligned}\mathbb{E}(U^{FG}) = \mathbb{E}(U^{FG}(\beta, G^\star)) &= \int_0^\infty \left\{ s^{(1)}(t) - \frac{s^{(1)}(t,\beta)}{s^{(0)}(t,\beta)} s^{(0)}(t) \right\} dt \\ &= \int_0^\infty \left\{ \mathbb{E}_X(X f_1(t|X)) - \frac{\mathbb{E}_X((1 - F_1(t|X))Xe^{\beta X})}{\mathbb{E}_X((1 - F_1(t|X))e^{\beta X})} \mathbb{E}_X(f_1(t|X)) \right\} dt \,. \end{aligned} \tag{B.3.1}$$

1. Under $H_0^{\lambda_1\lambda_2}$: $f_1(t|X) = f_1(t), F_1(t|X) = F_1(t)$

$$\begin{aligned} \therefore\ \mathbb{E}(U^{FG}) &= \int_0^\infty f_1(t) \left\{ \mathbb{E}_X(X) - \frac{\mathbb{E}_X(Xe^{\beta X})}{\mathbb{E}_X(e^{\beta X})} \right\} dt \\ &= P(T_1 < \infty) \left\{ \mathbb{E}_X(X) - \frac{\mathbb{E}_X(Xe^{\beta X})}{\mathbb{E}_X(e^{\beta X})} \right\} \\ &\stackrel{!}{=} 0 \iff \frac{\mathbb{E}_X(Xe^{\beta X})}{\mathbb{E}_X(e^{\beta X})} = \mathbb{E}_X(X), \end{aligned} \tag{B.3.2}$$

which holds iff $\beta = 0$.

To show: $\mathbb{E}_X(Xe^{\beta X}) = 0 \iff \beta = 0$.
Proof: We assume $\mathbb{E}_X(X) = 0$ without loss of generality.
$''\Longleftarrow''$: If $\beta = 0$, then $\mathbb{E}_X(Xe^{\beta X}) = \mathbb{E}_X(X) = 0$.
$''\Longrightarrow''$: Let $l(x) \geq 0$ be the probability density function of X and assume $\beta \neq 0$. Since

$$\int_{-\infty}^0 xe^{\beta x} l(x) dx = \int_0^\infty xe^{-\beta x} l(-x) dx \,,$$

we find that

$$\mathbb{E}_X(Xe^{\beta X}) = \int_{-\infty}^{\infty} xe^{\beta x} l(x)dx = \int_{-\infty}^{0} xe^{\beta x} l(x)dx + \int_{0}^{\infty} xe^{\beta x} l(x)dx$$
$$= \int_{0}^{\infty} x\big(\underbrace{e^{-\beta x}}_{>0 \text{ if } \beta \neq 0} \; l(-x) + \underbrace{e^{\beta x}}_{>0 \text{ if } \beta \neq 0} \; l(x)\big)dx \; . \tag{B.3.3}$$

If $\beta \neq 0$ the integrand in (B.3.3) is > 0 and, consequently, $\mathbb{E}_X(Xe^{\beta x}) > 0$. By proof of contradiction, β must be equal to 0.

2. Under H_0: This gives (B.3.2) also.

3. Under $H_0^{\lambda_1}$: In this case

$$f_1(t|X) = \lambda_{01}(t)S(t|X) = \lambda_{01}(t)e^{-\Lambda_{01}(t)-\Lambda_{02}(t|X)}, F_1(t|X) = \int_0^t \lambda_{01}(u)S(u|X)du \; ,$$

and (B.3.1) $\neq 0$ unless $\lambda_{02}(t|X) = \lambda_{02}(t)$.

B.3.2 DB-based Wald test

We assume $G^\star = G$, and consider (3.3.14) and (3.3.15) labelled as $\mathbb{E}(U_{\alpha_r}^{DB})$ and $\mathbb{E}(U_{\beta}^{DB})$ for the remainder of this section:

$$\mathbb{E}(U_{\alpha_r}^{DB}) = \mathbb{E}(U_1(\alpha_r, G^\star)) = \mathbb{E}_X\left(\frac{h'(\alpha_r + \beta X)\big(F_1(s_r|X) - h(\alpha_r + \beta X)\big)}{h(\alpha_r + \beta X)(1 - h(\alpha_r + \beta X))}\right)$$
$$\mathbb{E}(U_{\beta}^{DB}) = \mathbb{E}(U_1(\beta, G^\star)) = \sum_{r=1}^{R} \mathbb{E}_X\left(\frac{X \cdot h'(\alpha_r + \beta X)\big(F_1(s_r|X) - h(\alpha_r + \beta X)\big)}{h(\alpha_r + \beta X)(1 - h(\alpha_r + \beta X))}\right) .$$

1. Under $H_0^{\lambda_1\lambda_2}$: $F_1(s_r|X) = F_1(s_r)$ for all $r = 1, 2, \ldots, R$

$$\begin{aligned}
\therefore\ \mathbb{E}(U_{\alpha_r}^{DB}) &= F_1(s_r)\mathbb{E}_X\left(\frac{h'(\alpha_r+\beta X)}{h(\alpha_r+\beta X)(1-h(\alpha_r+\beta X))}\right) - \mathbb{E}_X\left(\frac{h'(\alpha_r+\beta X)}{(1-h(\alpha_r+\beta X))}\right) \\
&= \underbrace{\frac{P(X=0)h'(\alpha_r)}{(1-h(\alpha_r))}}_{>0}\left(\frac{F_1(s_r)}{h(\alpha_r)}-1\right) + \underbrace{\frac{P(X=1)h'(\alpha_r+\beta)}{(1-h(\alpha_r+\beta))}}_{>0}\left(\frac{F_1(s_r)}{h(\alpha_r+\beta)}-1\right) \\
&\stackrel{!}{=} 0 \iff F_1(s_r) = h(\alpha_r) \text{ and } F_1(s_r) = h(\alpha_r+\beta) \text{ for all } r = 1, 2, \ldots, R\,,
\end{aligned} \tag{B.3.4}$$

since $h(u) = g^{-1}(u) = 1 - \exp(-\exp(u)) > 0$ and $h'(u) = \exp(u - \exp(u)) > 0$ for all u. Similarly,

$$\begin{aligned}
\mathbb{E}(U_{\beta}^{DB}) &= \sum_{r=1}^{R}\Bigg\{F_1(s_r)\mathbb{E}_X\left(\frac{X\cdot h'(\alpha_r+\beta X)}{h(\alpha_r+\beta X)(1-h(\alpha_r+\beta X))}\right) - \\
&\qquad\qquad \mathbb{E}_X\left(\frac{X\cdot h'(\alpha_r+\beta X)}{(1-h(\alpha_r+\beta X))}\right)\Bigg\} \\
&= \sum_{r=1}^{R}\underbrace{\frac{P(X=1)h'(\alpha_r+\beta)}{(1-h(\alpha_r+\beta))}}_{>0}\left(\frac{F_1(s_r)}{h(\alpha_r+\beta)}-1\right) \\
&\stackrel{!}{=} 0 \iff F_1(s_r) = h(\alpha_r+\beta) \text{ for all } r = 1, 2, \ldots, R\,.
\end{aligned} \tag{B.3.5}$$

Note that $\beta = 0$ solves (B.3.4) and (B.3.5).

2. Under H_0: This gives (B.3.4) and (B.3.5) also.

3. Under $H_0^{\lambda_1}$: In this case (B.3.4) and (B.3.5) do not hold.

B.4 Additional results for the simulation study of Section 3.4.2

Here we present additional results for the simulation study of Section 3.4.2; the simulation setup is described in Section 3.4.2.1. Tables B.1 and B.2 report the empirical rejection rates for various tests for the setting where the true process is governed by intensity functions of the form $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X), k = 1, 2$ and where $P(T_1 < T_2|T \leq 1, X = 0) = 0.4$ and 0.8, along with the corresponding limiting values $\beta^\star_{FG}, \beta^\star_{DB_6}$ and $\beta^\star_{DB_3}$ under FG and DB estimation. Table B.3 represents the empirical rejection rates when data were generated according to models (3.4.7)-(3.4.8) and (3.4.9)-(3.4.10), and when $F_1(1|X = 0) = 0.24$ and 0.48.

The simulation results again confirm that $T^{F_1}_{FG}, T^{F_1}_{DB_6}$ and $T^{F_1}_{DB_3}$ are valid Wald tests of $\beta = 0$ under $H_0^{\lambda_1\lambda_2}$ and H_0, provided robust variance estimates are used. If the true data-generating process is implied by $H_0^{\lambda_1}$, these tests are no longer valid in the sense that they do not control the nominal 5% level. This is particularly the case when the treatment effect γ_2 on the $0 - 2$ intensity is reasonably large. If $\exp(\gamma_2)$ is close to 1, the type I error rate inflation is negligible. When comparing results presented in Tables B.1 and B.2, we see that the magnitude of inflation is also determined by the probability observed events are of type 1 in the control arm.

Simulation results based on trials with sample sizes of $n = 500$ and $n = 250$ are reported in Tables B.4 and B.7 for the intensity-based data-generating mechanism, and in Tables B.5 and B.6 for processes implied by models (3.4.7)-(3.4.8) and (3.4.9)-(3.4.10), respectively. For all scenarios considered, the type I error rate inflation of tests $T^{F_1}_{FG}, T^{F_1}_{DB_6}$ and $T^{F_1}_{DB_3}$ under $H_0^{\lambda_1}$ is at most 0.075 when $n = 500$ and 0.065 when $n = 250$. All tests for H_0 maintain the nominal 5% level under H_0, even for small sample sizes.

		REJECTION RATES							PROBABILITY LIMITS		
		$H_0^{\lambda_1}$		$H_0^{\lambda_1\lambda_2}$	H_0						
$\exp(\gamma_2)$	$\exp(\gamma_1)$	$T_{LR}^{\lambda_1}$	$T_{Cox}^{\lambda_1}$	$T_{Cox}^{\lambda_1\lambda_2}$	$T_{Gray}^{F_1}$	$T_{FG}^{F_1}$	$T_{DB_6}^{F_1}$	$T_{DB_3}^{F_1}$	$\beta_{FG}^{\star}$	$\beta_{DB_6}^{\star}$	$\beta_{DB_3}^{\star}$
0.5	1	0.0497	0.0487	0.9965	0.1324	0.1315	0.0901	0.0844	0.1183	0.0841	0.0797
	0.9	0.1154	0.1138	0.9986	0.0520	0.0515	0.0530	0.0542	0.0147	-0.0205	-0.0250
	0.75	0.4940	0.4910	0.9996	0.2003	0.2000	0.2291	0.2267	-0.1649	-0.2017	-0.2064
	0.6	0.9088	0.9072	1.0000	0.6973	0.6956	0.6946	0.6829	-0.3853	-0.4237	-0.4285
0.8	1	0.0517	0.0513	0.3423	0.0626	0.0624	0.0605	0.0574	0.0463	0.0331	0.0314
	0.9	0.1147	0.1133	0.4025	0.0693	0.0688	0.0742	0.0736	-0.057	-0.0712	-0.0730
	0.75	0.4749	0.4717	0.6471	0.3506	0.3482	0.3370	0.3267	-0.2362	-0.2520	-0.2540
	0.6	0.8941	0.8923	0.9216	0.8142	0.8122	0.7778	0.7621	-0.4562	-0.4736	-0.4757
0.9	1	0.0484	0.0477	0.1076	0.0494	0.0487	0.0485	0.0473	0.0230	0.0165	0.0156
	0.9	0.1129	0.1116	0.1603	0.0869	0.0862	0.0858	0.0842	-0.0802	-0.0878	-0.0887
	0.75	0.4709	0.4684	0.4477	0.3994	0.3954	0.3657	0.3508	-0.2593	-0.2684	-0.2695
	0.6	0.8839	0.8826	0.8423	0.8437	0.8440	0.8016	0.7866	-0.4792	-0.4899	-0.4912
1	1	0.0523	0.0508	0.0443	0.0516	0.0509	0.0484	0.0492	0.0000	0.0000	0.0000
	0.9	0.1108	0.1094	0.0883	0.1084	0.1078	0.1030	0.0987	-0.1031	-0.1041	-0.1043
	0.75	0.4652	0.4626	0.3640	0.4553	0.4526	0.4070	0.3879	-0.2821	-0.2846	-0.2850
	0.6	0.8890	0.8881	0.8154	0.8792	0.8776	0.8373	0.8193	-0.5019	-0.5060	-0.5065
1.1	1	0.0533	0.0525	0.1021	0.0561	0.0566	0.0547	0.0542	-0.0226	-0.0163	-0.0155
	0.9	0.1027	0.1019	0.1417	0.1266	0.1267	0.1108	0.1083	-0.1257	-0.1203	-0.1197
	0.75	0.4659	0.4635	0.4316	0.5115	0.5083	0.4461	0.4275	-0.3046	-0.3007	-0.3003
	0.6	0.8781	0.8768	0.8322	0.8917	0.8908	0.8473	0.8298	-0.5243	-0.5220	-0.5217
1.5	1	0.0435	0.0424	0.9227	0.1129	0.1112	0.0847	0.0781	-0.1100	-0.0797	-0.0760
	0.9	0.1099	0.1090	0.9340	0.2949	0.2933	0.2257	0.2090	-0.2129	-0.1835	-0.1799
	0.75	0.4411	0.4388	0.9684	0.6978	0.6951	0.5952	0.5584	-0.3915	-0.3635	-0.3601
	0.6	0.8654	0.8635	0.9940	0.9565	0.9551	0.9130	0.8962	-0.6110	-0.5844	-0.5812

Table B.1: Empirical rejection rates for different statistical tests of nominal size $\omega = 0.05$ when the true process has intensities $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X)$, along with the limiting values $\beta_{FG}^{\star}$ and $\beta_{DB}^{\star}$; $\gamma_1 : \log(0.9) = -0.1054, \log(0.75) = -0.2877, \log(0.6) = -0.5108$; $n = 1000$ individuals, $n_{sim} = 10000$, $\tau = 1$, $P(T \leq 1|X = 0) = 0.6$, $P(T_1 < T_2|T \leq 1, X = 0) = 0.4$, $\pi_r = 0.2$.

		REJECTION RATES							PROBABILITY LIMITS		
		$H_0^{\lambda_1}$		$H_0^{\lambda_1\lambda_2}$	H_0						
$\exp(\gamma_2)$	$\exp(\gamma_1)$	$T_{LR}^{\lambda_1}$	$T_{Cox}^{\lambda_1}$	$T_{Cox}^{\lambda_1\lambda_2}$	$T_{Gray}^{F_1}$	$T_{FG}^{F_1}$	$T_{DB_6}^{F_1}$	$T_{DB_3}^{F_1}$	$\beta_{FG}^{\star}$	$\beta_{DB_6}^{\star}$	$\beta_{DB_3}^{\star}$
0.5	1	0.0517	0.0512	0.7181	0.0748	0.0739	0.0592	0.0585	0.0436	0.0316	0.0297
	0.9	0.1730	0.1718	0.7887	0.0913	0.0911	0.0968	0.0977	-0.0604	-0.0732	-0.0751
	0.75	0.7710	0.7697	0.9562	0.6120	0.6115	0.5777	0.5618	-0.2406	-0.2546	-0.2567
	0.6	0.9963	0.9963	0.9994	0.9861	0.9862	0.9757	0.9719	-0.4615	-0.4768	-0.4791
0.8	1	0.0520	0.0516	0.1320	0.0552	0.0542	0.0548	0.0556	0.0172	0.0125	0.0118
	0.9	0.1794	0.1783	0.2355	0.1356	0.1363	0.1313	0.1266	-0.0864	-0.0919	-0.0928
	0.75	0.7704	0.7688	0.7358	0.7013	0.7003	0.6424	0.6227	-0.2661	-0.2729	-0.2740
	0.6	0.9964	0.9963	0.9923	0.9931	0.9930	0.9848	0.9810	-0.4866	-0.4948	-0.4960
0.9	1	0.0502	0.0500	0.0692	0.0515	0.0520	0.0504	0.0498	0.0086	0.0063	0.0059
	0.9	0.1728	0.1722	0.1557	0.1495	0.1501	0.1351	0.1323	-0.0949	-0.0981	-0.0986
	0.75	0.7616	0.7601	0.6789	0.7271	0.7264	0.6629	0.6406	-0.2745	-0.2790	-0.2797
	0.6	0.9949	0.9949	0.9883	0.9934	0.9934	0.9833	0.9788	-0.4948	-0.5007	-0.5016
1	1	0.0499	0.0498	0.0462	0.0500	0.0504	0.0479	0.0508	0.0000	0.0000	0.0000
	0.9	0.1736	0.1727	0.1331	0.1717	0.1703	0.1506	0.1492	-0.1034	-0.1043	-0.1044
	0.75	0.7569	0.7553	0.6555	0.7421	0.7415	0.6738	0.6484	-0.2829	-0.2851	-0.2854
	0.6	0.9950	0.9949	0.9864	0.9945	0.9946	0.9871	0.9842	-0.5031	-0.5067	-0.5072
1.1	1	0.0517	0.0516	0.0685	0.0514	0.0512	0.0488	0.0494	-0.0085	-0.0062	-0.0059
	0.9	0.1701	0.1691	0.1467	0.1885	0.1878	0.1594	0.1549	-0.1119	-0.1104	-0.1102
	0.75	0.7622	0.7610	0.6775	0.7746	0.7739	0.7035	0.6812	-0.2912	-0.2911	-0.2911
	0.6	0.9943	0.9942	0.9869	0.9939	0.9940	0.9879	0.9831	-0.5113	-0.5126	-0.5128
1.5	1	0.0527	0.0526	0.4758	0.0705	0.0703	0.0594	0.0579	-0.0422	-0.0308	-0.0291
	0.9	0.1684	0.1671	0.5613	0.2794	0.2795	0.2194	0.2074	-0.1451	-0.1347	-0.1331
	0.75	0.7578	0.7565	0.8807	0.8476	0.8473	0.7711	0.7447	-0.3240	-0.3149	-0.3135
	0.6	0.9944	0.9944	0.9971	0.9981	0.9979	0.9931	0.9902	-0.5437	-0.5360	-0.5348

Table B.2: Empirical rejection rates for different statistical tests of nominal size $\omega = 0.05$ when the true process has intensities $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X)$, along with the limiting values $\beta_{FG}^{\star}$ and $\beta_{DB}^{\star}$; $\gamma_1 : \log(0.9) = -0.1054, \log(0.75) = -0.2877, \log(0.6) = -0.5108$; $n = 1000$ individuals, $n_{sim} = 10000$, $\tau = 1$, $P(T \leq 1|X = 0) = 0.6$, $P(T_1 < T_2|T \leq 1, X = 0) = 0.8$, $\pi_r = 0.2$.

