Dynamic Prediction Models for Recurrent and Alternating Recurrent Events

by

Abigail G. Loe

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

Doctoral Committee:

Professor Susan Murray, Co-Chair
Associate Professor Zhenke Wu, Co-Chair
Associate Professor Kelly Bakulski
Professor Donglin Zeng

Abigail G. Loe
agloe@umich.edu
ORCID iD: 0009-0002-3123-067X

© Abigail G. Loe 2026

For my grandmothers, Monica and Carol.

ACKNOWLEDGEMENTS

Thank you to my advisors, Drs. Susan Murray and Zhenke Wu. It has been a joy to work through the challenges of a PhD with both of you. Zhenke, you took a chance and answered a cold email from a first-year graduate student. Your decision to work with me and fund me that first winter set me on the track that I am now. You helped me see the bigger picture in the tools I created for my thesis, and pushed me to think about the bigger picture, both in and outside of the classroom. Susan, you have been a constant support in and out of the classroom. It has been a privilege to work and learn under you (as well as to see you at protests) To know that I was cared for as a student as well as a person has been a wonderful gift. When I was discouraged, you would be the first person to tell me that all was not lost, and remind me that I would figure out the problem eventually. As I stand at the end of my time at Michigan, I realize now how lucky I was to have you as an advisor, in all of the ways that you have set me up for success in my next role; words cannot express how grateful I am.

I would also like to thank all of the collaborators I have had the pleasure of working with during my time at the University of Michigan, including (but not limited to): Drs. Imre Noth, John Kim, David O'Dwyer, and Adam Martin-Schwarz. Special thanks goes to Dr. Matt Zawistowski who helped nurture my love of teaching, and gave me the opportunity to mentor undergraduates. I am also grateful for the support and mentorship of Dr. Tom Braun, Dr. Kelly Bakulski, and Dr. Donglin Zeng.

Graduate school has been a time of immense growth out of the classroom. Friends from volleyball, roommates, and fellow classmates form a support system that I cannot begin to list both for space reasons, and for fear of missing someone critical. We have laughed together, commiserated together and studied together. Know that I am grateful for each and every one of you. And finally, to my family: my parents, Melissa and Brian, and my sisters (in a randomly selected order, as is our way), Maren, Katie and Caroline. When I have needed support, a shoulder to cry on (or couch to nap on), a person to run or laugh with, or a travel buddy, my family has been there for me. I love you all so very much.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

LIST OF FIGURES vi

LIST OF TABLES xv

LIST OF APPENDICES xix

ABSTRACT xx

CHAPTER

1 Introduction 1

1.1 Commonalities Across All Chapters 3
1.1.1 Survival Analysis and Censoring 3
1.1.2 Censored Longitudinal Data Structure 4
1.1.3 Pseudo-observations 5
1.1.4 History Covariates 6
1.2 Random Forest Algorithms 7
1.3 Prediction vs. Inference 9

2 Random Forest for Dynamic Risk Prediction of Recurrent Events: a Pseudo-Observation Approach* 10

2.1 Introduction 10
2.2 Notation and Censored Longitudinal Data Structure 13
2.2.1 Constructing Historical Recurrent Event Information as Predictors 14
2.3 Pseudo-Observations for Dynamic Prediction 18
2.4 Semiparametric Model for Dynamic Prediction (Modified XMT Model) . . . 19
2.5 Random Forest for Dynamic Prediction (RFRE.PO algorithm) 20
2.6 Evaluation Metrics 23
2.7 Simulation Study 23
2.8 Prediction of COPD Exacerbations 27
2.9 Discussion 33

3 A Random Forest Inverse Probability Weighted Pseudo-Observation Framework for Alternating Recurrent Events* 39

3.1 Introduction . . . 39
3.2 Censored Longitudinal Data Structure . . . 41
3.3 Construction of Pseudo-Observations . . . 43
3.4 Inverse Probability Weighted Pseudo-Observations . . . 45
3.5 Estimation of Weights via Random Forest . . . 47
3.6 Simulation Study . . . 49
3.7 Application to the CareQOL Study . . . 51
3.8 Discussion . . . 54

4 A Neural Network-Based Dynamic Prediction Model for Alternating Recurrent Events* . . . 57

4.1 Introduction . . . 57
4.2 Censored Longitudinal Data Structure . . . 59
4.3 Pseudo-Observation . . . 61
4.4 Statistical Review of Neural Networks . . . 61
4.4.1 Linear Regression as a Neural Network . . . 63
4.4.2 Long Short-term Memory Neural Networks . . . 66
4.4.3 Useful Extensions to Bare-bones LSTM . . . 67
4.5 Weight Estimation . . . 69
4.6 Weighted Semi-parametric Dynamic Prediction Model . . . 74
4.7 Simulation Study . . . 74
4.7.1 Evaluation Metrics . . . 76
4.7.2 Results . . . 77
4.8 Application to Intern Health Study Data . . . 81
4.9 Discussion . . . 82

5 Conclusion . . . 86

APPENDICES . . . 89

BIBLIOGRAPHY . . . 127

LIST OF FIGURES

FIGURE

1.1 For alternating recurrent events, an example of the correlated gap time problem. Event-times of interest for subject i are recorded as $T_{i,j}$, and return to risk population j for subject i is denoted $D_{i,j}$. Gap-time analysis uses $G_{i,j} = T_{i,j} - D_{i,j-1}$ as a response. To see the dependent gap-time issue, assume $T_{i,j}, D_{i,j} \perp C_i \forall j$, but assume correlation between event-times. At the first gap-time, C_i and $G_{i,1}$ are independent of each other. However, for subsequent events, time-to censoring and and gap-times now are functions of correlated variables, and thus are no longer independent. 4

1.2 A single random tree modeling binary response Y using covariates X_1 and X_2. The regression model that corresponds to this non-parametric tree is $\text{logit}(\pi_i) = \beta_1 I(X_1 < 0) + \beta_2 I(X_1 \geq 0) I(X_1 \geq 3) + \beta_3 I(X_1 \geq 0) I(X_1 < 3) I(X_2 \neq \text{Low}) + \beta_4 I(X_1 \geq 0) I(X_1 < 3) I(X_2 = \text{Low})$. Unlike regression, if X_1 and X_2 are strongly correlated, a random tree may still be fit as tree growth does not rely on Gram matrix inversion. 8

2.1 Panel (A) displays a visualization of recurrent events in both traditional format (black) and censored longitudinal format (blue). Start times of follow-up windows ($\mathcal{T} = \{0, 60, 120, 180 \text{ days}\}$, with $a = 60$; this step does not concern τ) are highlighted by vertical dashed lines. Traditional observed recurrent event data for individuals $i = 1, 2, 3$ are observed in Panel (B), while the observed censored longitudinal data for these individuals is displayed in Panel (C). All participants were administratively censored after 8 months (240 days) of follow-up. Participant 2 has no recorded events before C_2. Additionally, while $X^*_{3,3}$ is technically observed, it is not included in the censored longitudinal data set because it is the second observation in the window, and there is no t such that $X_i(t) = X^*_{3,3} - t$. Besides potential correlation between separate event times seen in the same individual, this data structure induces correlation due to the same event contributing to the final censored longitudinal values in more than one follow-up window. For example, $X^*_{1,2}$ produces two measures in the longitudinal data structure, $X_1(120) = 83$ and $X_1(180) = 23$. 15

2.2 A flowchart depicting the analysis pipeline of `RFRE.PO`. Labels above arrows correspond to the section which describes the corresponding step in greater detail. An analyst begins with data recorded as recurrent event times for subject i, $j = 1, 2, \ldots, J_i$ before constructing an appropriate $\mathcal{T}$, and transforming data into the censored longitudinal framework. Upon creating such a structure, the analyst calculates pseudo-observations, which are ultimately fed to a historical random forest. Recall that $p_{w,\text{test}}$ refers to the permutation test statistic described in Section 2.5, and w refers to a variable selected for permutation 24

2.3 Simulated model fit for (1) the modified XMT model using main effects terms only {m-XMT (Main Effects), shown in purple}, (2) our random forest algorithm for recurrent events (RFRE.PO, shown in green) and (3) the true modified XMT model {m-XMT (True), shown in yellow}. Results are based on 500 simulated datasets of n=500 participants per scenario. Panel (A) displays violin plots of C-statistics (indices of concordance) seen in simulation. For partial history information, average C-statistics from $m = 10$ imputed data sets are shown. Measures of center and spread are in Table 2.1. Within a box, historical covariate information increases from left to right. Panel (B) expands on the scenario boxed in red from panel (A), i.e., the low censoring setting with 0.3 correlation. Here violin plot C-statistic distributions are time-dependent, reflecting model fit for $\tau = \frac{1}{6}$ duration follow-up windows starting at times, t, marked on the horizontal axis. The quality of historical information increases from top to bottom. 28

2.4 Panel (A) displays the mean squared error (MSE) for simulation settings displayed in Figure 2.3. Censoring increases from top to bottom, and the amount of induced correlation increases from left to right. Within a box, the quality of historical information increases from left to right. Panel (B) shows the MSE at the start of each follow-up window for the setting boxed in red in Panel (A). From top to bottom, more historical information is added in Panel (B), and the MSE corresponding to `RFRE.PO` in $H_p^{(1)}(t)$ settings. 500 replicates were performed, each with $n = 500$ subjects. 29

2.5 A correlation heat map of continuous and ordinal variables in the Azithromycin for the Prevention of COPD cohort, measured at baseline. Variables derived from each other have instances of particularly high positive or negative correlation. These correlation patterns make random forest a more attractive modeling alternative compared to semi-parametric and parametric methods. 36

2.6 Wald Statistic Z-scores derived from the validation set for predictors that were included in the m-XMT models (Clinical Input model from Xia, Murray, Tayob 2020) and the m-XMT selected based on forward selection of the Wald statistic, after adjusting for multiple imputation using Rubin's Rule, where appropriate, and permutation test Z-scores derived from out-of-bag samples as part of the `RFRE.PO` algorithm, adjusting for multiple imputation using Rubin's Rule where appropriate. For the `RFRE.PO` method, 100 permutations were used in calculating the permutation Z-scores. Note that some Wald forward selection variables lost significance as additional terms were added. Colors along the y-axis characterize the type of predictor displayed. Predictors are grouped into one of five categories, displayed in the legend on the right. For the analysis, we created three history variables, $H_f(t), H_p^{(2)}(t)$ and $H_p^{(3)}(t)$. $H_f(t)$ was an indicator if the subject had experienced a severe exacerbation on study time prior to time t. $H_p^{(2)}(t)$ was assumed to take a categorical form based on whether the participant had an exacerbation in the past zero to 31 days ($H_p^{(2)}(t) = 4$), 31-92 days ($H_p^{(2)}(t) = 3$), 93-182 days ($H_p^{(2)}(t) = 2$), 183 days to 365 days ($H_p^{(2)}(t) = 1$), or finally, if that time was more than 365 days or never ($H_p^{(2)}(t) = 0$). $H_p^{(3)}(t)$ is the estimated 30-day event rate using all follow-up prior to time t. Both of these partial history variables are updated at later follow-up windows starting at $t > 0$ based on observed exacerbation data for each individual. 37

2.7 `RFRE.PO` predictions and history covariates from the Azithromycin for the Prevention of COPD validation cohort via a 3×3 panel heatmap display. Within each row of heatmaps, subjects are ordered vertically, and clustered according to the default settings of the `pheatmap` command in R applied to the fitted dynamic prediction values for that row of heatmaps. Heatmap rows marked (A), (B) and (C) vary values of follow-up window duration, τ, assuming $\tau = 90$, $\tau = 180$ and $\tau = 270$, respectively. Heatmap columns marked (1), (2) and (3) display: (1) dynamic predictions, $\widetilde{P}\{T_i(t) \geq \tau | Z_i(t)\}$, shown for follow up windows starting at t, with t spaced every 30 days until $t + \tau$ exceeds 360, (2) corresponding categorical $H_p^{(2)}(t)$ and (3) corresponding continuous $H_p^{(3)}(t)$. Both partial history variables are described in detail in Section 2.2, and the specification of $\mathcal{T}_i^*(t)$ and τ^* are described in Section 2.8. Although individuals are not clustered identically in the heatmaps shown in rows (A), (B) and (C), the heatmap color schemes (as expected) reflect higher probabilities of remaining event-free over shorter follow up windows, i.e., $\widetilde{P}\{T_i(t) \geq 90 | Z_i(t)\} > \widetilde{P}\{T_i(t) \geq 180 | Z_i(t)\} > \widetilde{P}\{T_i(t) \geq 270 | Z_i(t)\}$. The display of the history variables seems to reflect in the fitted values from the `RFRE.PO` method, as patterns in column (1) are roughly represented in columns (2) and (3). 38

3.1 A visualization of alternating recurrent events in both the longitudinal sense and the alternating censored longitudinal format. In Panel (A), events within a subject are marked with an "X." The primary and secondary event types are the start and the end of the red segments, respectively. Blue segments represent at-risk periods for the primary event type. Start times of potential follow-up windows of length $\tau = 120$ corresponding to $\mathcal{T} = \{0, 60, 180, 240 \text{ days}\}$ are represented with dashed vertical lines. Traditional observed alternating recurrent events are recorded in Panel (B), while Panel (C) corresponds to those same event converted to an alternating censored longitudinal format. All participants were administratively censored after 360 days of follow-up, thus $t = 300$ is not included in $\mathcal{T}$. Note that $X_{1,0}^{(1)}$ (the first 'X' in subject 1) produces multiple measures in the alternating censored longitudinal data format. Additionally, subject 2 produces three measures where $R_2(t) = 1$; subject 3 produces the same number of measures where $R_3(t) = 1$, but was removed from the at-risk population at $t = 60, 120$. 44

3.2 Violin plots of censored point estimates seen in simulation with censoring applied (500 replicates, $n = 750$). Coefficients β_1 and β_2 reflect those with coefficients $Z_i = 1$ and $Z_i = 0$ before $t = 6$ and coefficients β_3 and β_4 after and including $t = 6$, respectively. True coefficient values are indicated via horizontal lines. Because τ= 1 (month) in simulation, bias away from true parameter values multiplied by 100% can be interpreted as absolute percentage bias from the true paramter value. Features included in the random forest algorithm for estimating PAIR-GEE weights involve the seven history covariates described in Section 3.5 with varying degrees of availability that decreases from left to right for each displayed coefficient. Here, $\boldsymbol{H}_f$ indicates the full year of patient history was available prior to $t = 0$. $\boldsymbol{H}_{p_1}$ and $\boldsymbol{H}_{p_2}$ indicate that the patient history included information on the last two or one alternating recurrent events, respectively, if these occurred within the prior year. $\boldsymbol{H}_{p*}$ indicates that no historical information prior to $t = 0$ was available and that reflexive history covariates (as described in Section 3.5) were used. $\boldsymbol{H}_{p_3}$ indicates that complete history was available on 50% of individuals, but was otherwise unavailable. $\boldsymbol{H}_{p_4}$ indicates no availability of history covariates prior to $t = 0$. $\boldsymbol{H}_n$ indicates no history covariates were included in the random forest algorithm. 52

3.3 Point estimates and corresponding 95% confidence intervals from the CareQOL data analysis. The left and right panels correspond to the estimates of 2-week restricted mean time-to-entering and exiting depressed states, respectively. Significant coefficients, or coefficients that change significance between methods are displayed in a red font. Covariates that were standardized prior to inclusion in the model are denoted with an asterisk. Our analysis finds that after adjustment for common covariates, being in the intervention group was no longer significantly associated with a decrease in the number of days till entering a depressive state. 55

3.4 Distribution of inverse probability weights, based on random forest estimation of $\widehat{P}\{R_i(t)|\mathbf{Z}_i(t), \boldsymbol{H}_{p*}\}$ for subjects that are at-risk at each time point (Panel A), and subjects that are in a secondary state at time t (Panel B). Subjects are assumed to be at-risk at time $t = 0$. Weights were capped at 10. 56

4.1 A figure depicting the transformation from alternating recurrent events into the appropriate random variables of the alternating censored longitudinal data structure corresponding to $\mathcal{T} = \{0, 30, 60, 90, 120, 150\}$ and $\tau = 60$ days (represented by the curly braces). Panel (A) depicts the traditionally observed alternating recurrent events, with uncensored event times, $T_{i,j}^{(k)}$ listed, with corresponding observed event times $X_{i,j}^{(k)}$ and indicator functions, $\delta_{i,j}^{(k)}$. Periods where subject i is at-risk for a primary alternating recurrent event are depicted in light green, while secondary or recovery states are shown in purple. The τ-length follow-up window corresponding to $\tau = 60$ days and $\mathcal{T}_i = \{0, 30, 60, 90, 120, 150\}$ is shown with curly braces. Panel (B) depicts the same $\mathcal{T}_i$, τ, and subject, but records the random variables appropriate for use with the alternating censored longitudinal data structure described in Section 4.2. Potentially time-varying covariates $Z_i(t)$ may be recorded at each $t = \{0, 30, 60, 90, 120, 150\}$. 62

4.2 Three fully connected neural network architectures of increasing complexity, with squared error loss. Vectors of covariates are depicted in circles on the left, while nodes containing estimable parameters are displayed in rectangular boxes. Networks are called "fully connected" because each covariate is passed to each node in the first layer, and for networks with more than one layer, outputs from previous layers are passed to all subsequent layers. Panel (A) displays an architecture with one hidden layer containing one node, akin to linear regression in the special case where g is the identity link. Panel (B) displays a network architecture with three nodes in one hidden layer, that are ultimately combined to obtain the linear component, η_B. As Panel (C) contains multiple hidden layers, Panel (C) proposes a network that may also be termed a "deep learner." 70

4.3 A panel graphic displaying the network architecture termed "long short-term memory," or LSTM at time t_k. Panel (A) displays the functionals used in training an LSTM at time t_0 (base case), while Panel (B) displays the self-referential architecture for later $t \in \mathcal{T}$ (recursive). When connected recursively across time points, Panel A and Panel B create an LSTM at multiple $t \in \mathcal{T}$. Software package `torch` allows user-specification of links g_f, g_i, g_{C*}, g_o and g_C but defaults to g_f, g_i, g_{C*} and g_o as the log-link and g_C and g_{C*} as the inverse hyperbolic tangent; thus $g_{f,i,C*,o}^{-1}(x) = \frac{\exp(x)}{1+\exp(x)}$, and $g_{C,C*}^{-1}(x) = \tanh(x)$. In Section 4.7, we used defaults from the `torch` package in Python. In contrast to Figure 4.2, where outputs from one layer are combined via another generalized linear model functional, node outputs from $C_{t_k}^*$ and i_{t_k} are combined together via a dot product, as are f_{t_k} with $C_{t_{k-1}}$, and C_{t_k} with o_{t_k}. Not all covariates are used in all nodes in the first layer; covariate $C_{t_{k-1}}$ is only used in the second layer, while o_{t_k} is calculated in the first layer, before final combination in the third. At each time, inputs that depend on the algorithm at the prior time point itself are included ($C_{t_{k-1}}$ and $\eta_{t_{k-1}}$). At t_0, these quantities default to 0. In Panel (B), while ultimate output of interest is $g^{-1}(\eta_{t_k})$, loss from the previous time point factors in to the modeling of η_{t_k}. 71