				REJECTION RATES							
				$H_0^{\lambda_1}$		$H_0^{\lambda_1\lambda_2}$	$H_0^{F_1}$				PROBABILITY LIMITS
Models	$F_1(1\|X=0)$	$\exp(\beta)$	$\exp(\beta_2)$	$T_{LR}^{\lambda_1}$	$T_{Cox}^{\lambda_1}$	$T_{Cox}^{\lambda_1\lambda_2}$	$T_{Gray}^{F_1}$	$T_{FG}^{F_1}$	$T_{DB_6}^{F_1}$	$T_{DB_3}^{F_1}$	$\beta_{FG}^{\star}, \beta_{DB_6}^{\star}, \beta_{DB_3}^{\star}$
		0.8		0.3908	0.3882	0.3878	0.3134	0.3117	0.2816	0.2724	-0.2232
		0.9		0.1247	0.1232	0.1171	0.1076	0.1072	0.0988	0.0975	-0.1054
	0.24	1	-	0.0507	0.0495	0.0527	0.0512	0.0519	0.0539	0.0535	0.0000
(3.4.7)		1.1		0.1152	0.1144	0.1035	0.0998	0.0986	0.0932	0.0926	0.0953
-											
(3.4.8)		0.8		0.5319	0.5300	0.4480	0.5485	0.5472	0.4872	0.4670	-0.2231
		0.9		0.1704	0.1690	0.1348	0.1793	0.1799	0.1583	0.1544	-0.1054
	0.48	1	-	0.0500	0.0498	0.0479	0.0502	0.0502	0.0470	0.0471	0.0000
		1.1		0.1504	0.1495	0.1252	0.1623	0.1621	0.1439	0.1369	0.0953
(3.4.9)		0.8	0.8	0.5501	0.5489	0.4644	0.5636	0.5639	0.4984	0.4794	-0.2231
-	0.48		1	0.4465	0.4455	0.5584	0.5528	0.5513	0.4883	0.4680	-0.2231
(3.4.10)		1	0.8	0.0572	0.0570	0.1134	0.0501	0.0497	0.0499	0.0496	0.0000
			1	0.0484	0.0478	0.0466	0.0484	0.0495	0.0492	0.0481	0.0000

Table B.3: Empirical rejection rates for different statistical tests of nominal size $\omega = 0.05$ when the true process was implied by models (3.4.7)-(3.4.8) and (3.4.9)-(3.4.10), along with the limiting values $\beta_{FG}^{\star}$ and $\beta_{DB}^{\star}$; $\beta : \log(0.8) = -0.2231, \log(0.9) = -0.1054, \log(1.1) = 0.0953$; $n = 1000$ individuals, $n_{sim} = 10000$, $\tau = 1, F_1(1|X=0) = 0.24$ or 0.48, $F_2(1|X=0) = 0.39$ or $0.15, \pi_r = 0.2$.

		REJECTION RATES							PROBABILITY LIMITS		
		$H_0^{\lambda_1}$		$H_0^{\lambda_1\lambda_2}$	H_0						
$\exp(\gamma_2)$	$\exp(\gamma_1)$	$T_{LR}^{\lambda_1}$	$T_{Cox}^{\lambda_1}$	$T_{Cox}^{\lambda_1\lambda_2}$	$T_{Gray}^{F_1}$	$T_{FG}^{F_1}$	$T_{DB_6}^{F_1}$	$T_{DB_3}^{F_1}$	$\beta_{FG}^{\star}$	$\beta_{DB_6}^{\star}$	$\beta_{DB_3}^{\star}$
0.5	1	0.0504	0.0494	0.7312	0.0753	0.0749	0.0622	0.0609	0.0825	0.0593	0.0560
	0.9	0.0991	0.0974	0.7611	0.0527	0.0516	0.0575	0.0589	-0.0210	-0.0453	-0.0487
	0.75	0.3814	0.3784	0.8700	0.2136	0.2141	0.2210	0.2135	-0.2003	-0.2264	-0.2299
	0.6	0.7986	0.7951	0.9629	0.6375	0.6369	0.6141	0.5990	-0.4205	-0.4483	-0.4520
0.8	1	0.0503	0.0494	0.1350	0.0554	0.0545	0.0511	0.0527	0.0324	0.0234	0.0221
	0.9	0.0933	0.0921	0.1701	0.0683	0.0682	0.0691	0.0713	-0.0706	-0.0808	-0.0822
	0.75	0.3788	0.3744	0.3982	0.2957	0.2944	0.2777	0.2690	-0.2495	-0.2614	-0.2630
	0.6	0.7967	0.7935	0.7631	0.7274	0.7256	0.6765	0.6579	-0.4691	-0.4828	-0.4846
0.9	1	0.0508	0.0501	0.0680	0.0519	0.0514	0.0509	0.0510	0.0161	0.0117	0.0110
	0.9	0.0959	0.0941	0.0966	0.0804	0.0805	0.0766	0.0751	-0.0868	-0.0924	-0.0932
	0.75	0.3702	0.3671	0.3071	0.3259	0.3246	0.2915	0.2791	-0.2655	-0.2728	-0.2738
	0.6	0.7889	0.7856	0.7090	0.7490	0.7478	0.6936	0.6709	-0.4850	-0.4941	-0.4953
1	1	0.0477	0.0466	0.0478	0.0476	0.0471	0.0496	0.0519	0.0000	0.0000	0.0000
	0.9	0.0979	0.0962	0.0732	0.0946	0.0937	0.0889	0.0847	-0.1028	-0.1040	-0.1041
	0.75	0.3684	0.3648	0.2811	0.3577	0.3563	0.3141	0.3003	-0.2814	-0.2842	-0.2846
	0.6	0.7816	0.7785	0.6894	0.7655	0.7635	0.7112	0.6873	-0.5007	-0.5054	-0.5060
1.1	1	0.0512	0.0503	0.0628	0.0538	0.0532	0.0523	0.0525	-0.0159	-0.0116	-0.0109
	0.9	0.0925	0.0906	0.0947	0.1040	0.1039	0.0955	0.0923	-0.1187	-0.1154	-0.1150
	0.75	0.3707	0.3666	0.2991	0.3899	0.3881	0.3387	0.3258	-0.2971	-0.2956	-0.2954
	0.6	0.7847	0.7813	0.6948	0.7949	0.7926	0.7306	0.7066	-0.5163	-0.5166	-0.5166
1.5	1	0.0500	0.0489	0.4593	0.0759	0.0751	0.0649	0.0612	-0.0780	-0.0569	-0.0540
	0.9	0.0924	0.0906	0.5091	0.1839	0.1816	0.1440	0.1335	-0.1805	-0.1604	-0.1577
	0.75	0.3570	0.3538	0.6718	0.5136	0.5106	0.4314	0.4056	-0.3585	-0.3400	-0.3376
	0.6	0.7657	0.7625	0.8819	0.8613	0.8593	0.7931	0.7672	-0.5774	-0.5605	-0.5583

Table B.4: Empirical rejection rates for different statistical tests of nominal size $\omega = 0.05$ when the true process has intensities $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X)$, along with the limiting values $\beta_{FG}^{\star}$ and $\beta_{DB}^{\star}$; $\gamma_1 : \log(0.9) = -0.1054, \log(0.75) = -0.2877, \log(0.6) = -0.5108$; $n = 500$ individuals, $n_{sim} = 10000$, $\tau = 1$, $P(T \leq 1|X = 0) = 0.6$, $P(T_1 < T_2|T \leq 1, X = 0) = 0.6$, $\pi_r = 0.2$.

				REJECTION RATES							PROBABILITY LIMITS
				$H_0^{\lambda_1}$		$H_0^{\lambda_1\lambda_2}$	$H_0^{F_1}$				
n	Models	$\exp(\beta)$	$\exp(\beta_2)$	$T_{LR}^{\lambda_1}$	$T_{Cox}^{\lambda_1}$	$T_{Cox}^{\lambda_1\lambda_2}$	$T_{Gray}^{F_1}$	$T_{FG}^{F_1}$	$T_{DB_6}^{F_1}$	$T_{DB_3}^{F_1}$	$\beta_{FG}^{\star}, \beta_{DB_6}^{\star}, \beta_{DB_3}^{\star}$
500	(3.4.7) - (3.4.8)	0.8	-	0.2668	0.2633	0.1980	0.2467	0.2453	0.2187	0.2092	-0.2231
		0.9		0.0985	0.0972	0.0802	0.0955	0.0942	0.0872	0.0853	-0.1054
		1		0.0498	0.0487	0.0458	0.0511	0.0511	0.0498	0.0502	0.0000
		1.1		0.090	0.0882	0.0758	0.0873	0.0865	0.0800	0.0796	0.0953
	(3.4.9) - (3.4.10)	0.8	0.8	0.2706	0.2669	0.2004	0.2451	0.2441	0.2154	0.2062	-0.2231
			1	0.2011	0.1979	0.2176	0.2522	0.2506	0.2262	0.2123	
		0.9	0.8	0.1193	0.1168	0.1198	0.0984	0.0973	0.0911	0.0899	-0.1054
			1	0.0808	0.0800	0.0798	0.0914	0.0906	0.0835	0.0818	
		1	0.8	0.0561	0.0551	0.1203	0.0536	0.0538	0.0538	0.0523	0.0000
			1	0.0513	0.0501	0.0518	0.0500	0.0494	0.0500	0.0502	
		1.1	0.8	0.0662	0.0647	0.1957	0.0959	0.0955	0.0899	0.0857	0.0953
			1	0.0817	0.0804	0.0819	0.0887	0.0885	0.0842	0.0822	

Table B.5: Empirical rejection rates for different statistical tests of nominal size $\omega = 0.05$ when the true process is implied by models (3.4.7)-(3.4.8) and (3.4.9)-(3.4.10), along with the limiting values $\beta_{FG}^{\star}$ and $\beta_{DB}^{\star}$; $\beta : \log(0.8) = -0.2231, \log(0.9) = -0.1054, \log(1.1) = 0.0953$; $n = 500$ individuals, $n_{sim} = 10000$, $\tau = 1, F_1(1|X = 0) = 0.36, F_2(1|X = 0) = 0.27, \pi_r = 0.2$.

				REJECTION RATES							PROBABILITY LIMITS
				$H_0^{\lambda_1}$		$H_0^{\lambda_1\lambda_2}$	$H_0^{F_1}$				
n	Models	$\exp(\beta)$	$\exp(\beta_2)$	$T_{LR}^{\lambda_1}$	$T_{Cox}^{\lambda_1}$	$T_{Cox}^{\lambda_1\lambda_2}$	$T_{Gray}^{F_1}$	$T_{FG}^{F_1}$	$T_{DB_6}^{F_1}$	$T_{DB_3}^{F_1}$	$\beta_{FG}^{\star}, \beta_{DB_6}^{\star}, \beta_{DB_3}^{\star}$
250	(3.4.7) - (3.4.8)	0.8	-	0.1573	0.1533	0.1191	0.1468	0.1448	0.1310	0.1244	-0.2231
		0.9		0.0687	0.0661	0.0584	0.0692	0.0681	0.0664	0.0633	-0.1054
		1		0.0532	0.0506	0.0489	0.0529	0.0522	0.0507	0.0510	0.0000
		1.1		0.0699	0.0673	0.0598	0.0679	0.0678	0.0649	0.0624	0.0953
	(3.4.9) - (3.4.10)	0.8	0.8	0.1608	0.1560	0.1196	0.1500	0.1484	0.1338	0.1310	-0.2231
			1	0.1264	0.1225	0.1219	0.1487	0.1473	0.1322	0.1250	
		0.9	0.8	0.0818	0.0777	0.0704	0.0714	0.0707	0.0639	0.0605	-0.1054
			1	0.0639	0.0622	0.0604	0.0709	0.0687	0.0656	0.0648	
		1	0.8	0.0527	0.0510	0.0759	0.0521	0.0513	0.0491	0.0495	0.0000
			1	0.0489	0.0473	0.0465	0.0477	0.0470	0.0482	0.0469	
		1.1	0.8	0.0552	0.0526	0.1106	0.0705	0.0696	0.0676	0.0674	0.0953
			1	0.0656	0.0628	0.0649	0.0695	0.0691	0.0675	0.0669	

Table B.6: Empirical rejection rates for different statistical tests of nominal size $\omega = 0.05$ when the true process is implied by models (3.4.7)-(3.4.8) and (3.4.9)-(3.4.10), along with the limiting values $\beta_{FG}^{\star}$ and $\beta_{DB}^{\star}$; $\beta : \log(0.8) = -0.2231, \log(0.9) = -0.1054, \log(1.1) = 0.0953$; $n = 250$ individuals, $n_{sim} = 10000$, $\tau = 1, F_1(1|X = 0) = 0.36, F_2(1|X = 0) = 0.27, \pi_r = 0.2$.

		REJECTION RATES									
		$H_0^{\lambda_1}$		$H_0^{\lambda_1\lambda_2}$	H_0				PROBABILITY LIMITS		
$\exp(\gamma_2)$	$\exp(\gamma_1)$	$T_{LR}^{\lambda_1}$	$T_{Cox}^{\lambda_1}$	$T_{Cox}^{\lambda_1\lambda_2}$	$T_{Gray}^{F_1}$	$T_{FG}^{F_1}$	$T_{DB_6}^{F_1}$	$T_{DB_3}^{F_1}$	$\beta_{FG}^{\star}$	$\beta_{DB_6}^{\star}$	$\beta_{DB_3}^{\star}$
0.5	1	0.0526	0.0504	0.4126	0.0661	0.0645	0.0588	0.0587	0.0825	0.0593	0.0560
	0.9	0.0703	0.0678	0.4337	0.0472	0.0465	0.0511	0.0517	-0.0210	-0.0453	-0.0487
	0.75	0.2188	0.2114	0.5442	0.1278	0.1256	0.1341	0.1307	-0.2003	-0.2264	-0.2299
	0.6	0.5123	0.5037	0.7268	0.3709	0.3665	0.3514	0.3376	-0.4205	-0.4483	-0.4520
0.8	1	0.0451	0.0431	0.0841	0.0478	0.0461	0.0470	0.0453	0.0324	0.0234	0.0221
	0.9	0.0762	0.0724	0.1046	0.0596	0.0586	0.0596	0.0580	-0.0706	-0.0808	-0.0822
	0.75	0.2063	0.2010	0.2018	0.1657	0.1644	0.1533	0.1495	-0.2495	-0.2614	-0.2630
	0.6	0.4993	0.4912	0.4396	0.4328	0.4301	0.3927	0.3786	-0.4691	-0.4828	-0.4846
0.9	1	0.0498	0.0467	0.0566	0.0523	0.0518	0.0481	0.0509	0.0161	0.0117	0.0110
	0.9	0.0693	0.0672	0.0685	0.0642	0.0624	0.0615	0.0587	-0.0868	-0.0924	-0.0932
	0.75	0.2120	0.2055	0.1685	0.1839	0.1826	0.1703	0.1615	-0.2655	-0.2728	-0.2738
	0.6	0.4873	0.4781	0.3871	0.4509	0.4469	0.4077	0.3907	-0.4850	-0.4941	-0.4953
1	1	0.0501	0.0482	0.0456	0.0520	0.0519	0.0526	0.0523	0.0000	0.0000	0.0000
	0.9	0.0728	0.0705	0.0591	0.0705	0.0694	0.0680	0.0649	-0.1028	-0.1040	-0.1041
	0.75	0.2136	0.2085	0.1556	0.2073	0.2043	0.1814	0.1736	-0.2814	-0.2842	-0.2846
	0.6	0.4977	0.4888	0.3874	0.4828	0.4791	0.4303	0.4084	-0.5007	-0.5054	-0.5060
1.1	1	0.0508	0.0484	0.0539	0.0508	0.0496	0.0529	0.0500	-0.0159	-0.0116	-0.0109
	0.9	0.0702	0.0669	0.0672	0.0817	0.0811	0.0742	0.0744	-0.1187	-0.1154	-0.1150
	0.75	0.2063	0.2010	0.1618	0.2134	0.2110	0.1870	0.1805	-0.2971	-0.2956	-0.2954
	0.6	0.4984	0.4886	0.3919	0.5080	0.5040	0.4468	0.4246	-0.5163	-0.5166	-0.5166
1.5	1	0.0519	0.0502	0.2415	0.0623	0.0614	0.0530	0.0520	-0.0780	-0.0569	-0.0540
	0.9	0.0687	0.0660	0.2542	0.1141	0.1114	0.0909	0.0873	-0.1805	-0.1604	-0.1577
	0.75	0.1990	0.1925	0.3689	0.2869	0.2829	0.2345	0.2206	-0.3585	-0.3400	-0.3376
	0.6	0.4750	0.4662	0.5736	0.5729	0.5690	0.4975	0.4719	-0.5774	-0.5605	-0.5583

Table B.7: Empirical rejection rates for different statistical tests of nominal size $\omega = 0.05$ when the true process has intensities $\lambda_{0k}(t|X) = \lambda_k \exp(\gamma_k X)$, along with the limiting values $\beta_{FG}^{\star}$ and $\beta_{DB}^{\star}$; $\gamma_1 : \log(0.9) = -0.1054, \log(0.75) = -0.2877, \log(0.6) = -0.5108$; $n = 250$ individuals, $n_{sim} = 10000$, $\tau = 1$, $P(T \leq 1|X = 0) = 0.6$, $P(T_1 < T_2|T \leq 1, X = 0) = 0.6$, $\pi_r = 0.2$.

Appendix C

Appendix to Chapter 4

C.1 Details on the pseudo-Poisson procedures of Section 4.3.2

C.1.1 Software: an illustrative dataframe

For illustration, we consider four hypothetical individuals with the following sample paths:

- Individual 1 is alive and event-free at the administrative closure of study, i.e., $Z_1(s_r) = 0$ for all $r = 1, \ldots, R$ and $Z_1(\tau) = 0$.
- Individual 2 is alive at τ, with non-fatal events at times $T_{21} \in (s_1, s_2]$ and $T_{22} \in (s_4, s_5]$.
- Individual 3 has experienced a non-fatal event at $T_{31} \in (0, s_1]$ and withdrew earlier from study at $C_{r3} \in (s_2, s_3]$.
- Individual 4 died at time $T_4 \in (s_2, s_3]$ without having experienced non-fatal events prior to T_4, i.e., $Z_4(t) = 0$ for $0 < t < T_4$ and $Z_4(T_4) = D_0$.

Table C.1 shows the observed data in the usual dataframe format, whereas Table C.2 illustrates the stacked data structure required for fitting the pseudo-Poisson procedures of Section 4.3.2. Each individual in the stacked dataframe has R lines, one for each time point s_r. Listings C.1 and C.2 provide the `R` code for the call to the corresponding `geese()` function (Halekoh et al., 2006).

id	trt	from	to	start	stop	event.status	death.status
1	0	0	cens	0	τ	0	0
2	1	0	1	0	T_{21}	1	0
2	1	1	2	T_{21}	T_{22}	1	0
2	1	2	cens	T_{22}	τ	0	0
3	1	0	1	0	T_{31}	1	0
3	1	1	cens	T_{31}	C_{r3}	0	0
4	0	0	D_0	0	T_4	0	1

Table C.1: Standard dataframe with start-stop notation for four hypothetical individuals experiencing the recurrent and terminal event process in Figure 1.1(e).

Listing C.1: R code for pseudo-Poisson procedure based on weighted estimating equations

```
geese(Nsr ~ factor(sr) + as.factor(X) - 1, data=data, id=
    data$id, weights=data$weights, scale.fix=TRUE, family=
    gaussian(), mean.link="log", corstr="independence", variance
    ="poisson")
```

Listing C.2: R code for pseudo-Poisson procedure based on weighted counts

```
geese(weighted.Nsr ~ factor(sr) + as.factor(X) - 1, data=data,
id=data$id, scale.fix=TRUE, family=gaussian(), mean.link="log",
    corstr="independence", variance="poisson")
```

id	X_i	s_r	$\min(T_i, s_r)$	$N_i(s_r)$	$\mathbb{1}(C_i \geq \min(T_i, s_r))$	$\widehat{G}_i(\min(T_i, s_r))$	$\widehat{w}_i(\min(T_i, s_r))$	$\widetilde{N}_i(s_r)$
1	0	s_1	s_1	0	1	$\widehat{G}_1(s_1)$	$1/\widehat{G}_1(s_1)$	0
1	0	s_2	s_2	0	1	$\widehat{G}_1(s_2)$	$1/\widehat{G}_1(s_2)$	0
1	0	...	...	...	...	...	...	...
1	0	s_{R-1}	s_{R-1}	0	1	$\widehat{G}_1(s_{R-1})$	$1/\widehat{G}_1(s_R)$	0
1	0	s_R	s_R	0	1	$\widehat{G}_1(s_R)$	$1/\widehat{G}_1(s_{R-1})$	0
2	1	s_1	s_1	0	1	$\widehat{G}_2(s_1)$	$1/\widehat{G}_2(s_1)$	0
2	1	s_2	s_2	1	1	$\widehat{G}_2(s_2)$	$1/\widehat{G}_2(s_2)$	$1/\widehat{G}_2(s_2)$
2	1	s_3	s_3	1	1	$\widehat{G}_2(s_3)$	$1/\widehat{G}_2(s_3)$	$1/\widehat{G}_2(s_3)$
2	1	s_4	s_4	1	1	$\widehat{G}_2(s_4)$	$1/\widehat{G}_2(s_4)$	$1/\widehat{G}_2(s_4)$
2	1	s_5	s_5	2	1	$\widehat{G}_2(s_5)$	$1/\widehat{G}_2(s_5)$	$2/\widehat{G}_2(s_5)$
2	1	...	...	...	...	...	...	...
2	1	s_{R-1}	s_{R-1}	2	1	$\widehat{G}_2(s_{R-1})$	$1/\widehat{G}_2(s_{R-1})$	$2/\widehat{G}_2(s_{R-1})$
2	1	s_R	s_R	2	1	$\widehat{G}_2(s_R)$	$1/\widehat{G}_2(s_R)$	$2/\widehat{G}_2(s_R)$
3	1	s_1	s_1	1	1	$\widehat{G}_3(s_1)$	$1/\widehat{G}_3(s_1)$	$1/\widehat{G}_3(s_1)$
3	1	s_2	s_2	1	1	$\widehat{G}_3(s_2)$	$1/\widehat{G}_3(s_2)$	$1/\widehat{G}_3(s_2)$
3	1	s_3	s_3	NA	0	$\widehat{G}_3(s_3)$	0	0
3	1	s_4	s_4	NA	0	$\widehat{G}_3(s_4)$	0	0
3	1	...	...	...	...	...	...	...
3	1	s_R	s_R	NA	0	$\widehat{G}_3(s_R)$	0	0
4	0	s_1	s_1	0	1	$\widehat{G}_4(s_1)$	$1/\widehat{G}_4(s_1)$	0
4	0	s_2	s_2	0	1	$\widehat{G}_4(s_2)$	$1/\widehat{G}_4(s_2)$	0
4	0	s_3	T_4	0	1	$\widehat{G}_4(T_4)$	$1/\widehat{G}_4(T_4)$	0
4	0	s_4	T_4	0	1	$\widehat{G}_4(T_4)$	$1/\widehat{G}_4(T_4)$	0
4	0	...	...	...	...	...	...	...
4	0	s_R	T_4	0	1	$\widehat{G}_4(T_4)$	$1/\widehat{G}_4(T_4)$	0

Table C.2: Illustration of a stacked dataframe for fitting the pseudo-Poisson procedures of Sections 4.3.2.1 and 4.3.2.2 at time points $s_1, s_2, \ldots, s_R$; $C_i = \min(C_{ri}, \tau)$: net censoring time, T_i: time to death, weighted event count: $\widetilde{N}_i(s_r) = \widehat{w}_i(s_r)N_i(s_r)$ for $\widehat{w}_i(s_r) = \mathbb{1}(C_{ri} \geq \min(T_i, s_r))/\widehat{G}_i(\min(T_i, s_r))$.

C.2 Derivation of the cumulative mean function under scenario GL-1-2

Since $dN(u) = \sum_{k=1}^{K} Y_{k-1}(u) dN_{(k-1)k}(u)$ for $dN_{(k-1)k}(u) = \mathbb{1}(Z(u) = k, Z(u^-) = k-1)$ and $N(u^-) = k, T \geq u \iff Z(u^-) = k$, the conditional rate function $r(u|X)$ under scenario *GL-1-2* (see Section 4.4.1.2) can be derived as

$$\begin{aligned} r(u|X) &= \mathbb{E}(dN(u)|T \geq u, X) = \sum_{k=1}^{K} \mathbb{E}\big(\mathbb{1}(Z(u^-) = k-1) dN_{(k-1)k}(u)|T \geq u, X)\big) \\ &= \sum_{k=1}^{K} \underbrace{P(Z(u^-) = k-1|T \geq u, X)}_{P(Z(u^-)=k-1, T\geq u|X)/P(T\geq u|X)} \mathbb{E}(dN_{(k-1)k}(u)| \underbrace{Z(u^-) = k-1, T \geq u}_{Z(u^-)=k-1}, X) \\ &= \frac{1}{P(T \geq u|X)} \sum_{k=1}^{K} P(Z(u^-) = k-1|X) \lambda_{(k-1)k}(u|X) \ , \end{aligned}$$

which in turns give (4.4.6). If $a = 1$, then

$$r(u|X) = \frac{\lambda(u|X)}{P(T \geq u|X)} \underbrace{\sum_{k=1}^{K+1} P(Z(u^-) = k-1|X)}_{P(T\geq u|X)} = \lambda(u|X) \ ,$$

since $\lambda_{K(K+1)}(t|X) = 0$ per definition.

C.3 Limiting values of treatment effect estimators in mean function regression under different data-generating frameworks

C.3.1 The Ghosh-Lin estimator

C.3.1.1 Poisson processes for recurrent non-fatal events and death (GL-1-1)

The red line in Figure C.1 is patterned on the pamidronate trial (Hortobagyi et al., 1996) of Section 1.3.3, in which patients were administratively censored at 2 years and effects on the recurrent skeletal-related event intensity and the hazard for time to death were observed to be approximately $e^{\gamma} = 0.55$ and $e^{\delta} = 1.1$; see also Section 4.5.

C.3.1.2 Markov processes with event-specific baseline intensities (GL-1-2)

If the death intensities vary with the cumulative number of non-fatal events as in Figure C.2, the mean function ratio based on (4.4.6) is constant over time only if $\exp(\gamma) = \exp(\delta) = 1$ (not shown here). Figure C.3 contains plots of $\exp(\beta_P^{\star})$ as a function of $b \geq 1$ for different values of $\mathbb{E}(N(\tau)|X = 0)$ and $P(T \leq \tau|X = 0)$. Results from the pseudo-Poisson procedure are aligned with those from the GL procedure in Figure 4.5, with $\beta_P^{\star}$ closer to γ than $\beta_{GL}^{\star}$.

C.3.2 The Mao-Lin estimator

Figure C.4 contains contour plots depicting the dependence of the limiting value $\exp(\beta_P^{\dagger\star})$ on treatment and the process intensities.

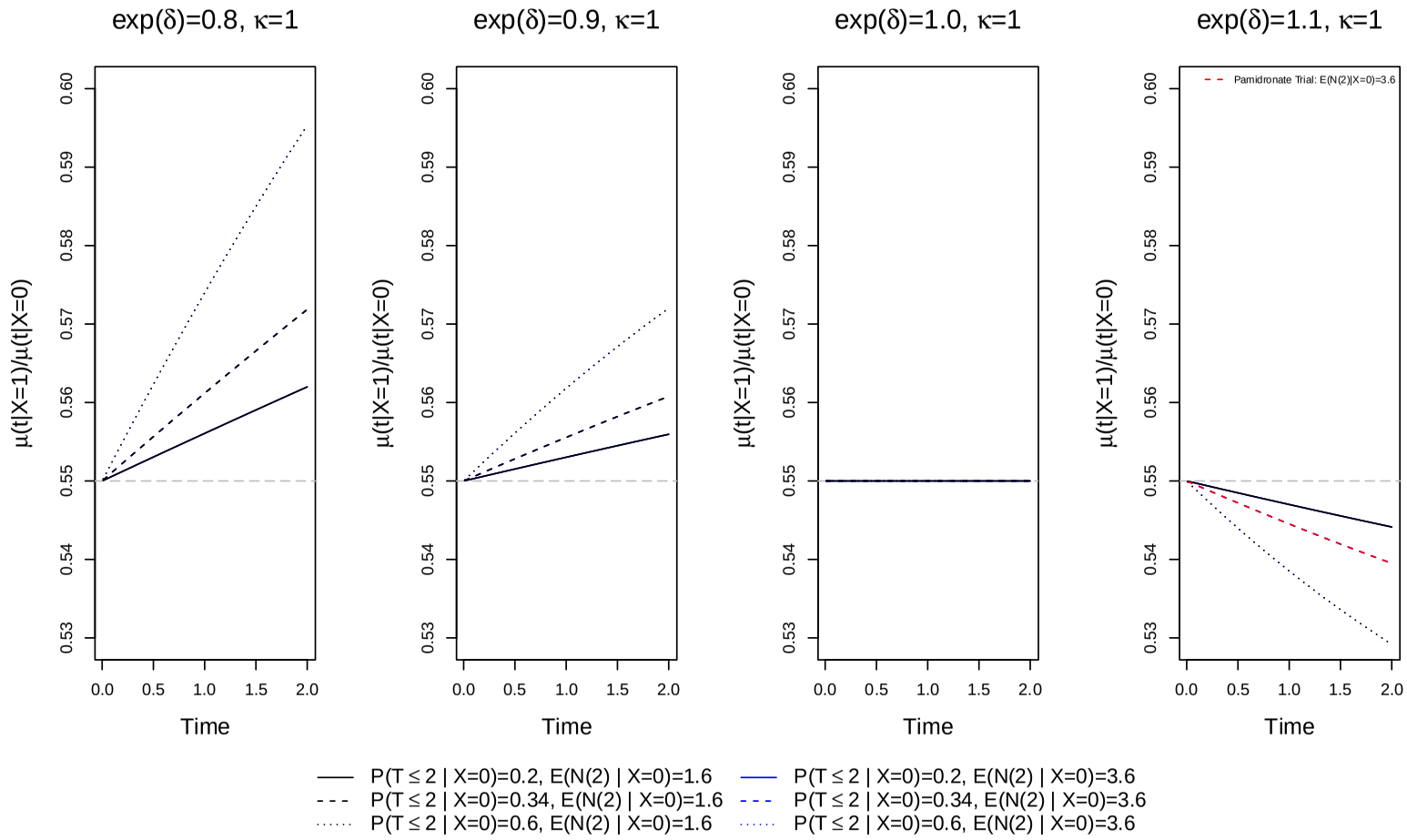


Figure C.1: *GL-1-1:* Plots of $\mu(t|X=1)/\mu(t|X=0)$ for $e^{\gamma}=0.55$ and different values of $P(T\leq\tau|X=0), \delta$ and $\mathbb{E}(N(\tau)|X=0)$ when the true process has intensities $\lambda(t|X)=\lambda_0\exp(\gamma X)$ and $\rho(t|X)=\rho_0\exp(\delta X)$; $\kappa=1, \tau=2$; the red line corresponds to the parameter setting observed in the pamidronate trial of Section 1.3.3 (Hortobagyi et al., 1996).

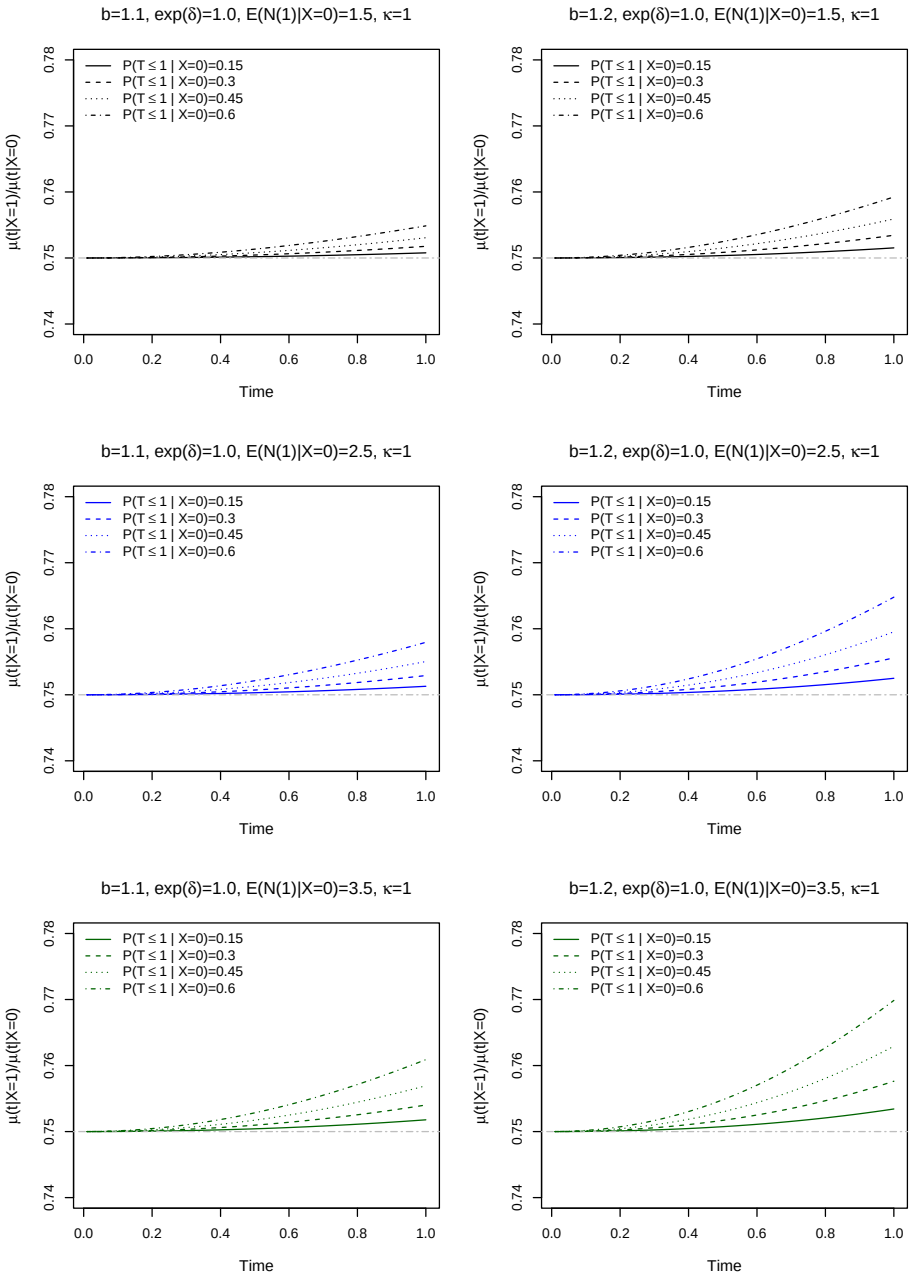


Figure C.2: *GL-1-2:* Plots of $\mu(t|X=1)/\mu(t|X=0)$ for $\exp(\gamma)=0.75, \exp(\delta)=1$ and different values of $P(T \leq \tau|X=0), b$ and $\mathbb{E}(N(\tau)|X=0)$ when the true process has intensities $\lambda_k(t|X)=\lambda(t|X)=\lambda_0\exp(\gamma X)$ $(a=1)$ and $\rho_k(t|X)=b^k\rho_0\exp(\delta X)$; $K=20, \tau=1, \kappa=1$.