4.4 A figure depicting the training, validation, and testing split for the rolling-origin algorithm, mapped onto the censored longitudinal data structure. The model building component contains data from $\{t_0, t_1, \ldots t_{k-1}\}$, while model testing comes from data observed at t_k. Within the model building cohort, the network is first trained on data from $\{t_0, t_1, \ldots t_{k-2}\}$, while data observed in the window starting at time t_{k-1} are used to validate. When the optimal number of epochs has been reached, n_{opt}, initial training stops. Re-train the model using data observed at window start times $\{t_0, t_1, \ldots t_{k-1}\}$. This model is then tested for data at t_k, to obtain dynamic predictions of $E[\min\{T_i(t_k), \tau\}]$. 72
4.5 A figure displaying the analysis pipeline for on-line (real-time) prediction using random forest or LSTM IPW weights and the rolling-origin training and validation split for both the LSTM and GEE-based models. 73
4.6 A panel graphic displaying the performance of the dynamic prediction models under consideration in Section 4.7 for the $\boldsymbol{H}_l$ setting. Counter clockwise from upper left is (A) c-statistic over time, (B) percent bias (in terms of τ) over time, (C) testing root mean squared error over time, and (D) mean absolute bias over time. The method used to predict $E[\min\{T_i(t), \tau\}] = \bar{\mathbf{Z}}^\top \beta$ is listed first (yellow for GEE Oracle, black for GEE Main Effects, and blue for LSTM), while the weight model is listed in parenthetical (with LSTM weights graphed in a solid line, while random forest weights are dashed). All results are from 200 replicates with $n = 2500$ each. 79
4.7 A panel graphic displaying the performance of the dynamic prediction models under consideration in Section 4.7 for the $\boldsymbol{H}_h$ setting. Counter clockwise from upper left is (A) c-statistic over time, (B) average bias over time, (C) testing root mean squared error over time, and (D) mean absolute bias over time. The method used to predict $E[\min\{T_i(t), \tau\}] = \bar{\mathbf{Z}}^\top \beta$ is listed first (yellow for GEE Oracle, black for GEE Main Effects, and blue for LSTM), while the weight model is listed in parenthetical (with LSTM weights graphed in a solid line, while random forest weights are dashed). All results are from 200 replicates with $n = 2500$ each. 80
4.8 A panel graphic displaying results from predicting periods of low mood for medical interns. The censored longitudinal data strucutre takes the form $\tau = 14$, $\mathcal{T} = \{0, 14, 28, \ldots, 350\}$. Panel (A) displays c-statistic, while Panel (B) displays testing RESS. The main effects model fit via GEE is graphed in black, while the LSTM models are shown in sky blue. The method used to calculate IPW weights is listed in parenthetical after the model name. 83

A.1 Panel (A) displays the mean squared error (MSE) for simulation settings displayed in Figure 2.3. Censoring increases from top to bottom, and the amount of induced correlation increases from left to right. Within a box, the quality of historical information increases from left to right. Panel (B) shows the MSE at the start of each follow-up window for the setting boxed in red in Panel (A). From top to bottom, more historical information is added in Panel (B), and the MSE corresponding to `RFRE.PO` in $H_p^{(1)}(t)$ settings. 500 replicates were performed, each with $n = 500$ subjects. 92

A.2 A panel graphic displaying time-dependent C-statistics (Panel A) and MSEs (Panel B) in the validation dataset of the Azithromycin for the Prevention of COPD study, calculated at each follow-up window $t \in \{0, 30, 60, 90, 120, 150, 180\}$ for $\tau = 180$; 95% confidence intervals also displayed. Compared to the other algorithms, `RFRE.PO` has fairly stable and attractive performance metrics across the various follow-up windows, likely because interactions with the various predictors over time are naturally incorporated if useful for prediction. 103

B.1 Six representative simulated subjects' alternating recurrent event data, following the model described in Section 3.6. Individuals $i = 1, 2, 3$ have $Z_i = 0$, and all other individuals have covariate profile $Z_i = 1$. Window start times, $t \in \mathcal{T}$, are displayed on the horizontal axis. Primary and secondary event types are the start and the end of the light salmon segments, respectively. Purple segments represent that at-risk periods of the primary event type. The time at which subjects experience a bend in hazards is distinguished with a vertical dashed line. Subjects 1 through 3 experience more time at risk relative to subjects $4, 5$, and 6. The latter subjects are more likely to be missing from the complete case analysis at window start times, $t \in \mathcal{T}$. 113

B.2 A graphic displaying the true τ-restricted mean compared to the empirical, or observed, τ-RMST in the highly correlated case for a single replicate. True parameter estimates are displayed using horizontal lines, while the empirical value is displayed on the y axis. Lines connecting the estimated τ-RMST are annotated with the number of subjects in the risk set for each window start time and each covariate profile. 115

B.3 Distribution of point estimates in the presence of censoring, when primary alternating recurrent events within a subject are independent This corresponds to the case of missing completely at random for exponentially distributed gap times. Results come from 300 replicates of $n = 750$ each. Features included in the random forest algorithm for estimating PAIR-GEE weights involve the seven history covariates described in Section 3.5 with varying degrees of availability that decreases from left to right for each displayed coefficient 116

B.4 Violin plots of uncensored point estimates seen in simulation (500 replicates, $n = 750$). Coefficients β_1 and β_2 reflect those with coefficients $Z_i = 1$ and $Z_i = 0$ before $t = 6$ and coefficients β_3 and β_4 after and including $t = 6$, respectively. Features included in the random forest algorithm for estimating PAIR-GEE weights involve the seven history covariates described in Section 3.5 with varying degrees of availability that decreases from left to right for each displayed coefficient. Here, $\boldsymbol{H}_f$ indicates the full year of patient history was available prior to $t = 0$. $\boldsymbol{H}_{p_1}$ and $\boldsymbol{H}_{p_2}$ indicate that the patient history included information on the last two or one alternating recurrent events, respectively, if these occurred within the prior year. $\boldsymbol{H}_{p*}$ indicates that no historical information prior to $t = 0$ was available and that $\boldsymbol{H}_{p^*}$ history covariates (as described in Section 3.5) were used. $\boldsymbol{H}_{p_3}$ indicates that complete history was available on 50% of individuals, but was otherwise unavailable. $\boldsymbol{H}_{p_4}$ indicates no availability of history covariates prior to $t = 0$. $\boldsymbol{H}_n$ indicates $\boldsymbol{H}_n$ were included in the random forest algorithm. . 117

B.5 Bias in point estimation for each $t \in \mathcal{T}$ when alternating recurrent events are correlated with each other. The top half of the graphic displays results from the uncensored simulation, with the bottom half displaying results from the censored results, described in the main paper. Notice the persistent bias for the unweighted methods, as well as the PAIR-GEE with $\boldsymbol{H}_n$, while other methods of incorporating history covariates are able, to varying extents, adjust for patient specific history in a more elegant manner, and approach non-zero bias in later follow-up windows. 118

C.1 A graphic displaying the generating model for simulation, where b_{pre} denotes burn-in, t is window start-time beyond burn-in, and $G_{i,j}^{(k)}$ denote the correlated multivariate normal gap-times. The primary alternating recurrent event of interest that forms $T_i(t)$ is circled in blue, while primary event gap times are displayed in green, and secondary event times are displayed in purple. 120

C.2 A panel graphic displaying the performance of the dynamic prediction models under consideration in Section 4.7 for the $\boldsymbol{H}_l$ setting. Confidence bands display the 97.5% percentile and 2.5% percentile respectively. Counter clockwise from upper left is (A) c-statistic over time, (B) percent bias (in terms of τ) over time, (C) testing root mean squared error over time, and (D) mean absolute bias over time. The method used to predict $E[\min\{T_i(t), \tau\}] = \bar{\mathbf{Z}}^\top \beta$ is listed first (yellow for GEE Oracle, black for GEE Main Effects, and blue for LSTM), while the weight model is listed in parenthetical (with LSTM weights graphed in a solid line, while random forest weights are dashed). All results are from 200 replicates with $n = 2500$ each. 125

C.3 A panel graphic displaying the performance of the dynamic prediction models under consideration in Section 4.7 for the $\boldsymbol{H}_h$ setting. Confidence bands display the 97.5% percentile and 2.5% percentile respectively. Counter clockwise from upper left is (A) c-statistic over time, (B) percent bias (in terms of τ) over time, (C) testing root mean squared error over time, and (D) mean absolute bias over time. The method used to predict $E[\min\{T_i(t), \tau\}] = \bar{\mathbf{Z}}^\top \beta$ is listed first (yellow for GEE Oracle, black for GEE Main Effects, and blue for LSTM), while the weight model is listed in parenthetical (with LSTM weights graphed in a solid line, while random forest weights are dashed). All results are from 200 replicates with $n = 2500$ each. 126

LIST OF TABLES

TABLE

2.1 Simulated C-statistics listed as Mean (ESD) for varying prediction algorithms when Z_1, Z_2 are correlated with each other, censoring percentages (none is 0%, light is 23%, moderate is 45%, and heavy is 63%), and gap-time correlations (ρ = 0, 0.3, 0.6, 0.9). 500 replicates were performed with n = 500 for each replicate, and the same data is reported in Figure 2.3. 34

2.2 Simulated MSE listed as Mean (ESD) for varying prediction algorithms considered in Section 2.7, with censoring percentages (None, Light, Moderate, and Heavy), and gap-time correlations (ρ = 0, 0.3, 0.6, 0.9). 500 replicates were performed with $n = 500$ for each replicate. 35

3.1 A table displaying results from simulation in the case where primary recurrent events have correlation of $\rho = 0.8$, and secondary recurrent events are independent of each other. The hazard of primary alternating recurrent event completely determines the hazard of the alternating secondary recurrent event, approximately 30% of subjects are censored before the final follow-up window, and results come from 500 replicates of $n = 750$ each. 53

4.1 A table displaying the hyperparameters selected for the LSTM models in Section 4.7. Hyperparameters were selected based on lowest testing loss in a rolling-origin grid search. Weight refers to the weighting strategy used in the neural network loss function, while "Subj. Embed Dim." and "Time Embed Dim." refer to the dimension of the vector that the network uses to learn time and subject characterizations. Hyperparameter h refers to the dimension of the "hidden state," or number of columns of the linear predictor η, where $\eta \in \mathbb{R}^{n(t) \times h}$. The number of LSTM layers refers to the number of LSTM architectures stacked on top of each other, and p_d is the proportion of slopes and intercept randomly set to 0 during each epoch of training. For further details on these parameters, see Section 4.4.2. 78

4.2 A table displaying the hyperparameters selected for the LSTM models in Section 4.8. Hyperparameters were selected based on lowest testing loss in a rolling-origin grid search at t_3. Weight refers to the weighting strategy used in the neural network loss function, while "Subj. Embed Dim." and "Time Embed Dim." refer to the dimension of the vector that the network uses to learn time and subject characterizations. Hyperparameter h refers to the dimension of the "hidden state," or number of columns of the linear predictor η, where $\eta \in \mathbb{R}^{n(t)\times h}$. The number of LSTM layers refers to the number of LSTM architectures stacked on top of each other, and p_d is the proportion of slopes and intercept randomly set to 0 during each epoch of training. 82

A.1 Simulated C-statistics listed as Mean (ESD) for varying prediction algorithms when Z_1, Z_2 are correlated with each other, censoring percentages (none is 0%, light is 23%, moderate is 45%, and heavy is 63%), and gap-time correlations (ρ = 0, 0.3, 0.6, 0.9). 500 replicates were performed with n = 500 for each replicate, and the same data is reported in Figure 2.3. 93

A.2 Simulated MSE listed as Mean (ESD) for varying prediction algorithms considered in Section 2.7, with censoring percentages (None, Light, Moderate, and Heavy), and gap-time correlations (ρ = 0, 0.3, 0.6, 0.9). 500 replicates were performed with $n = 500$ for each replicate. 94

A.3 Simulated proportion of times that the corresponding covariate appeared as statistically significant according to the Wald test for the m-XMT (Main Effects) method. History covariates considered were of the type τ^*-restricted mean survival time in previous windows, where τ^* is $\frac{1}{12}$ years (described in further detail in Section 2.2.1). In the $H_p^{(1)}(t)$ case, results were from $m = 10$ multiply imputed datasets combined via Rubin's rules. 500 replicates were performed with $n = 500$ for each replicate. Covariates with signal (Z_1, Z_2, Z_3, Z_4, Z_6) are displayed first, and noise covariates are shown second (Z_5, Z_7). Within a censoring and correlation designation, quality of historical information included in the model increases, from top to bottom. 95

A.4 Simulated proportion of times that the corresponding covariate appeared as statistically significant according to the random forest permutation test. 500 replicates were performed, with $n = 500$ subjects in each replicate. History covariates considered were of the type τ^*-restricted mean survival time in previous windows, where τ^* is $\frac{1}{12}$ years (described in further detail in Section 2.2.1). In the $H_p^{(1)}(t)$ case, results were from $m = 10$ multiply imputed datasets, combined via Rubin's rules. Covariates with signal (Z_1, Z_2, Z_3, Z_4, and Z_6) are displayed first, and noise covariates are shown second (Z_5, Z_7). Within a censoring and correlation designation, quality of historical information included in the model increases, from top to bottom. 97

A.5 m-XMT parameter estimates, based on covariates selected by Xia et al. [2020], applied to censored longitudinal data with $\tau = 180$, $a = 30$. 99

A.6 Parameter estimates from the m-XMT (Forward Selection) model based on the first of $m = 10$ imputed datasets. For the analysis, we created three history variables, $H_f(t), H_p^{(2)}(t)$ and $H_p^{(3)}(t)$. $H_f(t)$ was an indicator if the subject had experienced a severe exacerbation on study time prior to time t. $H_p^{(2)}(t)$ was assumed to take a categorical form based on whether the participant had an exacerbation in the past zero to 31 days ($H_p^{(2)}(t) = 4$), 31-92 days ($H_p^{(2)}(t) = 3$), 93-182 days ($H_p^{(2)}(t) = 2$), 183 days to 365 days ($H_p^{(2)}(t) = 1$), or finally, if that time was more than 365 days or never ($H_p^{(2)}(t) = 0$). $H_p^{(3)}(t)$ is the estimated 30-day event rate using all follow-up prior to time t. A very recent exacerbation, bronchitis, gender, treatment, the ratio of FEV to FVC %predicted, SGRQ symptoms, oxygen use, recent hospitalization, rate of exacerbations, and use of ICS medication remain significant across the $m = 10$ imputed datasets. 100

A.7 P-values from the `RFRE.PO` permutation test applied to the out-of-bag-samples combined across $m = 10$ multiply imputed datasets. Included predictors had a statistically significant permutation test z-score for at least one of the multiply imputed training datasets. Additional predictors that did not achieve statistical significance via the permutation test, but were used in the `RFRE.PO` algorithm in some form, are not shown. For analysis, we created three history variables, $H_f(t), H_p^{(2)}(t)$ and $H_p^{(3)}(t)$. $H_f(t)$ was an indicator if the subject had experienced a severe exacerbation on study time prior to time t. $H_p^{(2)}(t)$ was assumed to take a categorical form based on whether the participant had an exacerbation in the past zero to 31 days ($H_p^{(2)}(t) = 4$), 31-92 days ($H_p^{(2)}(t) = 3$), 93-182 days ($H_p^{(2)}(t) = 2$), 183 days to 365 days ($H_p^{(2)}(t) = 1$), or finally, if that time was more than 365 days or never ($H_p^{(2)}(t) = 0$). $H_p^{(3)}(t)$ is the estimated 30-day event rate using all follow-up prior to time t. 101

A.8 Comparison of the C-statistics and mean-squared errors (MSE) of models fit to the validation set of the Azithromycin in the Prevention of COPD cohort. The Wald forward selection model and the model corresponding to clinical input from Xia et al. [2020] are fit using the modified XMT approach. The forward selection model includes predictors through automated forward selection requiring $p < 0.05$ for entry into the model. Missing values in the forward selection model and `RFRE.PO` for time since most recent exacerbation are multiply imputed using the coding scheme described in Section 2.8. C-statistic and mean-squared error estimates and standard errors for all methods and models are averaged across results from multiply imputed data with Rubin's rule applied to bootstrapped ($b = 100$) standard errors. 102

B.1 A table displaying results from simulation in the case where primary recurrent events have correlation of $\rho = 0.8$, and secondary recurrent events are independent of each other. The hazard of primary alternating recurrent event completely determines the hazard of the alternating secondary recurrent event, no subjects are censored, and results come from 500 replicates of $n = 750$ each. 110

B.2 A table displaying results from simulation in the case where primary recurrent events and secondary recurrent events are independent of each other. The hazard of primary alternating recurrent event completely determines the hazard of the alternating secondary recurrent event, approximately 30% of subjects are censored, and results come from 300 replicates of $n = 750$ each. 111
B.3 A table displaying results from simulation in the case where primary recurrent events have correlation of $\rho = 0.8$, and secondary recurrent events are independent of each other. This is the same case as 3.1, yet simulation results are calculated using the weighted bootstrap summarized in Williams and Savitsky [2021]. The hazard of primary alternating recurrent events completely determines the hazard of the alternating secondary recurrent event, approximately 30% of subjects are censored before the final follow-up window, and results come from 500 replicates of $n = 750$ each. 112
B.4 A table displaying results from the same setting as results displayed in Table 3.1; however, the standard errors are calculated via a bootstrap for weighted, clustered data [Williams and Savitsky, 2021]. The mean of the variances of bootstrap replicates is reported as the estimated variance. 114

C.1 The point estimates with corresponding standard errors for the main effects model at time $t = 350$ for the IHS data with LSTM weights. The reference levels for marital status is ""ingle, device reference is "unknown", tobacco use at baseline is "none", and alcohol use is "none." The reference level for sex is "male." History covariates of the type $\boldsymbol{H}_\ell$ are denoted with a †. 120
C.2 The point estimates with corresponding standard errors for the main effects model at time $t = 350$ for the IHS data with RF weights. The reference levels for marital status is "single," device reference is "unknown", tobacco use at baseline is "none", and alcohol use is "none." The reference level for sex is "male." History covariates of the type $\boldsymbol{H}_\ell$ are denoted with a †. 122

LIST OF APPENDICES

APPENDIX

A Appendix for Chapter 2 . . . . . **89**

B Appendix for Chapter 3 . . . . . **104**

C Appendix for Chapter 4 . . . . . **119**

ABSTRACT

Recurrent events and alternating recurrent events pose interesting statistical and clinical challenges. These data are common across a multitude of disciplines, such as: pulmonary medicine, study of (re-) hospitalizations, criminal justice, social and behavioral health, and even political economy. Furthermore, with the advent of big data, data sources containing high dimensional and highly correlated covariates are increasingly common. Analysis of these events requires solutions to three problems (censored data, correlated outcomes, and correlated covariates), while maintaining statistical fidelity.