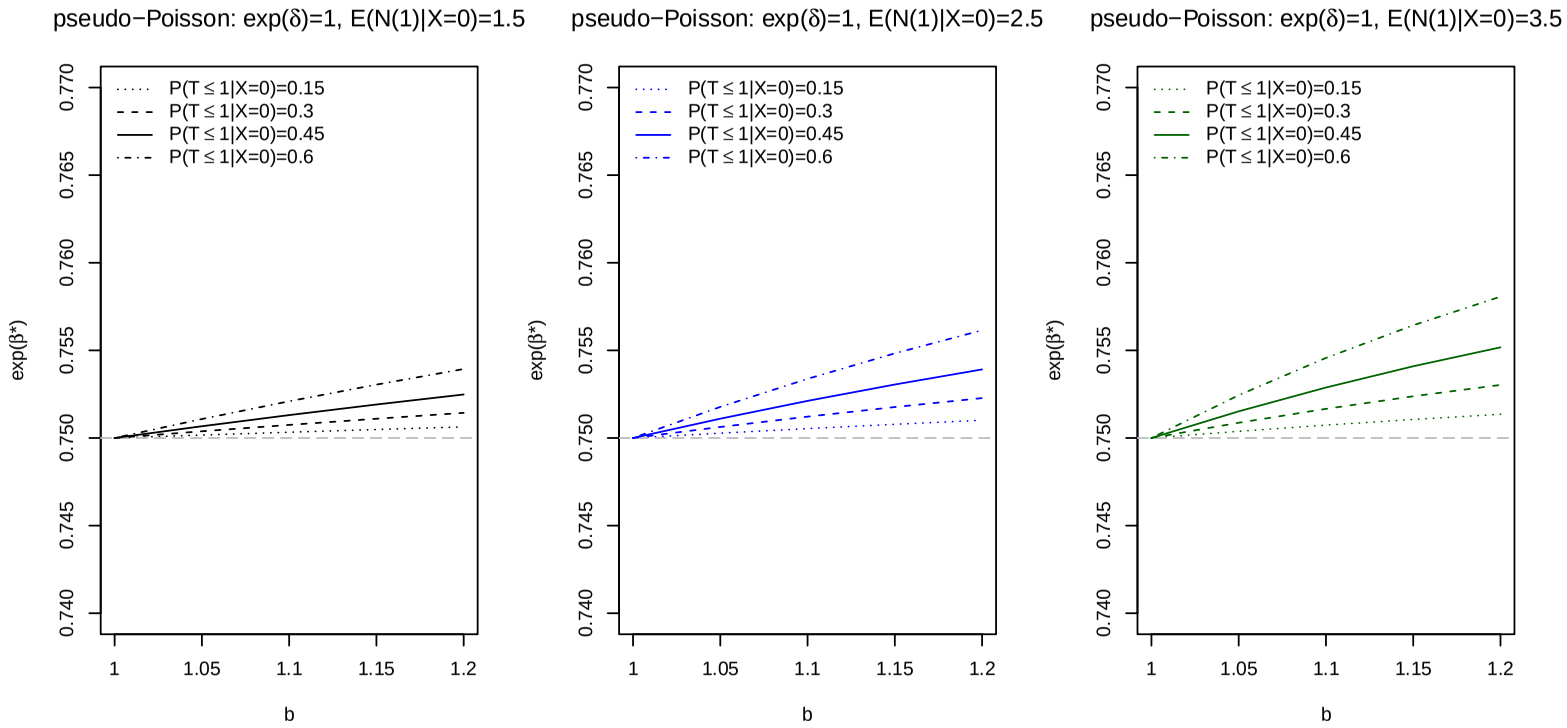
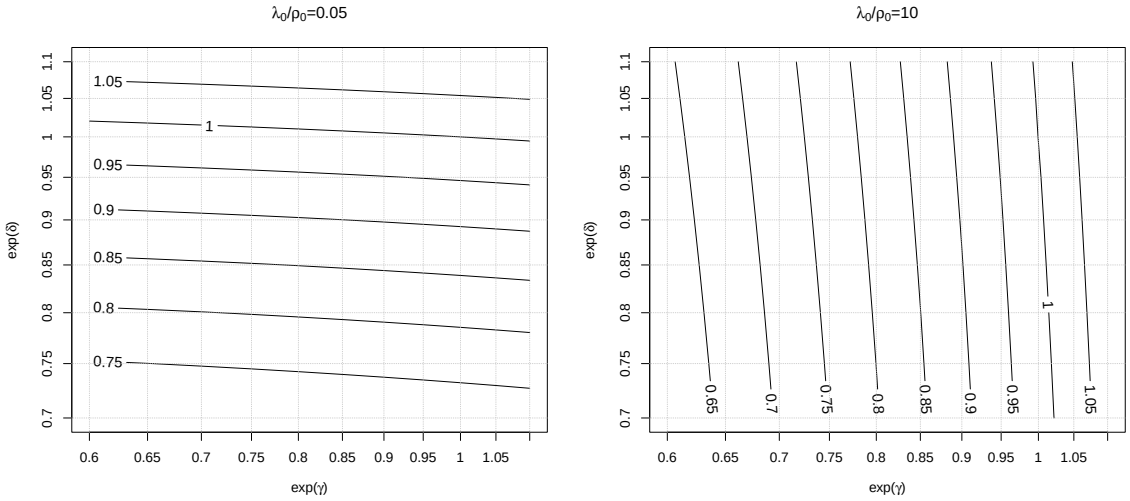
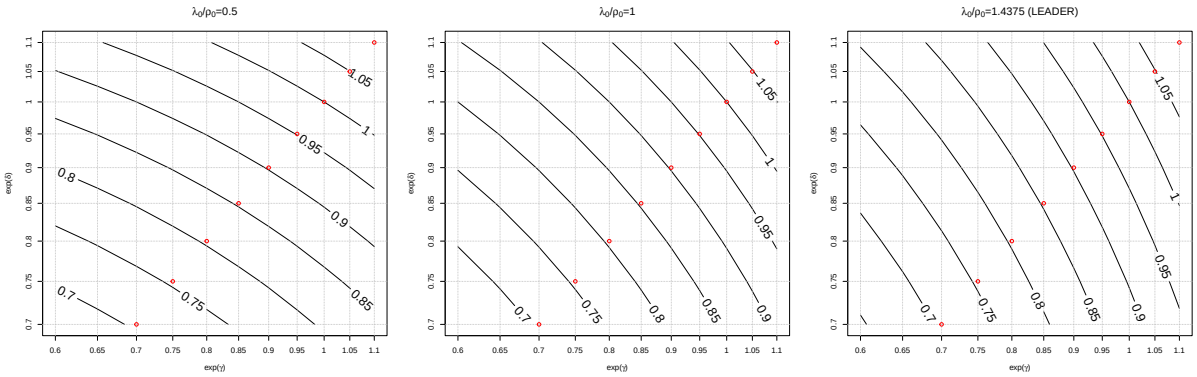


Figure C.3: *GL-1-2:* Limiting values $\exp(\beta_P^\star)$ as a function of b for different values of $P(T \leq \tau|X = 0)$ and $\mathbb{E}(N(\tau)|X = 0) = 1.5$ (left panel), 2.5 (middle panel), 3.5 (right panel) when the true process has intensities $\lambda_k(t|X) = \lambda(t|X) = \lambda_0 \exp(\gamma X)$ and $\rho_k(t|X) = b^k \rho_0 \exp(\delta X)$ for $\exp(\gamma) = 0.75$ and $\exp(\delta) = 1$; $K = 20, a = 1, \kappa = 1, \tau = 1$, $G^\star = G$.



(a) Settings where the Mao-Lin estimand is easy to interpret.
Left panel: $\lambda_0/\rho_0 \approx 0$; right panel: $\lambda_0/\rho_0 \nearrow \infty$.



(b) Settings where the Mao-Lin estimand is less easy to interpret.
Left panel: $\lambda_0/\rho_0 = 0.5$; middle panel: $\lambda_0/\rho_0 = 1$ and right panel: $\lambda_0/\rho_0 \approx 1.5$.

Figure C.4: *ML-1-1:* Plots showing the dependency of the limiting values $\exp(\beta_P^{\dagger\star})$ on treatment and the time-homogeneous process intensities $\lambda(t|X) = \lambda_0 \exp(\gamma X), \rho(t|X) = \rho_0 \exp(\delta X)$ and $\psi(t|X) = \psi_0$; $\kappa = 1, \tau = 5, P(T \leq \tau | X = 0) = 0.12$, pseudo-Poisson estimation based on $R = 6$ equi-spaced time points on $(0, \tau), G^\star = G$; the red circles indicate the points on the grid where $\exp(\gamma) = \exp(\delta)$.

C.4 Additional results for the application of Section 4.5

Figure C.5 depicts plots of the Nelson-Aalen estimates of Markov cumulative intensity functions for the occurrence of new skeletal complications, death and censoring for patients treated with pamidronate. Figure C.6 contains plots of the Nelson-Aalen estimates of Markov cumulative intensity functions for the occurrence of new skeletal-related events. Figure C.7 depicts plots of the Nelson-Aalen estimates of Markov cumulative intensity functions for the occurrence of death.

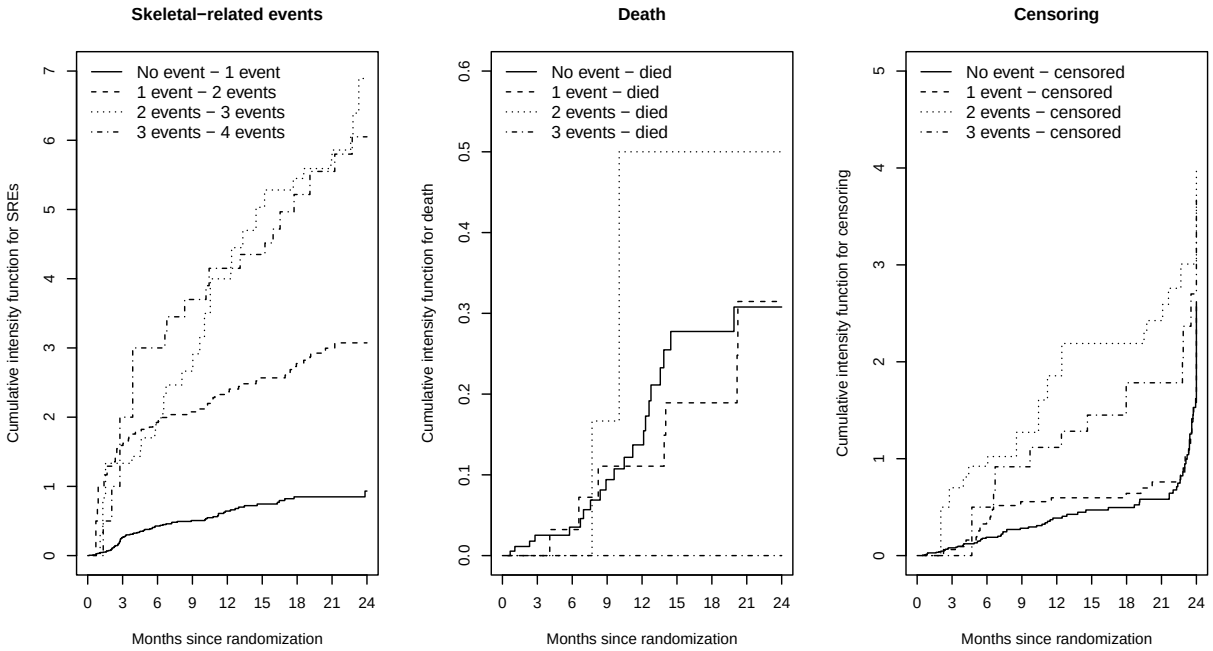


Figure C.5: Nonparametric Nelson-Aalen estimates of the cumulative intensity functions for the occurrence of new skeletal-related events (left panel), death (middle panel) and censoring (right panel) for pamidronate patients in Hortobagyi et al. (1996).

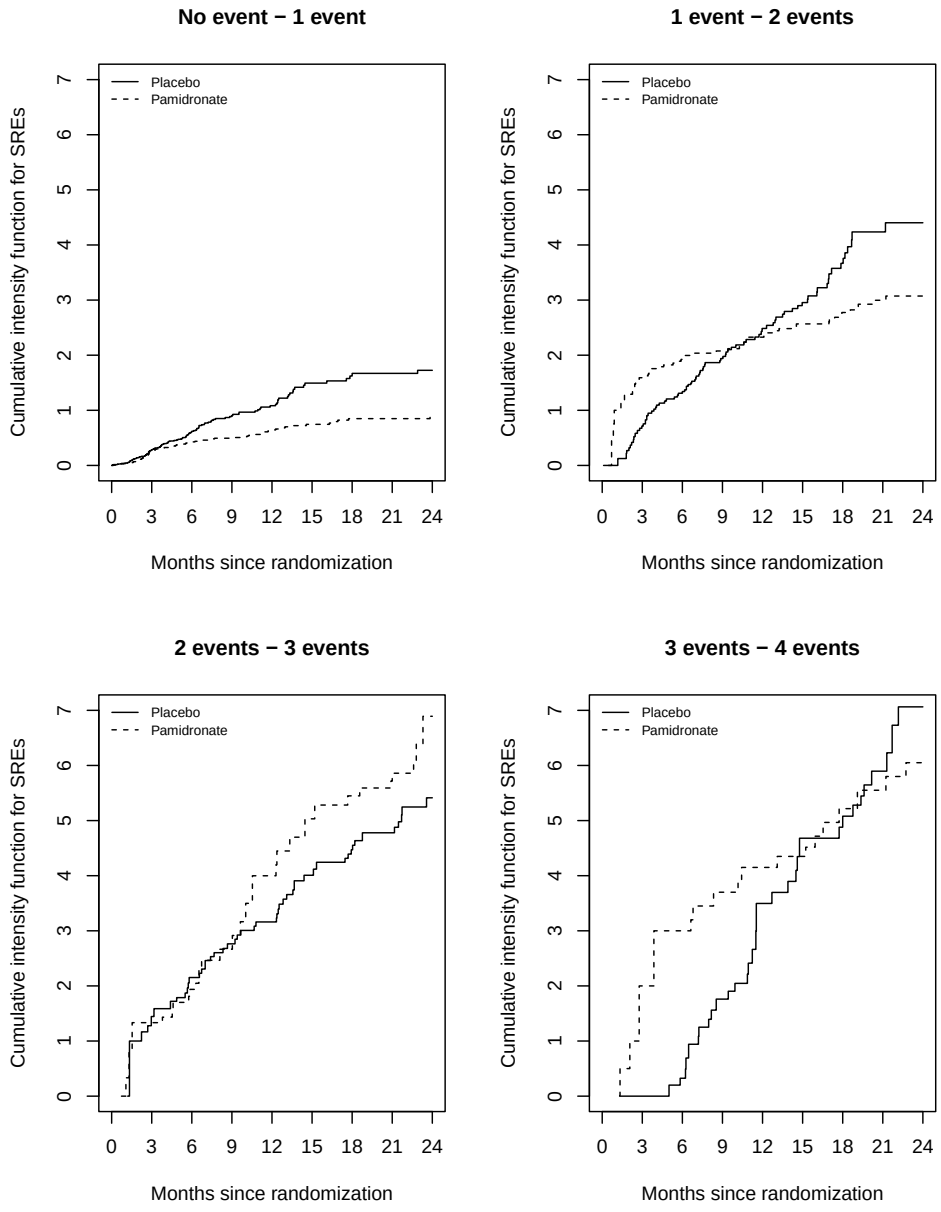


Figure C.6: Nonparametric Nelson-Aalen estimates of the cumulative intensity functions for the occurrence of one, two, three or four skeletal-related events (Hortobagyi et al., 1996); upper panels represent the NA estimates for $0 \longrightarrow 1$ and $1 \longrightarrow 2$ transitions in Figure 1.1(e); lower panels represent the $2 \longrightarrow 3$ and $3 \longrightarrow 4$ transitions.

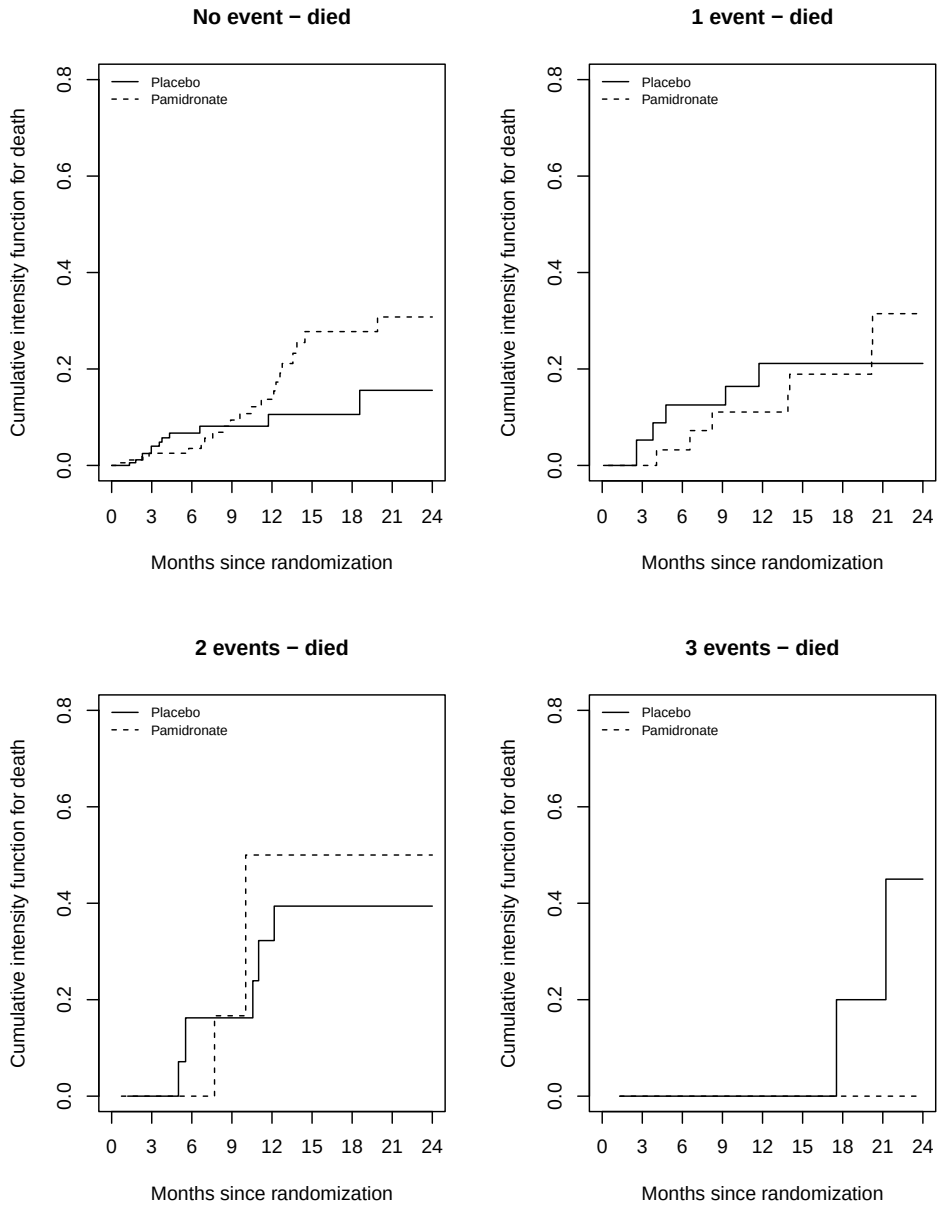


Figure C.7: Nonparametric Nelson-Aalen estimates of the cumulative intensity functions for the occurrence of death according to the cumulative number of previous skeletal-related events (Hortobagyi et al., 1996); upper panels represent the NA estimates for the $0 \longrightarrow D_0$ and $1 \longrightarrow D_1$ transitions in Figure 1.1(e); lower panels represent the $2 \longrightarrow D_2$ and $3 \longrightarrow D_3$ transitions.