In Chapter 2, we propose a random forest approach for dynamically predicting probabilities of remaining event-free during a subsequent τ-duration follow-up period from a reconstructed censored longitudinal data set. We compare the proposed random forest to longitudinal regression approach, modified from [Xia et al., 2020] to predict these same probabilities. We demonstrate how to incorporate past patient outcomes, even when this information is not always available at the start of follow-up. A highlight of simulation is the utility of these covariates in dynamic prediction settings, particularly when outcomes are correlated. The proposed random forest algorithm is demonstrated using recurrent exacerbation data from the trial of Azithromycin for the Prevention of Exacerbations of Chronic Obstructive Pulmonary Disease [Albert et al., 2011].

Chapter 3 introduces a wrinkle to the data structure of Chapter 2, by considering alternating recurrent events, where subjects experience two potentially correlated event types over time, the first, a primary event of interest that, when triggered, initiates a period of treatment and recovery measured via a secondary time-to-event. In this chapter, we develop a framework for regression analysis of censored alternating recurrent events that uses a random forest inverse probability weighting strategy to avoid bias in the analysis of the time to the primary event due to informative missingness from the alternate secondary state. The proposed regression model estimates τ-restricted mean time to the primary event of interest while taking into account complexities of censored data. Simulations show good performance of our method when the alternate times-to-event are either independent or correlated. We analyze mobile health study data [Carlozzi et al., 2025] to evaluate the impact of self-care push notifications on the mental state of caregivers of traumatic brain injury patients.

Chapter 4 returns to the setting of dynamic prediction for primary events, like in Chapter

2, but considers the complexity of dynamic outcome prediction in the presence of secondary events. After carefully introducing neural networks for a statistical audience, we develop a weighted neural network architecture appropriate for censored longitudinal data, and consider performance over time when compared to a semiparametric model of the type proposed by Chapter 3. Our model in Chapter 4 shows improved average predictive ability and lower average residual error sum of squares than a semiparametric model in simulation. Models are applied to dynamically predicting periods of lowered mood for medical interns based on a mobile health study. In Chapter 5, we summarize key takeaways and future directions.

CHAPTER 1

Introduction

With the advent of high-performance computing and cheap storage, machine learning algorithms have begun to permeate all disciplines from economics to criminal justice, including biostatistics. As a discipline, we are no longer constrained to analyzing covariates at baseline, or longitudinal models with pre-specified linear combinations of covariates; through rich data sources such as electronic health records, mobile health data, and insurance claims databases, models continue to gain complexity. Understanding and creating models for the advent of this big data, while remaining grounded in the scientific and medical questions are of great import to the field.

Simultaneously, with increasing computation power, models that used to be impossible to fit realistically are able to be fit within minutes, if not seconds. Particularly in the field of correlated and longitudinal data, where multiple outcomes are measured on a single subject, we now have the computational ability to model recurrent events on subjects. Furthermore, advanced predictive capabilities are often proposed and field tested by computer scientists and statisticians faster than a general audience can learn, adapt, and understand them. This thesis is concerned with models for these recurrent and alternating recurrent events. Briefly, recurrent events are events of the same type that happen to a subject over time. Some examples are asthma attacks, arthritis flareups, reinfections of flu, or blood clotting emergencies. Alternating recurrent events introduce an additional layer of complexity; after the onset of a recurrent event, there is a period of recovery, where subjects are no longer at-risk for an alternating recurrent event. For example, if the alternating recurrent event of interest is time-to hospitalization, then upon hospitalization, subjects are not at-risk for further hospitalization. If these secondary states are long, bias in analysis may be substantial.

In Chapter 2, we create a dynamic prediction model appropriate for analyzing recurrent events with the large number of patient-specific recorded covariates that are now available to modern analysts. We go beyond using data that are provided in the analytic sample, and propose the use of subject-level, individual covariates, termed "history covariates." These

covariates are highly correlated with each other, but highly informative for understanding health status of subjects. Prior to the development of methods in Chapter 2, most literature in the recurrent events space had been semi-parametric or parametric in nature, with specific assumptions about the nature of recurrent events [e.g., Andersen and Gill, 1982, Miloslavsky et al., 2004, Ang et al., 2009, Mao and Lin, 2016, Hengelbrock et al., 2016, Kong et al., 2016, Liang et al., 2017, Xu et al., 2017, Choi et al., 2017, Chan and Wang, 2017, Xia et al., 2020, Safari et al., 2023]. Semiparametric models have ease of interpretability, but often cannot accommodate highly correlated covariates, and may be subject to model misspecification. In Chapter 2, we propose a dynamic prediction model based on a longitudinal random forest [Breiman, 2001, Adler et al., 2011], as well as adapting a semiparametric model for dynamic survival probability predictions.

In Chapter 3, we introduce a complication in the recurrent events data structure; while Chapter 2 assumed negligible periods of recovery, Chapter 3 proposes a model appropriate for analyzing events when recovery times are substantial. The bias arising from recovery periods is a variation on length-sampling bias [Rao, 1965]; subjects who are healthier are at-risk for longer periods of time, while sicker subjects have longer periods of recovery. The majority of analysis of alternating recurrent events has been either state-based [such as Ionides et al., 2006, Bretó et al., 2009], parametric [such as Lee and Cook, 2019a,b], or semi-parametric in nature, focusing on the specification of hazards and intensities [see, for example, Cook et al., 2007, Hu et al., 2011]. Chapter 3 moves beyond the proportional hazards assumption, instead using the τ-restricted mean time-to first event of interest as a response. The τ-restricted mean has the advantage of clinical interpretability, as well as moving beyond the proportional hazards assumption. To correct for potential length-sampling bias, we propose an inverse weighted approach. The inverse weight is dynamically predicted, using the patient history information that was found to be so useful in prediction in Chapter 2.

Chapter 4 returns to the dynamic prediction setting of Chapter 2; however, we are concerned with dynamic prediction of alternating recurrent events. Rather than utilize a semi-parametric or parametric approach, [e.g., Fine et al., 2004, Yan and Fine, 2008, Hu et al., 2011, Loe et al., 2025b], we propose applying a long short-term memory neural network [Hochreiter and Schmidhuber, 1996, 1997], a variant of deep learning. Some statisticians have considered dynamic risk prediction [e.g., Yokota and Matsuyama, 2019, Loe et al., 2025a] or parametric and semi-parametric models [e.g., Cai and Schaubel, 2004, Mazroui et al., 2013b, Safari et al., 2023] in the recurrent event setting, while Zhao and Feng [2020] constructed a neural network and Zhao et al. [2020] applied random forest for dynamic predictions of a single time-to-event. To date, we are unaware of any authors who have proposed machine learning-based dynamic prediction models for alternating recurrent events.

1.1 Commonalities Across All Chapters

1.1.1 Survival Analysis and Censoring

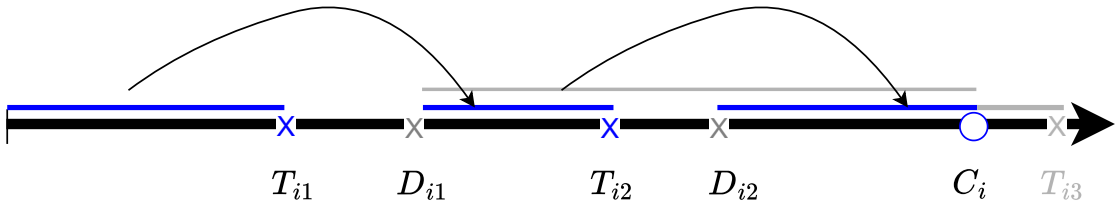
In this dissertation, outcomes are event-times, placing this dissertation in the realm of survival analysis. Complicating the analysis of time to event outcomes is the problem of censoring; that is, subjects may drop out of the study or move away. Classical survival analysis is generally concerned with time to a single event. Censoring for individual i, denoted in this thesis as C_i, means that we no longer observe subject event times, T_i, but rather the minimum of censoring time and event times. For analyzing single time-to event outcomes, modelers have historically used the Cox proportional hazards model, which models

$$\lambda_i(t) = \lambda_0(t) \exp(\mathbf{X}_i^\top \beta),$$

where $\lambda_i(t)$ is the hazard rate of an event for subject i, and $\lambda_0(t)$ is the (potentially) time-varying baseline hazard rate in the population. By modeling hazard rates, rather than event times, the Cox model provides interpretable and unbiased regression estimates in the presence of censoring. Furthermore, the Cox model allows for conditional independence of censoring and event-times, given covariates. However, this model assumes that hazard rates between different covariates levels are proportional to each other – a strong and often violated assumption

Andersen and Gill [1982] were some of the first statisticians to propose a model appropriate for analyzing multiple event times, using an extension of the proportional hazards model. They modeled the gap times between recurrent events through an intensity based framework. However, one of the key assumptions underlying their model is that events within a subject are independent. In the recurrent events setting, this is a strong assumption, more frail individuals are usually more susceptible to events than relatively healthier individuals within a disease population. When the independence assumption is violated, censoring times and gap-times becomes dependent on each other (see Figure 1.1).