Appendix D

Appendix to Chapter 5

D.1 Details on the win ratio approach of Section 5.3.1

D.1.1 Proof of equality of win probabilities

Here we prove the equality of (5.3.1), (5.3.2) and (5.3.3) under the assumption that $\{Z_i(t), t > 0\}$ is independent of $\{Z_j(t), t > 0\}$. We first show the equality of $p_{10}^{ID}(t)$=(5.3.1) and (5.3.3):

$$
\begin{aligned}
p_{10}^{ID}(t) &= P(T_{2j} \le t, T_{2j} \le T_{2i}) + P(T_{2j} > t, T_{2i} > t, T_{1j} \le t, T_{1j} \le T_{1i}) \\
&= P(T_{2j} \le t, T_{2i} \le t, T_{2j} \le T_{2i}) + P(T_{2j} \le t, T_{2i} > t) \\
&\qquad + P(T_{2j} > t, T_{2i} > t, T_{1j} \le t, T_{1i} \le t, T_{1j} \le T_{1i}) \\
&\qquad + P(T_{2j} > t, T_{2i} > t, T_{1j} \le t, T_{1i} > t) \\
&= P(Z_i(t) = 2, Z_j(t) = 2, T_{2j} \le T_{2i}) + P(Z_j(t) = 2, Z_i(t) \ne 2) \\
&\qquad + P(Z_i(t) = 1, Z_j(t) = 1, T_{1j} \le T_{1i}) + P(Z_j(t) = 1, Z_i(t) = 0) \\
&= (5.3.3)\ .
\end{aligned}
$$

Since

$$
\begin{aligned}
P(T_{2j} \le t, T_{2j} \le T_{2i}) &= P(Z_i(t) = 2, Z_j(t) = 2, T_{2j} \le T_{2i}) + P(Z_j(t) = 2, Z_i(t) \ne 2) \\
&= \int_0^t P(Z_i(t) = 2, T_{2i} > u) P(Z_j(t) = 2, T_{2j} = u) du + P(T_{2j} \le t) P(T_{2i} > t) \\
&= \int_0^t f_{2j}(u) \left\{ \int_u^t P(T_{2i} = s) ds \right\} du + P(T_{2j} \le t)(1 - P(T_{2i} \le t)) \\
&= \int_0^t f_{2j}(u)(P(T_{2i} \le t) - P(T_{2i} \le u)) du + P(T_{2j} \le t) - P(T_{2i} \le t) \int_0^t f_{2j}(u) du \\
&= \int_0^t f_{2j}(u) du - \int_0^t f_{2j}(u) P(T_{2i} \le u) du = \int_0^t f_{2j}(u) P(T_{2i} > u) du \ ,
\end{aligned}
$$

and

$$
\begin{aligned}
& P(T_{2i} > t) P(T_{2j} > t) \int_0^t g_j(u|t) P(T_{1i} > u | T_{2i} > t) du \\
&\quad = \int_0^t P(T_{1j} = u, T_{2j} \ge t) P(T_{1i} > u, T_{2i} > t) du \\
&\quad = \int_0^t P(T_{1j} = u, Z_j(t) = 1) \big[P(T_{1i} > u, Z_i(t) = 0) + P(T_{1i} > u, Z_i(t) = 1) \big] du \\
&\quad = \int_0^t P(T_{1j} = u, Z_j(t) = 1) P(T_{1i} > u, Z_i(t) = 1) du \\
&\qquad\quad + P(Z_i(t) = 0) \int_0^t P(T_{1j} = u, Z_j(t) = 1) du \\
&\quad = P(Z_i(t) = 1, Z_j(t) = 1, T_{1j} < T_{1i}) + P(Z_i(t) = 0) P(Z_j(t) = 1) \\
&\quad = P(T_{2j} > t, T_{2i} > t, T_{1j} \le t, T_{1j} \le T_{1i})
\end{aligned}
$$

we find that $p_{10}^{ID}(t)$=(5.3.2).

D.1.2 Derivation of win probabilities for illness-death processes

We find that

$$
\begin{aligned}
&P(Z_i(t) = Z_j(t) = 1, T_{1j} < T_{1i}) \\
&\quad = \int_0^t P(Z_i(t) = 1, Z_j(t) = 1, T_{1j} = u, T_{1i} > u) du \\
&\quad = \int_0^t P(Z_j(t) = 1, T_{1j} = u) P(Z_i(t) = 1, T_{1i} > u) du \\
&\quad = \int_0^t \Big\{ \underbrace{P(Z(t) = 1, T_1 = u | X = 0)}_{= P_{00}(0,u^-|X=0)\lambda_{01}(u|X=0)P_{11}(u,t^-|X=0)} \cdot \underbrace{P(Z(t) = 1, T_1 > u | X = 1)}_{= \int_u^t P_{00}(0,s^-|X=1)\lambda_{01}(s|X=1)P_{11}(s,t^-|X=1)ds} \Big\} du \\
&\quad = \int_0^t \Big\{ P_{00}(0, u^-|X = 0)\lambda_{01}(u|X = 0)P_{11}(u, t^-|X = 0) \cdot \\
&\qquad\qquad \int_u^t P_{00}(0, s^-|X = 1)\lambda_{01}(s|X = 1)P_{11}(s, t^-|X = 1)ds \Big\} du \\
&\quad = \int_0^t \big\{ P_{00}(0, u^-|X = 0)\lambda_{01}(u|X = 0)P_{11}(u, t^-|X = 0)h(u, t) \big\} du \, ,
\end{aligned}
$$

where

$$
h(u, t) = \int_u^t P_{00}(0, s^-|X = 1)\lambda_{01}(s|X = 1)P_{11}(s, t^-|X = 1)ds \, .
$$

Accordingly, we obtain

$$
\begin{aligned}
&P(Z_i(t) = Z_j(t) = 1, T_{1i} < T_{1j}) \\
&\quad = \int_0^t \big\{ P_{00}(0, u^-|X = 1)\lambda_{01}(u|X = 1)P_{11}(u, t^-|X = 1)\widetilde{h}(u, t) \big\} du
\end{aligned}
$$

for

$$
\widetilde{h}(u, t) = \int_u^t P_{00}(0, s^-|X = 0)\lambda_{01}(s|X = 0)P_{11}(s, t^-|X = 0)ds \, .
$$

Similarly,

$$
\begin{aligned}
P(Z_i(t) = Z_j(t) = 2, T_{2j} < T_{2i}) &= \int_0^t P(Z_i(t) = 2, Z_j(t) = 2, T_{2j} = u, T_{2i} > u)du \\
&= \int_0^t P(Z_j(t) = 2, T_{2j} = u) P(Z_i(t) = 2, T_{2i} > u)du \\
&= \int_0^t \Big\{ \underbrace{P(Z(t) = 2, T_2 = u | X = 0)}_{= P(Z(t)=2, T_{02}=u|X=0) + P(Z(t)=2, T_{12}=u|X=0)} \cdot \underbrace{P(Z(t) = 2, T_2 > u | X = 1)}_{= P(Z(t)=2, T_{02}>u|X=1) + P(Z(t)=2, T_{12}>u|X=1)} \Big\} du \\
&= \int_0^t \big\{ P(Z(t) = 2, T_{02} = u | X = 0) + P(Z(t) = 2, T_{12} = u | X = 0) \big\} \cdot \\
&\qquad \big\{ P(Z(t) = 2, T_{02} > u | X = 1) + P(Z(t) = 2, T_{12} > u | X = 1) \big\} du \\
&= \int_0^t \big\{ P_{00}(0, u^- | X = 0) \lambda_{02}(u | X = 0) + g_1(u) \big\} \cdot \big\{ g_2(u, t) + g_3(u, t) \big\} du \ ,
\end{aligned}
$$

where

$$
\begin{aligned}
g_1(u) &= P(Z(t) = 2, T_{12} = u | X = 0) \\
&= \int_0^u P_{00}(0, s^- | X = 0) \lambda_{01}(s | X = 0) P_{11}(s, u^- | X = 0) \lambda_{12}(u | X = 0) ds \\
&= \lambda_{12}(u | X = 0) \int_0^u P_{00}(0, s^- | X = 0) \lambda_{01}(s | X = 0) P_{11}(s, u^- | X = 0) ds \ ,
\end{aligned}
$$

and

$$
g_2(u, t) = P(Z(t) = 2, T_{02} > u | X = 1) = \int_u^t P_{00}(0, v^- | X = 1) \lambda_{02}(v | X = 1) dv \ ,
$$

$$
\begin{aligned}
g_3(u, t) &= P(Z(t) = 2, T_{12} > u | X = 1) \\
&= \int_u^t \int_0^v P_{00}(0, w^- | X = 1) \lambda_{01}(w | X = 1) P_{11}(w, v^- | X = 1) \lambda_{12}(v | X = 1) dw \, dv \\
&= \int_u^t \lambda_{12}(v | X = 1) \underbrace{\left(\int_0^v P_{00}(0, w^- | X = 1) \lambda_{01}(w | X = 1) P_{11}(w, v^- | X = 1) dw \right)}_{:= g_4(v)} dv \\
&= \int_u^t \lambda_{12}(v | X = 1) g_4(v) dv.
\end{aligned}
$$

Consequently,

$$g_2(u,t) + g_3(u,t) = \int_u^t \left\{ P_{00}(0, v^-|X=1)\lambda_{02}(v|X=1) + \lambda_{12}(v|X=1)g_4(v) \right\} dv \ .$$

The other three expressions are rather simple and can be obtained as

$$\begin{aligned}
P(Z_i(t)=0, Z_j(t)=1) &= P_{00}(0,t|X=1)P_{01}(0,t|X=0) \ , \\
P(Z_i(t)=0, Z_j(t)=2) &= P_{00}(0,t|X=1)P_{02}(0,t|X=0) \ , \\
P(Z_i(t)=1, Z_j(t)=2) &= P_{01}(0,t|X=1)P_{02}(0,t|X=0) \ .
\end{aligned}$$

D.1.3 Simulation study

A small simulation study was conducted to confirm the unbiasedness of $\widehat{p}_{ij}^{ID}(\tau)$ and $\widehat{p}_{ji}^{ID}(\tau)$ under nonparametric Nelson-Aalen and Aalen-Johansen estimation.
We set the administrative censoring time to $\tau = 1$ and generated data from a time-homogeneous illness-death process with intensities $\lambda_{jl}(t|X) = \lambda_{jl0}\exp(\gamma_{jl}X)$, $j < l$, following the procedure in Section 2.5 of Cook and Lawless (2018). The binary treatment indicator X was simulated from a binomial distribution with $P(X=1) = 0.5$. The baseline intensities $\lambda_{010}, \lambda_{020}$ and λ_{120} were set to satisfy the following constraints: (1) $\pi_T = P(T \leq \tau|X=0) = 0.6$, so 40% of individuals in the control group are alive and event-free (in state 0) at time τ, (2) $\pi_T^1 = P(T_{01} < T_{02}|T \leq \tau, X=0) = 0.7$, so 42% of individuals with $X=0$ are expected to experience the non-fatal event of interest (i.e. pass through state 1) and 18% die event-free ($0 \longrightarrow 2$ transition), and (3) $\pi_D = P(Z(\tau)=2|X=0) = 0.25$, so 7% of individuals in the control group are expected to die post-event ($1 \longrightarrow 2$ transition) by time τ. The random right-censoring time C_r was generated from an exponential distribution Exp(ρ) with the rate parameter ρ set to achieve $\pi_r = P(C_r < \tau|X=0) = 0.0, 0.2, 0.4$ and 0.6, with $\pi_r = 0$ corresponding to administrative censoring only. We set $e^{\gamma_{01}} = 0.75, e^{\gamma_{02}} = e^{\gamma_{12}} = 0.9$ or $e^{\gamma_{01}} = e^{\gamma_{02}} = e^{\gamma_{12}} = 1$.

For each parameter setting we consider the following approaches: the standard WR approach based on generalized pairwise comparisons (*unadj-D*, *unadj-M*), the IPCW-adjusted approach by Dong et al. (2020c) (*adj-D*), and the proposed multistate approach (5.3.4) based on nonparametric

π_T	π_T^1	π_D	$[e^{\gamma_{01}}, e^{\gamma_{02}}, e^{\gamma_{12}}]$	$p_{ij}(\tau)$	$p_{ji}(\tau)$	$\beta^\dagger(\tau)$
0.6	0.7	0.25	0.75, 0.9, 0.9	0.4430	0.3640	1.2171
			1, 1, 1	0.4200	0.4200	1.0000

Table D.1: True values for $p_{ij}(\tau), p_{ji}(\tau)$ and $\beta^\dagger(\tau)$ when $\tau = 1$.

Nelson-Aalen and Aalen-Johansen estimation (*MSM*). We simulated $n_{sim} = 2000$ datasets with $n = 250$ individuals and obtained the WR estimates using the `win.stat()` (*unadj-D*, *adj-D*), `WRrec()` (*unadj-M*) and `etm()` (*MSM*) functions in the respective `R` packages `WINS`, `WR` and `etm` (Dong et al., 2020c; Mao et al., 2022; Allignol et al., 2011).

Simulation results are provided in Table D.2. For each of the two parameter settings the true values for $p_{ij}(\tau), p_{ji}(\tau)$ and $\beta^\dagger(\tau)$ are reported on in Table D.1. Censoring introduce bias in the estimation of $\widehat{p}_{ij}^{ID}(\tau)$ and $\widehat{p}_{ji}^{ID}(\tau)$ for all methods, with the exception of the multistate model approach.

D.2 Reformulation of the utility-based estimand

Here we show that

$$\beta(\tau) = \sum_{k=1}^{K+1} \sum_{r<k} (u_r - u_k)\big(RMT_{rk}(\tau) - RMT_{kr}(\tau)\big) \ .$$

Consider a progressive multistate process $\{Z(t), 0 < t\}$ with state space $\mathcal{S} = \{0, 1, \ldots, K, K+1\}$ and utility values $u_s \in [0, 1]$ defined as in Section 5.2. Since $u_{K+1} = 0$ and

$$\mathbb{1}(Z_i(t) = s) - \mathbb{1}(Z_j(t) = s) = \begin{cases} 0, & Z_i(t) = Z_j(t) = s \\ 1, & Z_i(t) = s, Z_j(t) \neq s \\ -1, & Z_i(t) \neq s, Z_j(t) = s \end{cases} ,$$

$[e^{\gamma_{01}}, e^{\gamma_{02}}, e^{\gamma_{12}}]$	π_r	WR approach	MEAN($\widehat{p}_{ij}(\tau)$)	MEAN($\widehat{p}_{ji}(\tau)$)	MEAN($\widehat{\beta}^{\dagger}(\tau)$)
0.75, 0.9, 0.9	0.0	unadj-M	0.4429	0.3635	1.2393
		unadj-D	0.4429	0.3635	1.2393
		adj-D	0.4535	0.3530	1.3068
		MSM	0.4405	0.3611	1.2410
	0.2	unadj-M	0.3778	0.3114	1.2360
		unadj-D	0.3965	0.3268	1.2360
		adj-D	0.4054	0.3179	1.2992
		MSM	0.4384	0.3622	1.2334
	0.4	unadj-M	0.3161	0.2585	1.2488
		unadj-D	0.3499	0.2871	1.2442
		adj-D	0.3559	0.2811	1.2910
		MSM	0.4398	0.3619	1.2421
	0.6	unadj-M	0.2531	0.2053	1.2651
		unadj-D	0.2978	0.2426	1.2581
		adj-D	0.3025	0.2380	1.3019
		MSM	0.4398	0.3595	1.2565
1,1,1	0.0	unadj-M	0.4203	0.4194	1.0165
		unadj-D	0.4203	0.4194	1.0165
		adj-D	0.4205	0.4193	1.0178
		MSM	0.4174	0.4166	1.0166
	0.2	unadj-M	0.3615	0.3599	1.0207
		unadj-D	0.3789	0.3776	1.0199
		adj-D	0.3790	0.3776	1.0186
		MSM	0.4174	0.4161	1.0194
	0.4	unadj-M	0.3018	0.3040	1.0109
		unadj-D	0.3334	0.3357	1.0108
		adj-D	0.3333	0.3359	1.0092
		MSM	0.4149	0.4176	1.0123
	0.6	unadj-M	0.2446	0.2439	1.0249
		unadj-D	0.2867	0.2859	1.0234
		adj-D	0.2872	0.2854	1.0256
		MSM	0.4158	0.4148	1.0261

Table D.2: Empirical average of $\widehat{p}_{ij}(\tau), \widehat{p}_{ji}(\tau)$ and $\widehat{\beta}^{\dagger}(\tau)$ for different values of $\pi_r = P(C_r \leq \tau | X = 0) = 0, 0.2, 0.4$ and 0.6 when the true illness-death process is Markov with intensities $\lambda_{jl}(t|X) = \lambda_{jl0} \exp(\gamma_{jl} X)$; the true values for $p_{ij}(\tau), p_{ij}(\tau)$ and $\beta^{\dagger}(\tau)$ are provided in Table D.1; $\tau = 1, \pi_T = 0.6, \pi_T^1 = 0.7, \pi_D = 0.25, n = 250$ individuals, $n_{sim} = 2000$ simulation runs.

the utility-based estimand $\beta(\tau)$ can be rewritten as

$$\begin{aligned}
\beta(\tau) = \mathbb{E}(U_i(\tau) - U_j(\tau)) &= \mathbb{E}\Big(\sum_{s=0}^{K+1} u_s \int_0^\tau \big(\mathbb{1}(Z_i(t)=s) - \mathbb{1}(Z_j(t)=s)\big)dt\Big) \\
&= \mathbb{E}\Big(\sum_{s=0}^{K+1} u_s \int_0^\tau \big(\mathbb{1}(Z_i(t)=s, Z_j(t)\neq s) - \mathbb{1}(Z_i(t)\neq s, Z_j(t)=s)\big)dt\Big) \\
&= \sum_{s=0}^{K+1}\sum_{r>s} u_s \int_0^\tau \big(P(Z_i(t)=s, Z_j(t)=r) - P(Z_i(t)=r, Z_j(t)=s)\big)dt\Big) \\
&\qquad + \sum_{s=0}^{K+1}\sum_{r<s} u_s \int_0^\tau \big(P(Z_i(t)=s, Z_j(t)=r) - P(Z_i(t)=r, Z_j(t)=s)\big)dt \\
&= \sum_{s=0}^{K+1}\sum_{r>s} u_s \int_0^\tau \big(P(Z_i(t)=s, Z_j(t)=r) - P(Z_i(t)=r, Z_j(t)=s)\big)dt \\
&\qquad - \sum_{s=0}^{K+1}\sum_{r<s} u_s \int_0^\tau \big(P(Z_i(t)=r, Z_j(t)=s) - P(Z_i(t)=s, Z_j(t)=r)\big)dt \\
&= \sum_{k=1}^{K+1}\sum_{r<k} u_r \int_0^\tau \big(P(Z_i(t)=k, Z_j(t)=r) - P(Z_i(t)=r, Z_j(t)=s)\big)dt \\
&\qquad - \sum_{k=1}^{K}\sum_{r<k} u_k \int_0^\tau \big(P(Z_i(t)=r, Z_j(t)=k) - P(Z_i(t)=k, Z_j(t)=r)\big)dt\ .
\end{aligned}$$

D.3 Illustration of the recursive infinitesimal jackknife variance formula for $\widehat{D}(\tau)$ in illness-death processes

Here we apply the recursive infinitesimal jackknife variance formula for $\widehat{D}(\tau)$ to a specific multistate process. We consider for illustration the extended illness-death model in Figure D.1, with separate states for event-free death (state 2) and post-event death (state $2'$); thus $\mathcal{S} = \{0, 1, 2', 2\}$. Assume $P(Z(0) = 0) = 1$ so that state 0 is the initial state. In what follows, we use $k, l \in \mathcal{S}$ to index states of the illness-death process and $i, i', i'' = 1, 2, \ldots, n$ to index individuals. We further assume that there are no ties, i.e. the counting process $dN_\cdot(t) =$

$(dN_{\cdot kl}(t) = \left(\sum_{i=1}^{n} dN_{ikl}(t) : k < l\right)$ can only jump once over $[t, t+dt)$. For notational convenience we let $\overline{Y}_{\cdot k}(t) = \sum_{i=1}^{n} \overline{Y}_{ik}(t)$ be the number of individuals under observation and at risk for transitions out of state $k = 0, 1$.