Of the many estimands available in survival analysis, the τ-restricted mean was developed to aid decision making for patients and practitioners. It measures event-free time from baseline to a pre-specified τ, a clinically meaningful value [Andersen et al., 2004, Royston and Parmar, 2013]. For single time-to-event outcomes, the restricted mean is reported in the context of an event horizon, rather than conventional measures such as hazard ratios, which are often more opaque [Kim et al., 2021, Uno et al., 2014]. Clinicians have begun recommending reporting the restricted mean in conjunction with other epidemiological measures (such as the hazard ratio, or relative risk reduction) to contextualize treatment decisions



$$G_{i1} = T_{i1} \qquad G_{i,2} = T_{i,2} - D_{i,1} = T_{i,2} - \{(D_{i,1} - T_{i,1}) + T_{i,1}\}$$
$$C_i - 0 \qquad C_i - D_{i,1} = C_i - \{(D_{i,1} - T_{i,1}) + T_{i,1}\}$$

Figure 1.1: For alternating recurrent events, an example of the correlated gap time problem. Event-times of interest for subject i are recorded as $T_{i,j}$, and return to risk population j for subject i is denoted $D_{i,j}$. Gap-time analysis uses $G_{i,j} = T_{i,j} - D_{i,j-1}$ as a response. To see the dependent gap-time issue, assume $T_{i,j}, D_{i,j} \perp C_i \forall j$, but assume correlation between event-times. At the first gap-time, C_i and $G_{i,1}$ are independent of each other. However, for subsequent events, time-to censoring and and gap-times now are functions of correlated variables, and thus are no longer independent.

[Zipkin et al., 2014, Kim et al., 2017, McCaw et al., 2019, e.g.]. Furthermore, analysis of the restricted mean requires no assumptions about proportionality of hazards [Royston and Parmar, 2013]. Moreover, in a recurrent events context, the restricted mean is particularly appropriate, as patients are necessarily followed beyond the onset of the first illness. As a way to provide context for clinical decision making, the three chapters in this dissertation consider models using various τ-thresholds. Chapter 2 considers recurrent-event-free survival probability beyond a τ-length time, while Chapters 3 and 4 consider average τ-restricted mean survival. Together, these models provide dynamic prediction tools for more informed medical decision making, and more interpretable parameter estimates.

1.1.2 Censored Longitudinal Data Structure

In order to bypass the statistical issue of dependent censoring, Tayob and Murray [2015] proposed transforming data into multiple prespecified follow-up windows of length τ. They term this data structure the censored longitudinal data structure; rather than analyze gap-times, the censored longitudinal dataset analyzes time-to next recurrent event within a window. Thoughtfully chosen windows capture most recurrent events [Xia and Murray, 2019], while also avoiding the complication of dependent censoring. By creating a structure composed of repeated follow-up windows over follow-up, the censored longitudinal data structure increases efficiency, records more information on the analytic population, and can induce further cor-

relation in repeated outcomes research.

In applications, the censored longitudinal data framework naturally mirrors clinical outcomes. For example, in clinical trials and even some observational studies, it is common practice for subjects to return to the clinic for updated health status measures. At that time, a clinician or analyst may provide a dynamic prediction for the time to next recurrent event (Chapter 2), associations between updated covariates and outcomes (Chapter 3), or a dynamic prediction of the time-to next alternating recurrent event (Chapter 4). These statistical metrics can inform patient decision making from a clinical perspective, while also guiding researchers who design interventions for various subgroups of interest.

1.1.3 Pseudo-observations

The fundamental challenge in analyzing time-to-event data is censoring; each of the papers in this dissertation apply a technical device known as pseudo-observations in order to bypass concerns around censoring. Pseudo-observations are a variation of jack-knife methods, first proposed by Andersen et al. [2003] and Andersen et al. [2004]. Let $\theta = E[f(X_i)]$ be the marginal mean of an individual for which we have an unbiased estimator, $\widehat{\theta}$. While θ is a marginal quantity, it may be expressed as a mixture of conditional expectations; that is, $\theta = E[f(X_i)] = E_Z\{E_X[f(X_i)|Z_i]\} = E[\theta_i(Z_i)]$, where Z_i are measured covariates and $\theta_i(Z_i) = E_X[f(X_i)|Z_i]$. The quantity $E[f(X_i)|Z_i]$ is conditional, and one typically of interest in regression settings.

Now, instead of considering the theoretical distribution of Z, we consider θ and $E[\theta]$ over two different empirical cohorts: (1) a cohort of n subject, and (2) a cohort excluding the i^{th} subject. Specifically,

$$\widetilde{\theta} = \frac{1}{n}\sum_i \theta_i(Z_i) = \frac{1}{n}\sum_i E[f(X_i) \mid Z_i] \tag{1.1}$$

$$\widetilde{\theta}^{-i} = \frac{1}{n-1}\sum_{j\neq i} \theta_j(Z_j) = \frac{1}{n-1}\sum_{j\neq i} E[f(X_j) \mid Z_j] \tag{1.2}$$

There are two important consequences of Equations 1.1 and 1.2. First, as $n \to \infty$, $\widetilde{\theta} \to E[\theta_i(Z_i)] = \theta$ and $\widetilde{\theta}^{-i} \to E[\theta_i(Z_i)] = \theta$. Second, we may isolate the quantity of interest, $E[f(X_i) \mid Z_i]$, by clever linear combination of $\widetilde{\theta}$ and $\widetilde{\theta}^{-i}$, specifically:

$$n\widetilde{\theta} - (n-1)\widetilde{\theta}^{-i} = E[f(X_i) \mid Z_i] \tag{1.3}$$

Recall that $\widehat{\theta}$ is unbiased for θ; using $\widehat{\theta}$ in place of θ, we may use a marginal, nonparametric estimator for a (survival) quantity of interest to estimate regression outcomes.

While historically pseudo-observations have been used for diagnostic tests to study bias and variance of estimators such as the $\widehat{\beta}$ terms in ordinary least squares regression, various analysts have addressed the bivariate nature of survival outcomes using pseudo-observations in their regression models [Andersen et al., 2004, Andrei, 2005, Andrei and Murray, 2007, Xia et al., 2020], a technique each chapter in this dissertation uses. Each of the survival quantities in this dissertation have consistent, marginal nonparametric estimators based on Kaplan-Meier estimates of the survival curve. By using pseudo-observations, we are able to bypass the bivariate outcome, $\min(T_i, C_i)$, and instead use regression methods built for longitudinal, univariate outcomes, with all of their well-established asymptotic guarantees [Robins et al., 1995].

1.1.4 History Covariates

One of the unique contributions of this dissertation is to consider pertinent past subject-specific outcomes. For example, in predicting time-to-next recurrent event, some of the most useful covariates may be time since the last recurrent event, or frequency of recurrent events over a prespecified amount of time. These covariates are not always available at the start of follow-up, but the information is collected as follow-up proceeds. In Chapter 2, we introduce the concept of history covariates, and in particular, explore settings where this information is always available or partially available (after follow-up commences), and compare them to a case where an analyst never uses them. Chapter 2 introduces a multiple imputation algorithm for use in settings where history is only partially available.

Chapter 3 expands upon these covariates. Rather than using the history covariates as a means to an end for dynamic prediction purposes, Chapter 3 considers use of history covariates in predicting at-risk status for bias correction, and expands on the definitions of partial history. For example, in some settings, some patients may have full history, while others do not. Chapter 3 makes use of these history covariates in fitting an inverse weight for a semi-parametric longitudinal model fit via generalized estimating equations. The setting of Chapter 3 is not dynamic prediction, but rather post-hoc analysis of covariates on longitudinal trajectories. As this goal is retrospective in nature, we introduce a new type of history, known as "reflective" history. In fitting an inverse weight, reflective history covariates will use outcomes after or outcomes prior to time t_k but not both in crafting history covariates. This is similar to Sun and Tchetgen Tchetgen [2018], who considered the use of missing patterns in (augmented) inverse probability weighted modeling. As we shall see in Chapter 3, the implementation of partial history and reflective history have clear uses in settings where dynamic prediction of a first-stage model is of interest. Finally, Chapter 4

utilizes these well-develoepd history covariates in an online dynamic prediction framework. As higher quality history information and observations at more time points accrete, the performance of each modeling strategy considered in Chapter 4 begins to reflect the superior covariate information recorded. In each chapter, we shall see that history covariates are some of the most important and useful covariates, whether for predicting outcomes or debiasing estimates. Moreover, the modern machine learning algorithms that we develop in this dissertation easily incorporate history covariates.

1.2 Random Forest Algorithms

Each of the various chapters in this dissertation make use of random forest algorithms. Random forests were introduced by Breiman [2001] as a way to model an unknown relationship between potentially many correlated covariates and an outcome. Random forests are a non-parametric regression strategy that relies on bagging. While the various algorithms implemented vary from chapter to chapter, roughly random forest algorithms may be described (in a highly non-technical manner) as follows:

1. First, sample with replacement from the observations dataset. This is also known as bootstrapping, a strategy common in non-parametric statistics.

    - Bootstrapping allows an analyst to continually resample a population, by treating their data sample as a pseudo-population from which to derive estimators and their variances.

2. Sample from the predictors to obtain a smaller subset of predictors. This ensures that highly correlated covariates have the chance to contribute to the overall model.

3. With the bootstrapped data and the sampled covariates, create a binary regression tree by:

    (a) Recursively splitting data into two groups that minimizes an objective function.

    (b) Keep splitting the data until no further reduction in objective function, or loss, can be made, or one of several hyperparameters – number of splits, or the minimal number of subjects in a grouping has been reached.

4. Fitted values corresponding to a single tree come from the data not selected into the bootstrap sample.

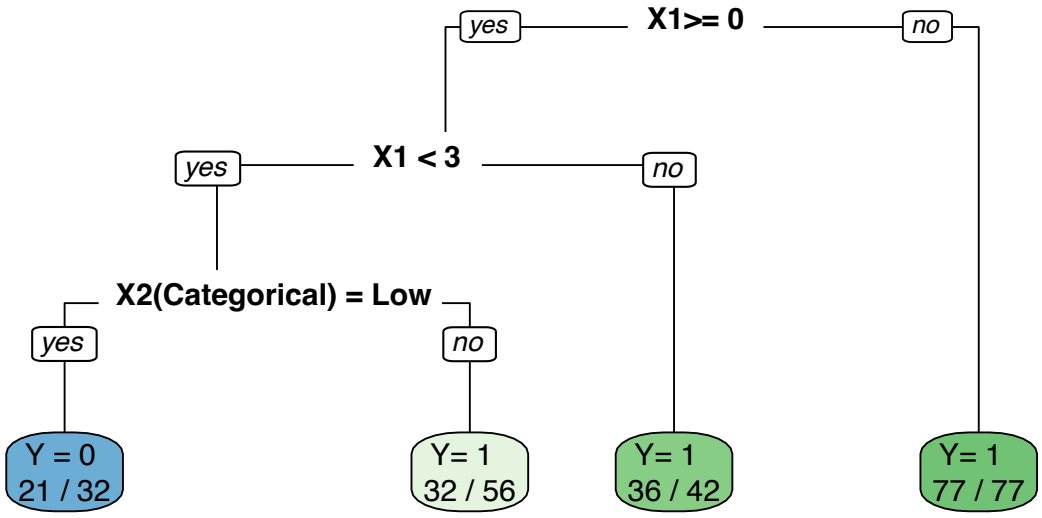


Figure 1.2: A single random tree modeling binary response Y using covariates X_1 and X_2. The regression model that corresponds to this non-parametric tree is $\text{logit}(\pi_i) = \beta_1 I(X_1 < 0) + \beta_2 I(X_1 \geq 0) I(X_1 \geq 3) + \beta_3 I(X_1 \geq 0) I(X_1 < 3) I(X_2 \neq \text{Low}) + \beta_4 I(X_1 \geq 0) I(X_1 < 3) I(X_2 = \text{Low})$. Unlike regression, if X_1 and X_2 are strongly correlated, a random tree may still be fit as tree growth does not rely on Gram matrix inversion.

The above (loose) algorithm describes a random tree; a random forest is made up of many such random trees. Figure 1.2 displays a single random tree, which makes splits on continuous covariate X_1 and categorical covariate X_2, to model binary response Y. A corresponding regression model would be $\text{logit}(\pi_i) = \beta_1 I(X_1 < 0) + \beta_2 I(X_1 \geq 0) I(X_1 \geq 3) + \beta_3 I(X_1 \geq 0) I(X_1 < 3) I(X_2 \neq \text{Low}) + \beta_4 I(X_1 \geq 0) I(X_1 < 3) I(X_2 = \text{Low})$, with estimates coming from the $p \times 2$ tables of the dichotomous covariates $I(X_1 < 0)$, $I(X_1 \geq 3)$, $I(0 \leq X_1 < 3)$ and $I(X_2 = \text{“}Low\text{”})$. Whereas a parametric modeler must pre-specify these cut points (X_1 at 0 and 3, X_2 at “Low” level), a random forest modeler may rely on the algorithm to detect the appropriate dichotomizations.

Random forests – while nonparametric, highly flexible, excellent at prediction tasks, and able to fit models in the case where $n >>> p$ – are often described as black-box, a term meant to indicate that one cannot see into the process of fitting a forest. Random forest algorithms do not provide an interpretable point estimate (i.e. $\widehat{\beta}$) with corresponding standard error. However, in an effort to provide some kind of insight into the modeling process, statisticians have proposed metrics known as variable importance scores. As we shall see, these methods depend on the type of forest and software used to fit to data, but all scores aim to provide an idea of what variables are most useful in fitting to a response. Since the introduction of the algorithm, many variations of random forests have sprung up. There are random forests for

survival data, random forests with random effects, and random forests for longitudinal data [see, for example Ishwaran et al., 2004, Pickett et al., 2021, Wager and Athey, 2018, Gerber et al., 2021, Capitaine et al., 2021, Adler et al., 2011]. As our recurrent event outcome is longitudinal, we focus on algorithms that are appropriate for longitudinal data.

1.3 Prediction vs. Inference

At various points throughout this dissertation, we will be concerned with prediction, and at other times, we will be concerned with inference. Inference and models for inference are the ones that readers who have taken a basic statistics class will be most familiar with. For example, univariate models that regress clinical outcomes on randomized treatments can infer treatment effect to a population with a disease. Models that result in causality are not the only models concerning inference. For example, any kind of point estimate with associated variance estimate with interpretable parameters is a model that provides inference about associations.

On the other hand, models for dynamic prediction are only concerned with obtaining the best possible fit values for future time points. Unlike classical (bio)statistics methods, they are not concerned with, say, estimation of treatment effect, or interpretable associations between covariates and point estimates. Prior to the advent of advanced computational methods (around the turn of the twenty-first century), models for inference may have also been used as models for prediction; more recently, advanced computing allows practitioners to utilize black box algorithms which adjust for many more covariates than traditional models were able to incorporate. In developing these models, it becomes permissible to use (1) covariates that may otherwise adjust away treatment effects, (2) covariates that are highly correlated and (3) more covariates than observations.

The distinction between prediction and inference is not always cut and dried. The goals of Chapters 2 and 4 are dynamic prediction models; as many covariates as possible are incorporated into the final models. However, each of these chapters also proposes a semiparametric model with interpretable parameters; it is the covariates that are incorporated into the models that make them appropriate for inference or for prediction. In between these dynamic prediction models, in Chapter 3, we are concerned with a semiparametric model, which yields interpretable parameters. However, the semiparametric model requires the dynamic prediction of an inverse weight for debiasing; this first stage model then becomes a dynamic prediction model. In other words, it does not matter which ingredients go into the model per se, only that the inverse probability model is highly accurate.

CHAPTER 2

Random Forest for Dynamic Risk Prediction of Recurrent Events: a Pseudo-Observation Approach*

2.1 Introduction

Recurrent events are common in healthcare settings, such as exacerbations of pulmonary diseases, or bleeding events for chemotherapy patients. Clinicians are often interested in predicting the probability of remaining event-free until a patient's next check-up. Information from events over time helps improve precision of predictions relative to analyses of a single time-to-event. Prediction is also greatly enhanced by the large number of patient-specific predictors that are collected now, for instance, in the form of electronic health records data, gene expression levels, metabolomic data, etc. The very large dimensions of the predictors and the multivariate nature of recurrent event outcomes require careful thought for analysis, particularly when censoring of recurrent events is present. A method that can utilize many potentially highly correlated risk factors of recurrent events for dynamic prediction would be of interest to modelers and practitioners.

For dynamic prediction of an upcoming recurrent event, patient history of such events is often one of the most useful predictors (typically recurrent event-times within a patient are positively correlated). A feature of recurrent event data is that each patient's event history at study time, t, becomes richer for making predictions as t increases. For instance in the Azithromycin for the Prevention of Exacerbation of COPD study [Albert et al., 2011], recurrent exacerbations were monitored over a year of follow-up; a cohort that we would like to use for making dynamic predictions of exacerbations. Unfortunately, patient history data for the most recent exacerbation time prior to study enrollment was not collected in this study. As a result, this patient history data would not be available for dynamically predicting future events until later in study. How to include patient history data in dynamic

[0]Published as: Abigail Loe, Susan Murray and Zhenke Wu. "Random Forest for Dynamic Risk Prediction of Recurrent Events," Biostatistics, vol. 26, no. 1, pp. 1–20, 2025. (See also [Loe et al., 2025a].)

predictions, when this data is only partially available at early follow-up times, has not been addressed in the recurrent-event setting.

The majority of literature in recurrent event analysis has been parametric or semi-parametric in nature, with specific modeled relationships between risk factors and recurrent event outcomes. Models have been designed for the analysis of recurrent events, gap times, and times-to-first-event across different follow-up windows, each with different assumptions imposed by the modeling paradigms [e.g., Andersen and Gill, 1982, Miloslavsky et al., 2004, Ang et al., 2009, Mao and Lin, 2016, Hengelbrock et al., 2016, Kong et al., 2016, Liang et al., 2017, Xu et al., 2017, Choi et al., 2017, Chan and Wang, 2017, Xia et al., 2020, Safari et al., 2023]. Such models have interpretable parameter estimates that can inform medical literature, generate hypotheses, and to some degree, predict outcomes. Disadvantages include limits to estimating these same parameters when there are many predictors or collinearity between important predictors that make unique solutions for corresponding parameter estimates difficult to obtain (often due to difficulty with inversion of nearly singular matrices; see, for example, Hoerl and Kennard [1970], Gunst and Webster [1975]). In addition, methods that incorrectly assume independence of gap times between recurrent events measured on an individual give biased results due to induced dependent censoring.

A few authors propose pseudo-observation approaches to parametrically or semi-parametrically model censored recurrent event data [Yokota and Matsuyama, 2019, Xia et al., 2020], replacing censored outcomes with tailor-made pseudo-observations suitable for obtaining desired parameter estimates from models available for uncensored data. Yokota and Matsuyama [2019] developed pseudo-observations for modeling probabilities that the k-th recurrent event occurred in a particular follow-up window. Xia et al. [2020] developed pseudo-observations for modeling the restricted mean event-free time in a particular follow-up window. Both of these approaches are versions of landmark analysis, where outcomes for analysis are measured relative to landmark times. Xia et al. [2020]'s landmarking approach transforms censored recurrent event data into a censored longitudinal data structure with restricted mean pseudo-observations defined over a series of follow-up windows that can be analyzed using standard longitudinal data analysis methods.