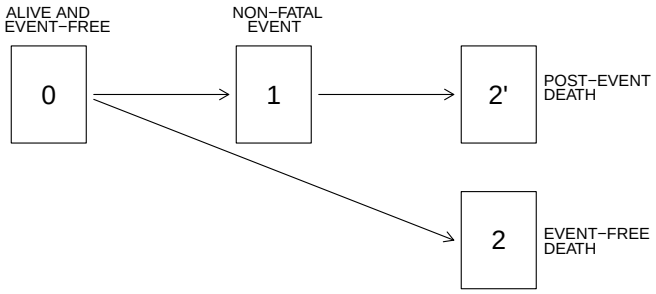


Figure D.1: Multistate diagram of a four-state illness-death process.

The 4×4 transition intensity matrix for this model is

$$d\mathbb{\Lambda}(t) = \begin{pmatrix} d\Lambda_{00}(t) & d\Lambda_{01}(t) & 0 & d\Lambda_{02}(t) \\ 0 & d\Lambda_{11}(t) & d\Lambda_{12'}(t) & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} ,$$

where $d\Lambda_{kl}(t) \geq 0$ for $k \neq l$ and $d\Lambda_{kk}(t) = -\sum_{l,l\neq k} d\Lambda_{kl}(t)$ for $(k,l) \in \{(0,1),(0,2),(1,2')\}$. Nelson-Aalen estimation in the full sample $\mathcal{O} = (\mathcal{O}_1, \ldots, \mathcal{O}_{i-1}, \mathcal{O}_i, \mathcal{O}_{i+1}, \ldots, \mathcal{O}_n)$ gives

$$d\widehat{\mathbb{\Lambda}}(t) = d\widehat{\mathbb{\Lambda}}(t;\widetilde{w}) = \begin{pmatrix} d\widehat{\Lambda}_{00}(t;\widetilde{w}) & d\widehat{\Lambda}_{01}(t;\widetilde{w}) & 0 & d\widehat{\Lambda}_{02}(t;\widetilde{w}) \\ 0 & d\widehat{\Lambda}_{11}(t;\widetilde{w}) & d\widehat{\Lambda}_{12'}(t;\widetilde{w}) & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} ,$$

with entries of the form

$$\begin{aligned} d\widehat{\Lambda}_{kl}(t;\widetilde{w}) &= \frac{\sum_{i=1}^{n} \overline{Y}_{ik}(t) dN_{ikl}(t)}{\overline{Y}_{\cdot k}(t)} , \qquad (k,l) \in \{(0,1),(0,2),(1,2')\} \\ d\widehat{\Lambda}_{00}(t;\widetilde{w}) &= -(d\widehat{\Lambda}_{01}(t;\widetilde{w}) + d\widehat{\Lambda}_{02}(t;\widetilde{w})) , \\ d\widehat{\Lambda}_{11}(t;\widetilde{w}) &= -d\widehat{\Lambda}_{12'}(t;\widetilde{w}) . \end{aligned}$$

The 4×4 derivative matrix of $d\widehat{\mathbb{A}}(t;w)$ evaluated at $w = \widetilde{w} = (1, \ldots, 1, 1, 1, \ldots 1)'$, i.e.

$$\left[\frac{\partial}{\partial w_i}\big(d\widehat{\mathbb{A}}(t;w)\big)\right]_{|_{w=\widetilde{w}}} = \begin{pmatrix} \dfrac{\partial d\widehat{\Lambda}_{00}(t;w)}{\partial w_i} & \dfrac{\partial d\widehat{\Lambda}_{01}(t;w)}{\partial w_i} & 0 & \dfrac{\partial d\widehat{\Lambda}_{02}(t;w)}{\partial w_i} \\ 0 & \dfrac{\partial d\widehat{\Lambda}_{11}(t;w)}{\partial w_i} & \dfrac{\partial d\widehat{\Lambda}_{12'}(t;w)}{\partial w_i} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}\Bigg|_{w=\widetilde{w}},$$

has $(k,l) \in \{(0,1),(0,2),(1,2')\}$ entries

$$\left[\frac{\partial d\widehat{\Lambda}_{kl}(t;w)}{\partial w_i}\right]_{|_{w=\widetilde{w}}} = \frac{\overline{Y}_{ik}(t)(dN_{ikl}(t) - d\widehat{\Lambda}_{kl}(t;\widetilde{w}))}{\overline{Y}_{\cdot k}(t)}$$

$$= \begin{cases} \dfrac{1 - d\widehat{\Lambda}_{kl}(t;\widetilde{w})}{\overline{Y}_{\cdot k}(t)} = \dfrac{\overline{Y}_{\cdot k}(t) - 1}{\overline{Y}_{\cdot k}(t)^2} & , \text{ if } \overline{Y}_{ik}(t) = 1, dN_{ikl}(t) = 1, dN_{i'kl}(t) = 0 \\ \dfrac{-d\widehat{\Lambda}_{kl}(t;\widetilde{w})}{\overline{Y}_{\cdot k}(t)} = \dfrac{-1}{\overline{Y}_{\cdot k}(t)^2} & , \text{ if } \overline{Y}_{ik}(t) = 1, dN_{ikl}(t) = 0, dN_{i'kl}(t) = 1 \\ 0 & , \text{ if } \overline{Y}_{ik}(t) = 1, dN_{ikl}(t) = 0, dN_{i'kl}(t) = 0 \\ 0 & , \text{ if } \overline{Y}_{ik}(t) = 0 \end{cases} \tag{D.3.1}$$

for some individual $i' \neq i$. The entries of individual i's derivative matrix can thus be of three unique forms. If individual i is observed to make a $k \longrightarrow l$ transition at time t, then the (k,l) entry is equal to $(\overline{Y}_{\cdot k}(t) - 1)/\overline{Y}_{\cdot k}(t)^2$; see first case in (D.3.1). If an individual other than i is observed to make a $k \longrightarrow l$ transition at time t but i is still at-risk for a $k \longrightarrow l$ transition at t (i.e. $\overline{Y}_{ik}(t) = 1$), the entry equals $-1/\overline{Y}_{\cdot k}(t)^2$; see second case. Otherwise, the entry is 0; see third and fourth case.

We now let $0 < t_{(1)} < \ldots < t_{(s)} < \ldots < t_{(S)} \leq \tau$ be the S distinct ordered observed transition times (of any type) less than τ, $s = 1, \ldots, S$, and let

$$\widehat{I}_{ik}(t_{(s)}) = \left.\frac{\partial \widehat{P}_{0k}(0, t_{(s)}; w)}{\partial w_i}\right|_{w=\widetilde{w}} \qquad i = 1, \ldots, n,\ k = 0, 1, 2, 2'$$

be individual i's influence value on $\widehat{P}_{0k}(0, t_{(s)})$ evaluated at $w = \widetilde{w} = (1, \ldots, 1, 1, 1, \ldots, 1)'$. We next derive the explicit expressions for $\widehat{I}_{ik}(t_{(s)})$ when $s = 1, 2$, followed by an illustrative example in Appendix D.4. This example demonstrates how the influence values can be easily obtained using standard software for survival and event history analysis (Therneau, 2024).

D.3.1 Expressions for $\widehat{I}_{ik}(t_{(s)})$ at $t_{(1)}$

Since $\partial\widehat{\mathbb{P}}(0, t_{(1)}; w)/\partial w_i = \partial d\widehat{\mathbb{A}}(t_{(1)}; w)/\partial w_i$ for $s = 1$, we obtain

$$\begin{aligned}
\widehat{I}_{i1}(t_{(1)}) &= \left[\frac{\partial d\widehat{\Lambda}_{01}(t_{(1)}; w)}{\partial w_i}\right]_{|_{w=\widetilde{w}}}, \qquad \widehat{I}_{i2}(t_{(1)}) = \left[\frac{\partial d\widehat{\Lambda}_{02}(t_{(1)}; w)}{\partial w_i}\right]_{|_{w=\widetilde{w}}}, \\
\widehat{I}_{i0}(t_{(1)}) &= \left[\frac{\partial d\widehat{\Lambda}_{00}(t_{(1)}; w)}{\partial w_i}\right]_{|_{w=\widetilde{w}}} = -(\widehat{I}_{i1}(t_{(1)}) + \widehat{I}_{i2}(t_{(1)})), \\
\widehat{I}_{i2'}(t_{(1)}) &= 0\ .
\end{aligned}$$

D.3.2 Expressions for $\widehat{I}_{ik}(t_{(s)})$ at $t_{(2)}$

Since

$$\frac{\partial\widehat{\mathbb{P}}(0, t_{(2)}; w)}{\partial w_i} = \frac{\partial\widehat{\mathbb{P}}(0, t_{(1)}; w)}{\partial w_i}(\mathbb{I} + d\widehat{\mathbb{A}}(t_{(2)}; w)) + \widehat{\mathbb{P}}(0, t_{(1)}; w)\frac{\partial d\widehat{\mathbb{A}}(t_{(2)}; w)}{\partial w_i}\ ,$$

we obtain

$$\widehat{I}_{i0}(t_{(2)}) = \widehat{I}_{i0}(t_{(1)})\big(1 + d\widehat{\Lambda}_{00}(t_{(2)}; \widetilde{w})\big) + \big(1 + d\widehat{\Lambda}_{00}(t_{(1)}; \widetilde{w})\big)\left[\frac{\partial d\widehat{\Lambda}_{00}(t_{(2)}; w)}{\partial w_i}\right]_{|_{w=\widetilde{w}}}$$

$$\widehat{I}_{i1}(t_{(2)}) = \widehat{I}_{i0}(t_{(1)})d\widehat{\Lambda}_{01}(t_{(2)};\widetilde{w}) + (1 + d\widehat{\Lambda}_{00}(t_{(1)};\widetilde{w}))\left[\frac{\partial d\widehat{\Lambda}_{01}(t_{(2)};w)}{\partial w_i}\right]_{|_{w=\widetilde{w}}}$$
$$+ d\widehat{\Lambda}_{01}(t_{(1)};\widetilde{w})\left[\frac{\partial d\widehat{\Lambda}_{11}(t_{(2)};w)}{\partial w_i}\right]_{|_{w=\widetilde{w}}} + \widehat{I}_{i1}(t_{(1)})(1 + d\widehat{\Lambda}_{11}(t_{(2)};\widetilde{w}))$$
$$\widehat{I}_{i2}(t_{(2)}) = \widehat{I}_{i0}(t_{(1)})d\widehat{\Lambda}_{02}(t_{(2)};\widetilde{w}) + \big(1 + d\widehat{\Lambda}_{00}(t_{(1)};\widetilde{w})\big)\left[\frac{\partial d\widehat{\Lambda}_{02}(t_{(2)};w)}{\partial w_i}\right]_{|_{w=\widetilde{w}}} + \widehat{I}_{i2}(t_{(1)})$$
$$\widehat{I}_{i2'}(t_{(2)}) = \widehat{I}_{i1}(t_{(1)})d\widehat{\Lambda}_{12'}(t_{(2)};\widetilde{w}) + d\widehat{\Lambda}_{01}(t_{(1)};\widetilde{w})\left[\frac{\partial d\widehat{\Lambda}_{12'}(t_{(2)};w)}{\partial w_i}\right]_{|_{w=\widetilde{w}}} + \widehat{I}_{i2'}(t_{(1)})\ .$$

We now take a closer look at $\widehat{I}_{i1}(t_{(2)})$ and consider four different cases:

1. If there is neither a $0 \longrightarrow 1$ transition at time $t_{(1)}$ nor at time $t_{(2)}$ (i.e. $d\widehat{\Lambda}_{01}(t_{(1)};\widetilde{w}) = d\widehat{\Lambda}_{01}(t_{(2)};\widetilde{w}) = 0, \widehat{I}_{i1}(t_{(1)}) = 0$), then $\widehat{I}_{i1}(t_{(2)}) = 0$.

2. If a $0 \longrightarrow 1$ transition happens at time $t_{(1)}$ but not at $t_{(2)}$ (i.e. $d\widehat{\Lambda}_{01}(t_{(2)};\widetilde{w}) = 0$), then

$$\widehat{I}_{i1}(t_{(2)}) = d\widehat{\Lambda}_{01}(t_{(1)};\widetilde{w})\left[\frac{\partial d\widehat{\Lambda}_{11}(t_{(2)};w)}{\partial w_i}\right]_{|_{w=\widetilde{w}}} + \widehat{I}_{i1}(t_{(1)})(1 + d\widehat{\Lambda}_{11}(t_{(2)};\widetilde{w}))$$

$$= \begin{cases} d\widehat{\Lambda}_{01}(t_{(1)};\widetilde{w})(\overline{Y}_{\cdot 1}(t_{(2)}) - 1)/\overline{Y}_{\cdot 1}(t_{(2)})^2 & , \text{ if } dN_{i01}(t_{(1)}) = 1, dN_{i12'}(t_{(2)}) = 1 \\ \quad + \widehat{I}_{i1}(t_{(1)})(1 + d\widehat{\Lambda}_{11}(t_{(2)};\widetilde{w})) & \\ \widehat{I}_{i1}(t_{(1)}) & , \text{ if } dN_{i01}(t_{(1)}) = 1, dN_{i12'}(t_{(2)}) = 0, \\ & \quad dN_{i'02}(t_{(2)}) = 1 \\ \widehat{I}_{i1}(t_{(1)})(1 + d\widehat{\Lambda}_{11}(t_{(2)};\widetilde{w})) & , \text{ if } dN_{i'01}(t_{(1)}) = 1, dN_{i'12'}(t_{(2)}) = 1 \\ \widehat{I}_{i1}(t_{(1)}) & , \text{ if } dN_{i'01}(t_{(1)}) = 1, dN_{i'12'}(t_{(2)}) = 0, \\ & \quad dN_{i02}(t_{(2)}) = 1 \\ \widehat{I}_{i1}(t_{(1)}) & , \text{ if } dN_{i'01}(t_{(1)}) = 1, dN_{i'12'}(t_{(2)}) = 0, \\ & \quad dN_{i''02}(t_{(2)}) = 1 \end{cases}.$$

3. If a $0 \longrightarrow 1$ transition happens at time $t_{(2)}$ but not at time $t_{(1)}$ (i.e. $d\widehat{\Lambda}_{01}(t_{(1)}; \widetilde{w}) = 0$ and $\widehat{I}_{i1}(t_{(1)}) = 0$), then

$$\widehat{I}_{i1}(t_{(2)}) = \widehat{I}_{i0}(t_{(1)}) d\widehat{\Lambda}_{01}(t_{(2)}; \widetilde{w}) + (1 + d\widehat{\Lambda}_{00}(t_{(1)}; \widetilde{w})) \left[\frac{\partial d\widehat{\Lambda}_{01}(t_{(2)}; w)}{\partial w_i} \right]_{|_{w=\widetilde{w}}} + \widehat{I}_{i1}(t_{(1)})$$

$$= \begin{cases} \widehat{I}_{i0}(t_{(1)}) d\widehat{\Lambda}_{01}(t_{(2)}; \widetilde{w}) + \widehat{I}_{i1}(t_{(1)}) & , \text{ if } dN_{i02}(t_{(1)}) = 1, dN_{i'01}(t_{(2)}) = 1 \\ \widehat{I}_{i0}(t_{(1)}) d\widehat{\Lambda}_{01}(t_{(2)}; \widetilde{w}) + \widehat{I}_{i1}(t_{(1)}) & \\ \quad + (1 + d\widehat{\Lambda}_{00}(t_{(1)}; \widetilde{w})) \dfrac{(\overline{Y}_{\cdot 0}(t_{(2)}) - 1)}{\overline{Y}_{\cdot 0}(t_{(2)})^2} & , \text{ if } dN_{i'02}(t_{(1)}) = 1, dN_{i01}(t_{(2)}) = 1 \\ \widehat{I}_{i0}(t_{(1)}) d\widehat{\Lambda}_{01}(t_{(2)}; \widetilde{w}) + \widehat{I}_{i1}(t_{(1)}) & \\ \quad - (1 + d\widehat{\Lambda}_{00}(t_{(1)}; \widetilde{w})) / \overline{Y}_{\cdot 0}(t_{(2)})^2 & , \text{ if } dN_{i'02}(t_{(1)}) = 1, dN_{i''01}(t_{(2)}) = 1 \end{cases} .$$

4. If a $0 \longrightarrow 1$ transition happens at both times $t_{(1)}$ and $t_{(2)}$, then

$$\widehat{I}_{i1}(t_{(2)}) = \begin{cases} \widehat{I}_{i0}(t_{(1)}) d\widehat{\Lambda}_{01}(t_{(2)}; \widetilde{w}) + \widehat{I}_{i1}(t_{(1)}) & , \text{ if } dN_{i01}(t_{(1)}) = 1, dN_{i'01}(t_{(2)}) = 1 \\ \widehat{I}_{i0}(t_{(1)}) d\widehat{\Lambda}_{01}(t_{(2)}; \widetilde{w}) + \widehat{I}_{i1}(t_{(1)}) & \\ \quad + (1 + d\widehat{\Lambda}_{00}(t_{(1)}; \widetilde{w})) \dfrac{(\overline{Y}_{\cdot 0}(t_{(2)}) - 1)}{\overline{Y}_{\cdot 0}(t_{(2)})^2} & , \text{ if } dN_{i'01}(t_{(1)}) = 1, dN_{i01}(t_{(2)}) = 1 \\ \widehat{I}_{i0}(t_{(1)}) d\widehat{\Lambda}_{01}(t_{(2)}; \widetilde{w}) + \widehat{I}_{i1}(t_{(1)}) & \\ \quad - \dfrac{(1 + d\widehat{\Lambda}_{00}(t_{(1)}; \widetilde{w}))}{\overline{Y}_{\cdot 0}(t_{(2)})^2} & , \text{ if } dN_{i'01}(t_{(1)}) = 1, dN_{i''01}(t_{(2)}) = 1 \end{cases} .$$

Similar expressions can be obtained for $\widehat{I}_{i0}(t_{(2)}), \widehat{I}_{i2'}(t_{(2)})$ and $\widehat{I}_{i2}(t_{(2)})$.

D.4 Software for the infinitesimal jackknife procedure: an illustrative example for the illness-death model

The objective of this section is to illustrate that the `survfit()` function of the `R` package `survival` provides the influence values $\widehat{I}_{ik}(t_{(s)})$ needed to calculate $\widehat{\text{var}}_{\text{IJK}}(\widehat{\beta}(\tau))$ (Therneau, 2024). We consider for illustration the following data generated according to the extended illness-death model in Figure D.1.

```
simdataID

##    id   start    stop          event
## 2   1 0.00000 0.22532 eventFreeDeath
## 4   2 0.00000 0.25898       nonFatal
## 5   2 0.25898 1.00000           cens
## 7   3 0.00000 1.00000           cens
## 9   4 0.00000 0.40288 eventFreeDeath
## 10  5 0.00000 0.04080       nonFatal
## 11  5 0.04080 1.00000           cens
## 14  6 0.00000 1.00000           cens
## 15  7 0.00000 0.02646       nonFatal
## 16  7 0.02646 0.20500           cens
## 17  8 0.00000 0.02978       nonFatal
## 18  8 0.02978 0.68341 postEventDeath
## 19  9 0.00000 0.45689       nonFatal
## 20  9 0.45689 1.00000           cens
## 25 10 0.00000 0.52290       nonFatal
## 26 10 0.52290 1.00000           cens
## 27 11 0.00000 0.53147       nonFatal
## 28 11 0.53147 1.00000           cens
## 29 12 0.00000 1.00000           cens
```

This illustrative dataset contains $n = 12$ observations, with a non-fatal event (i.e. $0 \to 1$ transition) experienced by individual $i = 7$ at time $t_{(1)} = 0.02646$. Since $s = 1$ and $P(Z(0) = 0) = 1$, we have $\overline{Y}_{\cdot 0}(t_{(1)}) = 12$ and the $\widehat{I}_{ik}(t_{(1)})$ values are obtained as follows (c.f. Section D.3.1):

$$\begin{aligned}
\widehat{I}_{i1}(t_{(1)}) &= \begin{cases} 0.07639 = (12-1)/12^2 & , \text{ if } i = 7 \\ -0.00694 = -1/12^2 & , \text{ if } i \neq 7 \end{cases} \\
\widehat{I}_{i2}(t_{(1)}) &= 0 \\
\widehat{I}_{i0}(t_{(1)}) &= \begin{cases} -0.07639 = (12-1)/12^2 & , \text{ if } i = 7 \\ 0.00694 = -1/12^2 & , \text{ if } i \neq 7 \end{cases} \\
\widehat{I}_{i2'}(t_{(1)}) &= 0 \ .
\end{aligned}$$

At $t_{(2)} = 0.02978$ individual $i = 8$ is observed to experience a non-fatal event (i.e. $0 \to 1$ transition). Thus, according to the fourth case in Section D.3.2, we find

$$\begin{aligned}
\widehat{I}_{i1}(t_{(2)}) &= \begin{cases} 0.06944 = -\widehat{I}_{i1}(t_{(1)})/11 + \widehat{I}_{i1}(t_{(1)}) & , \text{ if } i = 7 \\ 0.06944 = -\widehat{I}_{i1}(t_{(1)})/11 + \widehat{I}_{i1}(t_{(1)}) + (1-1/12)(11-1)/11^2 & , \text{ if } i = 8 \\ -0.01389 = -\widehat{I}_{i1}(t_{(1)})/11 + \widehat{I}_{i1}(t_{(1)}) - (1-1/12)/11^2 & , \text{ if } i \neq 7, i \neq 8 \end{cases} \\
\widehat{I}_{i2}(t_{(2)}) &= 0 \\
\widehat{I}_{i0}(t_{(2)}) &= \begin{cases} -0.06944 = -\widehat{I}_{i1}(t_{(1)})(1-1/11) & , \text{ if } i = 7 \\ -0.06944 = -\widehat{I}_{i1}(t_{(1)})(1-1/11) - (1-1/12)(11-1)/11^2 & , \text{ if } i = 8 \\ 0.01389 = -\widehat{I}_{i1}(t_{(1)})(1-1/11) + (1-1/12)/11^2 & , \text{ if } i \neq 7, i \neq 8 \end{cases} \\
\widehat{I}_{i2'}(t_{(2)}) &= 0 \ .
\end{aligned}$$

The `survfit()` function below returns the influence values $\widehat{I}_{ik}(t_{(s)})$ for any $k = 0, 1, 2, 2', s = 0, 1, \ldots, S$ and $i = 1, \ldots, n$ (Therneau, 2024), which are in agreement with the values calculated by hand in Sections D.3.1 and D.3.2.

Listing D.1: R code

```
1 library(survival)
2 fit.surv <- survfit(Surv(start, stop, event) ~ 1, data=
    simdataID, id=id, stype=1, influence=TRUE)
3 fit.surv$influence.pstate
```

Influence matrix $\widehat{\mathbb{J}}_0 = \left[\widehat{I}_{i0}(t_{(s)})\right]$ for state $k = 0$ and time points $s = 0, 1, 2, 3$:

```
##    0  0.02646  0.02978   0.0408
## 1  0  0.00694  0.01389  0.02083
## 2  0  0.00694  0.01389  0.02083
## 3  0  0.00694  0.01389  0.02083
## 4  0  0.00694  0.01389  0.02083
## 5  0  0.00694  0.01389 -0.06250
## 6  0  0.00694  0.01389  0.02083
## 7  0 -0.07639 -0.06944 -0.06250
## 8  0  0.00694 -0.06944 -0.06250
## 9  0  0.00694  0.01389  0.02083
## 10 0  0.00694  0.01389  0.02083
## 11 0  0.00694  0.01389  0.02083
## 12 0  0.00694  0.01389  0.02083
```

Influence matrix $\widehat{\mathbb{J}}_1 = \left[\widehat{I}_{i1}(t_{(s)})\right]$ for state $k = 1$ and time points $s = 0, 1, 2, 3$:

```
##    0  0.02646  0.02978   0.0408
## 1  0 -0.00694 -0.01389 -0.02083
## 2  0 -0.00694 -0.01389 -0.02083
## 3  0 -0.00694 -0.01389 -0.02083
## 4  0 -0.00694 -0.01389 -0.02083
## 5  0 -0.00694 -0.01389  0.06250
## 6  0 -0.00694 -0.01389 -0.02083
## 7  0  0.07639  0.06944  0.06250
```

```
## 8  0 -0.00694  0.06944  0.06250
## 9  0 -0.00694 -0.01389 -0.02083
## 10 0 -0.00694 -0.01389 -0.02083
## 11 0 -0.00694 -0.01389 -0.02083
## 12 0 -0.00694 -0.01389 -0.02083
```

Influence matrix $\widehat{\mathbb{J}}_{2'} = \left[\widehat{I}_{i2'}(t_{(s)})\right]$ for state $k = 0$ and time points $s = 0, 1, 2, 3$:

```
##    0 0.02646 0.02978 0.0408
## 1  0       0       0      0
## 2  0       0       0      0
## 3  0       0       0      0
## 4  0       0       0      0
## 5  0       0       0      0
## 6  0       0       0      0
## 7  0       0       0      0
## 8  0       0       0      0
## 9  0       0       0      0
## 10 0       0       0      0
## 11 0       0       0      0
## 12 0       0       0      0
```

Influence matrix $\widehat{\mathbb{J}}_{2} = \left[\widehat{I}_{i2}(t_{(s)})\right]$ for state $k = 2$ and time points $s = 0, 1, 2, 3$:

```
##    0 0.02646 0.02978 0.0408
## 1  0       0       0      0
## 2  0       0       0      0
## 3  0       0       0      0
## 4  0       0       0      0
## 5  0       0       0      0
## 6  0       0       0      0
## 7  0       0       0      0
```

```
## 8  0       0       0       0
## 9  0       0       0       0
## 10 0       0       0       0
## 11 0       0       0       0
## 12 0       0       0       0
```

D.5 Additional results for the simulation study of Section 5.4.2

Here we provide additional results for the simulation study of Section 5.4.2. Tables D.3 and D.4 summarize the simulation results for the Markov models when $\pi_T = 0.6, \pi_T^1 = 0.7$ and $\pi_D = 0.4$. With this parameter setting we mimic a clinical trial in which $40\%, 20\%$ and 40% of individuals in the control group are in states $0, 1$ and 2 by the administrative end of study $\tau = 1$, respectively; since $\pi_D = 0.4$, 22% died post-event and 18% died event-free. Results for the mixed-Markov models are presented in Tables D.5-D.6.

		MARKOV PROCESSES														
		$n = 100$					$n = 250$					$n = 1000$				
$[e^{\gamma_{01}}, e^{\gamma_{02}}, e^{\gamma_{12}}]$	$\beta(\tau)$	BIAS	ESE	ASE	RR	CP	BIAS	ESE	ASE	RR	CP	BIAS	ESE	ASE	RR	CP
							$\pi_r = 0.20$ and $\exp(\gamma_r) = 1$									
1,1,1	0.0000	0.0005	0.0623	0.0617	0.0551	0.9449	0.0003	0.0387	0.0392	0.0486	0.9514	-0.0003	0.0197	0.0196	0.0517	0.9483
1,1,0.8	0.0066	0.0005	0.0617	0.0611	0.0675	0.9449	0.0003	0.0383	0.0388	0.0683	0.9508	-0.0003	0.0195	0.0195	0.0949	0.9487
0.9,1,1	0.0120	0.0001	0.0624	0.0616	0.0778	0.9441	0.0004	0.0391	0.0391	0.0918	0.9503	-0.0002	0.0195	0.0196	0.1467	0.9511
1,0.8,1	0.0169	0.0002	0.0621	0.0610	0.0907	0.9438	0.0004	0.0386	0.0387	0.1167	0.9482	-0.0002	0.0195	0.0194	0.2173	0.9498
0.75, 1.1, 1.2	0.0174	0.0003	0.0626	0.0622	0.0891	0.9471	0.0004	0.0395	0.0395	0.1171	0.9468	-0.0002	0.0198	0.0198	0.2212	0.9510
1,0.8,0.8	0.0236	0.0002	0.0615	0.0603	0.1099	0.9441	0.0004	0.0382	0.0383	0.1565	0.9480	-0.0002	0.0192	0.0192	0.3336	0.9499
0.75,1,1.2	0.0262	0.0003	0.0626	0.0618	0.1134	0.9452	0.0004	0.0392	0.0392	0.1673	0.9479	-0.0001	0.0198	0.0197	0.3796	0.9480
0.75,1,1	0.0308	0.0003	0.0621	0.0613	0.1295	0.9456	0.0004	0.0389	0.039	0.2002	0.9485	-0.0001	0.0197	0.0195	0.4668	0.9484
0.75,0.9,1.2	0.0353	0.0004	0.0628	0.0613	0.1516	0.9405	0.0005	0.0387	0.039	0.2302	0.9499	-0.0001	0.0197	0.0195	0.5665	0.9497
0.75,1,0.8	0.0359	0.0003	0.0616	0.0608	0.1489	0.9460	0.0004	0.0386	0.0386	0.2394	0.9493	-0.0001	0.0195	0.0194	0.5812	0.9487
0.75,0.9,0.9	0.0423	0.0003	0.0621	0.0606	0.1808	0.9406	0.0005	0.0383	0.0385	0.2957	0.9507	-0.0001	0.0195	0.0193	0.7025	0.9503
0.75,0.8,1	0.0490	0.0000	0.0611	0.0604	0.2053	0.9455	0.0005	0.0381	0.0384	0.3613	0.9495	0.0000	0.0194	0.0192	0.8156	0.9480
							$\pi_r = 0.40$ and $\exp(\gamma_r) = 1$									
1,1,1	0.0000	0.0003	0.0652	0.0640	0.0540	0.9460	0.0004	0.0405	0.0406	0.0525	0.9475	-0.0003	0.0205	0.0204	0.0467	0.9533
1,1,0.8	0.0066	0.0004	0.0646	0.0634	0.0671	0.9448	0.0004	0.0401	0.0403	0.0688	0.9475	-0.0003	0.0203	0.0202	0.0904	0.9533
0.9,1,1	0.0120	0.0001	0.0656	0.0639	0.0802	0.9418	0.0005	0.0405	0.0406	0.0871	0.9499	-0.0002	0.0206	0.0203	0.1449	0.9499
1,0.8,1	0.0169	0.0004	0.0644	0.0633	0.0884	0.9436	0.0005	0.0403	0.0402	0.1145	0.9451	-0.0001	0.0203	0.0201	0.2113	0.9473
0.75, 1.1, 1.2	0.0174	0.0004	0.0653	0.0645	0.0900	0.9460	0.0003	0.0411	0.0410	0.1142	0.9489	-0.0002	0.0207	0.0205	0.2144	0.9507
1,0.8,0.8	0.0236	0.0005	0.0637	0.0626	0.1083	0.9444	0.0005	0.0399	0.0398	0.1499	0.9455	-0.0001	0.0201	0.0199	0.3266	0.9475
0.75,1,1.2	0.0262	0.0007	0.0654	0.0641	0.1173	0.9398	0.0002	0.0409	0.0407	0.1628	0.9476	-0.0001	0.0204	0.0204	0.3614	0.9514
0.75,1,1	0.0308	0.0006	0.0649	0.0637	0.1311	0.9404	0.0002	0.0406	0.0404	0.1930	0.9480	-0.0001	0.0203	0.0203	0.4486	0.9512
0.75,0.9,1.2	0.0353	0.0008	0.0648	0.0637	0.1504	0.9414	0.0003	0.0407	0.0404	0.2278	0.9482	-0.0002	0.0204	0.0203	0.5372	0.9514
0.75,1,0.8	0.0359	0.0006	0.0644	0.0631	0.1507	0.9401	0.0002	0.0403	0.0401	0.2325	0.9482	-0.0001	0.0201	0.0201	0.5532	0.9507
0.75,0.9,0.9	0.0423	0.0008	0.0641	0.0629	0.1770	0.9415	0.0004	0.0402	0.0400	0.2892	0.9481	-0.0002	0.0202	0.0200	0.6757	0.9511
0.75,0.8,1	0.0490	0.0001	0.0643	0.0627	0.2027	0.9433	0.0004	0.0397	0.0398	0.3429	0.9491	-0.0002	0.0201	0.0200	0.7886	0.9441

Table D.3: Empirical properties of the utility-based estimator $\widehat{\beta}(\tau)$ and associated tests of $H_0 : \beta(\tau) \leq 0$ vs. $H_1 : \beta(\tau) > 0$ when the illness-death process is Markov with intensities $\lambda_{kl}(t|X) = \lambda_{kl0} \exp(\gamma_{kl} X)$; $\tau = 1, \pi_T = 0.6, \pi_T^1 = 0.7, \pi_D = 0.4, \pi_r = P(C_r \leq \tau | X = 0) = 0.2$ and $0.4, \exp(\gamma_r) = 1, u_0 = 1, u_1 = 0.5, u_2 = 0, n_{sim} = 10000$ simulation runs.

		MARKOV PROCESSES														
		$n = 100$					$n = 250$					$n = 1000$				
$[e^{\gamma_{01}}, e^{\gamma_{02}}, e^{\gamma_{12}}]$	$\beta(\tau)$	BIAS	ESE	ASE	RR	CP	BIAS	ESE	ASE	RR	CP	BIAS	ESE	ASE	RR	CP
$\pi_r = 0.20$ and $\exp(\gamma_r) = 0.85$																
1,1,1	0.0000	0.0003	0.0623	0.0615	0.0536	0.9464	0.0001	0.0390	0.0391	0.0483	0.9517	-0.0003	0.0197	0.0196	0.0524	0.9476
1,1,0.8	0.0066	0.0003	0.0617	0.0610	0.0650	0.9454	0.0001	0.0386	0.0387	0.0703	0.9520	-0.0003	0.0195	0.0194	0.0979	0.9471
0.9,1,1	0.0120	0.0001	0.0620	0.0614	0.0773	0.9461	0.0003	0.0386	0.0390	0.0898	0.9497	-0.0002	0.0004	0.0196	0.1460	0.9519
1,0.8,1	0.0169	0.0001	0.0619	0.0608	0.0920	0.9430	0.0004	0.0385	0.0386	0.1143	0.9488	-0.0002	0.0195	0.0194	0.2214	0.9470
0.75, 1.1, 1.2	0.0174	0.0004	0.0628	0.0620	0.0923	0.9432	0.0004	0.0392	0.0394	0.1174	0.9512	-0.0002	0.0200	0.0198	0.2251	0.9489
1,0.8,0.8	0.0236	0.0001	0.0613	0.0602	0.1120	0.9430	0.0004	0.0381	0.0382	0.1519	0.9481	-0.0002	0.0193	0.0192	0.3340	0.9476
0.75,1,1.2	0.0262	0.0004	0.0630	0.0616	0.1207	0.9418	0.0003	0.0390	0.0392	0.1688	0.9487	-0.0001	0.0197	0.0196	0.3764	0.9491
0.75,1,1	0.0308	0.0004	0.0626	0.0612	0.1351	0.9422	0.0004	0.0388	0.0389	0.1971	0.9488	-0.0001	0.0196	0.0195	0.4702	0.9494
0.75,0.9,1.2	0.0353	0.0004	0.0624	0.0612	0.1483	0.9411	0.0004	0.0386	0.0389	0.2323	0.9504	-0.0001	0.0196	0.0195	0.5679	0.9479
0.75,1,0.8	0.0359	0.0003	0.0621	0.0607	0.1536	0.9427	0.0004	0.0385	0.0386	0.2417	0.9494	-0.0001	0.0194	0.0193	0.5864	0.9499
0.75,0.9,0.9	0.0423	0.0004	0.0617	0.0605	0.1783	0.9412	0.0004	0.0382	0.0384	0.2924	0.9498	-0.0001	0.0194	0.0193	0.7116	0.9486
0.75,0.8,1	0.0490	-0.0001	0.0614	0.0603	0.2059	0.9422	0.0005	0.0380	0.0383	0.3568	0.9514	-0.0001	0.0194	0.0192	0.8154	0.9455
$\pi_r = 0.40$ and $\exp(\gamma_r) = 0.85$																
1,1,1	0.0000	0.0006	0.0649	0.0637	0.0545	0.9455	0.0002	0.0406	0.0404	0.0535	0.9465	-0.0003	0.0203	0.0203	0.0489	0.9511
1,1,0.8	0.0066	0.0005	0.0643	0.0631	0.0659	0.9454	0.0002	0.0402	0.0401	0.0705	0.9467	-0.0003	0.0201	0.0201	0.0908	0.9512
0.9,1,1	0.0120	0.0004	0.0645	0.0636	0.0782	0.9461	0.0003	0.0403	0.0404	0.0898	0.9504	-0.0002	0.0203	0.0202	0.1421	0.9504
1,0.8,1	0.0169	0.0007	0.0643	0.0630	0.0900	0.9443	0.0004	0.0397	0.0400	0.1113	0.9483	-0.0001	0.0203	0.0200	0.2111	0.9481
0.75, 1.1, 1.2	0.0174	0.0004	0.0652	0.0642	0.0936	0.9411	0.0004	0.0406	0.0408	0.1131	0.9502	-0.0002	0.0204	0.0204	0.2118	0.9514
1,0.8,0.8	0.0236	0.0007	0.0636	0.0623	0.1083	0.9436	0.0004	0.0393	0.0396	0.1469	0.9477	-0.0001	0.0200	0.0198	0.3228	0.9479
0.75,1,1.2	0.0262	0.0007	0.0651	0.0638	0.1142	0.9417	0.0002	0.0404	0.0405	0.1604	0.9500	-0.0001	0.0203	0.0203	0.3600	0.9486
0.75,1,1	0.0308	0.0006	0.0647	0.0633	0.1260	0.9426	0.0002	0.0401	0.0402	0.1898	0.9501	-0.0001	0.0202	0.0202	0.4524	0.9490
0.75,0.9,1.2	0.0353	0.0007	0.0647	0.0633	0.1453	0.9399	0.0005	0.0403	0.0402	0.2262	0.9497	-0.0001	0.0204	0.0202	0.5414	0.9479
0.75,1,0.8	0.0359	0.0006	0.0642	0.0628	0.1453	0.9417	0.0003	0.0398	0.0399	0.2310	0.9497	-0.0001	0.0200	0.0200	0.5583	0.9501
0.75,0.9,0.9	0.0423	0.0006	0.0639	0.0626	0.1735	0.9416	0.0005	0.0398	0.0398	0.2888	0.9499	-0.0001	0.0201	0.0199	0.6805	0.9482
0.75,0.8,1	0.0490	0.0003	0.0632	0.0624	0.1990	0.9431	0.0003	0.0394	0.0396	0.3453	0.9493	-0.0001	0.0202	0.0199	0.7866	0.9499

Table D.4: Empirical properties of the utility-based estimator $\widehat{\beta}(\tau)$ and associated tests of $H_0 : \beta(\tau) \leq 0$ vs. $H_1 : \beta(\tau) > 0$ when the illness-death process is Markov with intensities $\lambda_{kl}(t|X) = \lambda_{kl0} \exp(\gamma_{kl} X)$; $\tau = 1, \pi_T = 0.6, \pi_T^1 = 0.7, \pi_D = 0.4, \pi_r = P(C_r \leq \tau | X = 0) = 0.2$ and $0.4, \exp(\gamma_r) = 0.85, u_0 = 1, u_1 = 0.5, u_2 = 0, n_{sim} = 10000$ simulation runs.

		MIXED-MARKOV PROCESSES														
		$n = 100$					$n = 250$					$n = 1000$				
$[e^{\gamma_{01}}, e^{\gamma_{02}}, e^{\gamma_{12}}]$	$\beta(\tau)$	BIAS	ESE	ASE	RR	CP	BIAS	ESE	ASE	RR	CP	BIAS	ESE	ASE	RR	CP
$\phi = 0.5:\ \pi_r = 0.20$ and $\exp(\gamma_r) = 1$																
1,1,1	0.0000	-0.0008	0.0642	0.0643	0.0486	0.9514	-0.0002	0.0414	0.0408	0.0517	0.9483	0.0000	0.0205	0.0205	0.0514	0.9486
1,1,0.8	0.0066	-0.0007	0.0636	0.0638	0.0595	0.9510	-0.0002	0.0410	0.0404	0.0723	0.9481	0.0000	0.0204	0.0203	0.0945	0.9489
0.9,1,1	0.0106	-0.0007	0.0647	0.0642	0.0708	0.9473	-0.0003	0.0411	0.0408	0.0843	0.9482	0.0000	0.0204	0.0204	0.1306	0.9499
0.75, 1.1, 1.2	0.0146	-0.0009	0.0656	0.0649	0.0809	0.9470	-0.0001	0.0415	0.0412	0.0999	0.9446	0.0000	0.0207	0.0206	0.1764	0.9482
1,0.8,1	0.0161	-0.0006	0.0643	0.0636	0.0879	0.9446	-0.0003	0.0407	0.0404	0.1083	0.9501	0.0001	0.0201	0.0202	0.1952	0.9491
1,0.8,0.8	0.0228	-0.0006	0.0637	0.0630	0.1033	0.9446	-0.0003	0.0403	0.0400	0.1424	0.9498	0.0001	0.0200	0.0200	0.3002	0.9479
0.75,1,1.2	0.0231	-0.0010	0.0655	0.0645	0.1007	0.9449	-0.0002	0.0413	0.0409	0.1423	0.9497	0.0000	0.0206	0.0205	0.3023	0.9489
0.75,1,1	0.0276	-0.0010	0.0651	0.0641	0.1144	0.9453	-0.0002	0.0411	0.0407	0.1694	0.9488	0.0000	0.0205	0.0204	0.3847	0.9491
0.75,0.9,1.2	0.0319	-0.0010	0.0648	0.0641	0.1244	0.9455	-0.0004	0.0407	0.0407	0.1950	0.9495	-0.0001	0.0204	0.0204	0.4652	0.9488
0.75,1,0.8	0.0329	-0.0010	0.0645	0.0636	0.1332	0.9457	-0.0002	0.0408	0.0404	0.2042	0.9489	0.0000	0.0203	0.0202	0.4906	0.9489
0.75,0.9,0.9	0.0390	-0.0010	0.0640	0.0634	0.1508	0.9450	-0.0004	0.0403	0.0402	0.2478	0.9502	-0.0001	0.0202	0.0202	0.6147	0.9491
0.75,0.8,1	0.0455	-0.0013	0.0641	0.0632	0.1737	0.9471	-0.0002	0.0407	0.0401	0.3070	0.9459	-0.0001	0.0201	0.0201	0.7308	0.9484
$\phi = 0.5:\ \pi_r = 0.40$ and $\exp(\gamma_r) = 1$																
1,1,1	0.0000	-0.0003	0.0675	0.0666	0.0552	0.9448	-0.0004	0.0420	0.0422	0.0488	0.9512	0.0000	0.0212	0.0212	0.0521	0.9479
1,1,0.8	0.0066	-0.0003	0.0669	0.0660	0.0667	0.9461	-0.0004	0.0416	0.0419	0.0648	0.9502	0.0000	0.0210	0.0210	0.0917	0.9483
0.9,1,1	0.0106	-0.0003	0.0682	0.0665	0.0744	0.9460	-0.0004	0.0425	0.0422	0.0842	0.9467	0.0001	0.0209	0.0212	0.1254	0.9521
0.75, 1.1, 1.2	0.0146	-0.0003	0.0681	0.0672	0.0849	0.9437	-0.0004	0.0429	0.0426	0.1014	0.9488	0.0001	0.0214	0.0214	0.1697	0.9466
1,0.8,1	0.0161	0.0000	0.0668	0.0658	0.0864	0.9469	-0.0004	0.0422	0.0418	0.1040	0.9473	0.0002	0.0208	0.0210	0.1889	0.9492
1,0.8,0.8	0.0228	-0.0001	0.0661	0.0652	0.1016	0.9464	-0.0004	0.0418	0.0414	0.1363	0.9467	0.0002	0.0206	0.0208	0.2936	0.9496
0.75,1,1.2	0.0231	-0.0007	0.0684	0.0668	0.1026	0.9448	-0.0004	0.0427	0.0424	0.1387	0.9487	0.0000	0.0211	0.0212	0.2879	0.9489
0.75,1,1	0.0276	-0.0007	0.0680	0.0664	0.1159	0.9450	-0.0004	0.0424	0.0421	0.1627	0.9486	0.0000	0.0210	0.0211	0.3626	0.9491
0.75,0.9,1.2	0.0319	-0.0006	0.0675	0.0664	0.1293	0.9430	-0.0006	0.0424	0.0421	0.1887	0.9516	0.0000	0.0209	0.0211	0.4443	0.9521
0.75,1,0.8	0.0329	-0.0008	0.0674	0.0659	0.1333	0.9452	-0.0004	0.0421	0.0418	0.1949	0.9487	0.0000	0.0209	0.0210	0.4698	0.9490
0.75,0.9,0.9	0.0390	-0.0006	0.0667	0.0657	0.1519	0.9441	-0.0006	0.0419	0.0417	0.2397	0.9509	0.0000	0.0207	0.0209	0.5860	0.9526
0.75,0.8,1	0.0455	-0.0007	0.0670	0.0654	0.1822	0.9472	-0.0006	0.0418	0.0415	0.2879	0.9479	0.0000	0.0205	0.0208	0.7065	0.9521

Table D.5: Empirical properties of the utility-based estimator $\widehat{\beta}(\tau)$ and associated tests of $H_0 : \beta(\tau) \leq 0$ vs. $H_1 : \beta(\tau) > 0$ when the illness-death process is mixed-Markov with conditional intensities $\lambda_{kl}(t|X,U) = U\lambda_{kl0}\exp(\gamma_{kl}X)$ for $\phi = 0.5$; $\tau = 1, \pi_T = 0.6, \pi_T^1 = 0.7, \pi_D = 0.4, \pi_r = P(C_r \leq \tau|X = 0) = 0.2$ and $0.4, \exp(\gamma_r) = 1, u_0 = 1, u_1 = 0.5, u_2 = 0, n_{sim} = 10000$ simulation runs.

		MIXED-MARKOV PROCESSES														
		$n = 100$					$n = 250$					$n = 1000$				
$[e^{\gamma_{01}}, e^{\gamma_{02}}, e^{\gamma_{12}}]$	$\beta(\tau)$	BIAS	ESE	ASE	RR	CP	BIAS	ESE	ASE	RR	CP	BIAS	ESE	ASE	RR	CP
$\phi = 1.0: \ \pi_r = 0.20$ and $\exp(\gamma_r) = 1$																
1,1,1	0.0000	0.0009	0.0675	0.0667	0.0564	0.9436	0.0005	0.0423	0.0424	0.0525	0.9475	-0.0004	0.0212	0.0212	0.0481	0.9519
1,1,0.8	0.0067	0.0010	0.0669	0.0661	0.0679	0.9437	0.0006	0.0419	0.0420	0.0721	0.9481	-0.0003	0.0211	0.0211	0.0914	0.9507
0.9,1,1	0.0092	0.0013	0.0679	0.0667	0.0743	0.9430	0.0007	0.0424	0.0424	0.0798	0.9485	-0.0002	0.0214	0.0212	0.1129	0.9471
0.75, 1.1, 1.2	0.0114	0.0011	0.0685	0.0674	0.0752	0.9444	0.0008	0.0429	0.0428	0.0887	0.9448	-0.0003	0.0215	0.0214	0.1293	0.9497
1,0.8,1	0.0152	0.0007	0.0673	0.0660	0.0895	0.9437	0.0006	0.0420	0.0419	0.1046	0.9442	-0.0002	0.0210	0.0210	0.1763	0.9497
0.75,1,1.2	0.0196	0.0012	0.0676	0.067	0.0939	0.9446	0.0007	0.0423	0.0426	0.1225	0.9444	-0.0002	0.0212	0.0213	0.2356	0.9535
1,0.8,0.8	0.0221	0.0007	0.0667	0.0654	0.1055	0.9430	0.0006	0.0417	0.0415	0.1393	0.9440	-0.0001	0.0208	0.0208	0.2775	0.9493
0.75,1,1	0.0243	0.0012	0.0672	0.0666	0.1064	0.9449	0.0007	0.0421	0.0423	0.1444	0.9442	-0.0001	0.0211	0.0212	0.3093	0.9535
0.75,0.9,1.2	0.0281	0.0011	0.0677	0.0666	0.1196	0.9426	0.0006	0.0426	0.0423	0.1692	0.9435	-0.0001	0.0213	0.0212	0.3735	0.9496
0.75,1,0.8	0.0297	0.0013	0.0668	0.0661	0.1241	0.9446	0.0008	0.0417	0.0420	0.1790	0.9453	0.0000	0.0209	0.0210	0.4086	0.9534
0.75,0.9,0.9	0.0354	0.0012	0.0670	0.0660	0.1434	0.9420	0.0007	0.0421	0.0419	0.2176	0.9438	0.0000	0.0211	0.0210	0.5159	0.9486
0.75,0.8,1	0.0416	0.0012	0.0670	0.0658	0.1686	0.9427	0.0009	0.0418	0.0418	0.2670	0.9469	0.0001	0.0211	0.0209	0.6321	0.9470
$\phi = 1.0: \ \pi_r = 0.40$ and $\exp(\gamma_r) = 1$																
1,1,1	0.0000	0.0006	0.0697	0.0689	0.0529	0.9471	0.0003	0.0440	0.0437	0.0510	0.9490	-0.0001	0.0221	0.0219	0.0523	0.9477
1,1,0.8	0.0067	0.0007	0.0691	0.0683	0.0632	0.9471	0.0004	0.0437	0.0434	0.0703	0.9493	0.0000	0.0219	0.0217	0.0941	0.9482
0.9,1,1	0.0092	0.0010	0.0700	0.0688	0.0721	0.9449	0.0004	0.0438	0.0437	0.0764	0.9495	0.0000	0.0221	0.0219	0.1122	0.9504
0.75, 1.1, 1.2	0.0114	0.0010	0.0707	0.0696	0.0737	0.9460	0.0007	0.0442	0.0442	0.0895	0.9466	-0.0001	0.0224	0.0221	0.1350	0.9477
1,0.8,1	0.0152	0.0005	0.0692	0.0681	0.0816	0.9475	0.0005	0.0431	0.0433	0.0987	0.9494	0.0002	0.0220	0.0217	0.1799	0.9444
0.75,1,1.2	0.0196	0.0009	0.0702	0.0692	0.0924	0.9444	0.0006	0.0438	0.0440	0.1163	0.9505	0.0000	0.0222	0.0220	0.2303	0.9499
1,0.8,0.8	0.0221	0.0005	0.0686	0.0675	0.0991	0.9464	0.0005	0.0428	0.0429	0.1277	0.9491	0.0002	0.0218	0.0215	0.2722	0.9442
0.75,1,1	0.0243	0.0010	0.0697	0.0688	0.1066	0.9441	0.0006	0.0436	0.0437	0.1421	0.9501	0.0000	0.0220	0.0219	0.2986	0.9494
0.75,0.9,1.2	0.0281	0.0011	0.0700	0.0688	0.1154	0.9427	0.0006	0.0437	0.0437	0.1632	0.9469	0.0001	0.0220	0.0219	0.3551	0.9487
0.75,1,0.8	0.0297	0.0010	0.0692	0.0683	0.1233	0.9449	0.0006	0.0433	0.0434	0.1762	0.9494	0.0001	0.0219	0.0217	0.3948	0.9495
0.75,0.9,0.9	0.0354	0.0013	0.0693	0.0681	0.1406	0.9435	0.0007	0.0432	0.0433	0.2125	0.9466	0.0002	0.0217	0.0217	0.4899	0.9475
0.75,0.8,1	0.0416	0.0011	0.0688	0.0679	0.1627	0.9475	0.0008	0.0430	0.0432	0.2522	0.9485	0.0002	0.0218	0.0216	0.6127	0.9477

Table D.6: Empirical properties of the utility-based estimator $\widehat{\beta}(\tau)$ and associated tests of $H_0 : \beta(\tau) \leq 0$ vs. $H_1 : \beta(\tau) > 0$ when the illness-death process is mixed-Markov with conditional intensities $\lambda_{kl}(t|X,U) = U\lambda_{kl0}\exp(\gamma_{kl}X)$ for $\phi = 1$; $\tau = 1, \pi_T = 0.6, \pi_T^1 = 0.7, \pi_D = 0.4, \pi_r = P(C_r \leq \tau|X = 0) = 0.2$ and $0.4, \exp(\gamma_r) = 1, u_0 = 1, u_1 = 0.5, u_2 = 0, n_{sim} = 10000$ simulation runs.

D.6 Additional results for the application of Section 5.5

Figure D.2 depicts plots of the Nelson-Aalen estimates of the cumulative intensity functions out of state 0 by transfusion type.

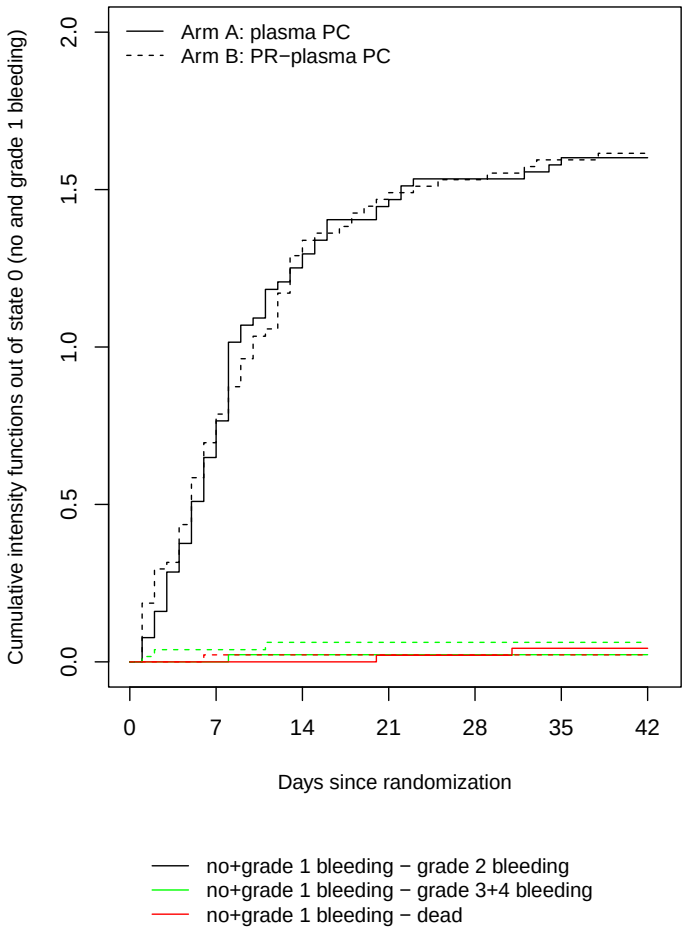


Figure D.2: Nonparametric Nelson-Aalen estimates of the cumulative intensity functions for transitions out of state 0 by transfusion type.