As to the goal of dynamically predicting recurrent event risk based on traditional censored recurrent event data with many potentially colinear predictors, perhaps with non-linear associations with recurrent event outcomes, we are drawn to the nonparametric nature of random forest algorithms. Machine learning methods, and in particular random forest algorithms, have gained widespread acceptance as a method of building predictions based on complex relationships of biomarkers (first proposed by Breiman et al. [2017], applied more recently in statistical literature such as Wager and Athey [2018], Zhao et al. [2020], Gerber et al.

[2021], Ishwaran et al. [2004], and reviewed in Hu and Szymczak [2023]). These algorithms naturally accommodate non-canonical relationships between covariates and outcomes, an attractive feature in an age of increasingly high-dimensional data where a traditional modeler cannot realistically find all the interactions, higher-order covariates and non-linear patterns that may be predictive of an outcome. By using information across different sets of covariates within each of the generated trees of a random forest, a modeler can side-step several traditional pitfalls of parametric and semi-parametric models: (1) multicolinearity of predictors and (2) limited ability to include a large number of covariates. In particular, Zhao et al. [2020] showed that for a single event time, imposing a censored longitudinal data structure introduced by Tayob and Murray [2015] and incorporating ideas from pseudo-observation literature, random forest regression can address nonlinear relationships between covariates and outcomes for dynamic prediction of event-free periods. Machine learning methods for censored recurrent event data are still in their infancy; we are not aware of any machine learning methods available for the analysis of recurrent event data.

In our manuscript we incorporate ideas from Xia et al. [2020] on building a censored longitudinal structure for censored recurrent event data using pseudo-observations. Our methods make several important advances from this point: (1) To better reflect the goals of dynamic prediction, we develop pseudo-observations for analysis of event-free probabilities over τ-duration follow-up windows measured from regularly spaced landmark times over the duration of the entire follow-up period. (2) We develop a modified version of the Xia, Murray and Tayob (XMT) model where the estimand of interest is the probability of being event-free over the subsequent τ-duration window, rather than a τ-restricted mean. (3) We propose and evaluate different methods for including historical event rate information in these analyses with a multiple imputation approach for addressing missing event history data prior to the initiation of follow-up, and finally, (4), we describe a novel random forest algorithm with increased prediction accuracy to perform dynamic prediction of event-free probabilities over τ-duration follow-up windows applicable to censored recurrent event data using elements from advances (1) and (3). This final random forest tool addresses many difficult issues facing analysts of recurrent event data mentioned above and we compare the performance of dynamic prediction using our modified-XMT model versus our random forest algorithm.

The rest of this chapter is organized as follows. In Section 2.2, we describe how to transform censored recurrent event data to a censored longitudinal data framework for analysis and offer examples of useful patient history covariates. In Section 2.3, we describe how to construct pseudo-observations based on the censored longitudinal data structure given in Section 2.2. Using these pseudo-observations, we describe dynamic prediction based on our modified version of the Xia, Murray and Tayob (XMT) model in Section 2.4. In Section 2.5,

again using pseudo-observations defined in Section 2.3, we describe how to perform dynamic prediction via a random forest algorithm applicable to longitudinal data (i.e., the constructed pseudo-observations). Model evaluation metrics are summarized in Section 2.6. In Section 2.7, we use simulations to compare our modified XMT regression approach and our proposed random forest approach for the analysis of censored recurrent event data, and evaluate different methods for including historical event rate information in these analyses. We apply both the modified XMT and our recurrent event random forest algorithm to data from the study of Azithromycin for the Prevention of Exacerbation of COPD study in Section 2.8. We conclude with a discussion in Section 2.9.

2.2 Notation and Censored Longitudinal Data Structure

For individual i, denote recurrent event times $T_{i,1} < T_{i,2} < \ldots$, that are potentially censored at time C_i, where C_i is independent of $T_{i,j}$, $i = 1, ..., n$, and $j \geq 1$. Independent censoring mechanisms are common, for instance, in settings where participants are identified during a finite accrual period and followed for a predetermined period (often based on research funding). Observed data for each individual is $X^*_{i,j} = \min(T_{i,j}, C_i)$, $i = 1, \ldots, n$, and $j \geq 1$. These event times are sometimes converted to gap times between recurrent events $G_{i,j} = T_{i,j+1} - T_{i,j}$, $i = 1, \ldots, n$, $j \geq 1$ for analyses. However, a well-known challenge is the dependence between gap time random variables, $G_{i,j}$, and corresponding censoring times, $C_i - T_{i,j-1}$ [Zhu, 2014].

One approach to avoiding dependent censoring in analyses of recurrent events was proposed by Tayob and Murray [2015], who transformed correlated recurrent event times into a censored longitudinal data structure. First, a series of follow-up window start times are pre-specified to $\mathcal{T} = \{t_0, t_1, t_2, \ldots\}$, with $t_j - t_{j-1} = a$, $j \geq 1$; the start times can be as close as $a = 1$ day apart, if desired. Computational speed of the analysis is usually the limiting factor in choosing a. Xia and Murray [2019] found that a equal to one-third of the mean gap time in the control group resulted in capturing 90% of the recurrent events from the original data set in at least one follow-up window. In practice, more frequent follow-up window start times have the potential for more precise estimation, but with diminishing returns once all recurrent events have been captured in at least one follow up window.

In this chapter, at each of the pre-specified "check-in time points" in $\mathcal{T}$, we will focus on estimating the probability of remaining recurrent-event-free over a subsequent follow-up window of length τ, where τ is user-specified and ideally informed by clinical expertise. Although the check-in time points are shared across subjects by design, potential censoring may result in different numbers of effective check-ins; we therefore denote the subject-specific

sets of check-in time points by $\{\mathcal{T}_i\}_{i=1}^n$, where $\mathcal{T}_i = \{t_0, t_1, ...t_{b_i}\}$. In the case of no censoring, we chose t_{b_i} so that the final follow-up interval from t_{b_i} to $(t_{b_i} + \tau)$ does not exceed the study period; in the case with censoring, we choose $t_{b_i} < C_i$. We follow Tayob and Murray [2015] and construct the censored longitudinal data that is essential for creating pseudo-observations in Section 2.3. We repeat the following two steps at all the check-in time points $t \in \mathcal{T}_i$, $i = 1, \ldots, n$:

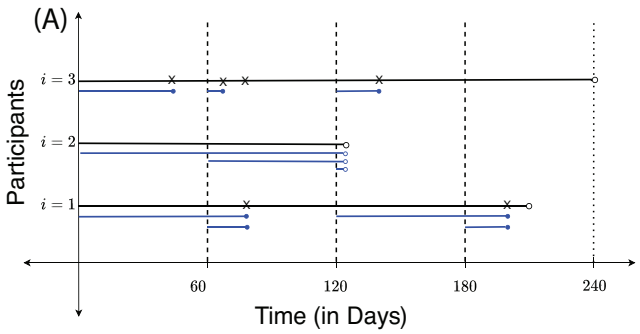
i) Identify the first recurrent event occurred after t; let $\eta_i(t) = \min\{j : X^*_{i,j} \geq t, j = 1, \ldots, n_i\}$ be its index among all the observed recurrent events for subject i. Define the residual recurrent-event-free time by $T_i(t) = T_{i,\eta_i(t)} - t$ with the corresponding censoring random variable $C_i(t) = C_i - t$, which importantly remains independent of $T_i(t)$;

ii) Construct the observed censored longitudinal data by $\{X_i(t) = X^*_{i,\eta_i(t)} - t, \delta_i(t) = I\{X_i(t) \leq C_i(t)\} : i = 1, 2, \ldots, n\}$.

Figure 2.1 uses three hypothetical subjects to illustrate the above two steps of converting traditional recurrent event data $\{X^*_{i,j}, \delta_{i,j}\}$ into the censored longitudinal data format $\{X_i(t), \delta_i(t) : t \in \mathcal{T}_i\}$. Our prediction target can now be formally expressed as $P(T_i(t) \geq \tau | \mathbf{Z}_i(t))$, i.e., at a pre-specified time point t, the probability of remaining recurrent-event-free over a subsequent follow-up window of length τ given $\mathbf{Z}_i(t)$, a set of time-invariant or time-varying predictors observed up to time t. In the following, we will occasionally suppress the subscript i and write $P(T(t) \geq \tau | \mathbf{Z}(t))$ for ease of presentation.

2.2.1 Constructing Historical Recurrent Event Information as Predictors

An important time-dependent covariate available in longitudinal studies is past recurrent event history, particularly when recurrent events within a subject are highly correlated. However, it is not always the case that history covariates are fully known at the beginning of study follow-up. Thus we consider two different types of history covariates (1) *full* history covariates, where $\mathbf{Z}_i(t)$ is known for each follow-up window starting at $t_0 \in \mathcal{T}_i$ or (2) *partial* history covariates where $\mathbf{Z}_i(t)$ is unavailable for analysis at t_0, but recorded for all $t > t_0$. In setting (2), we perform a multiple imputation analysis where, for each imputed dataset, $m = 1, \ldots, M$, random variables of interest prior to t_0 for individual i are independently sampled from an appropriate distribution, controlled by parameters that are themselves sampled from uniform distributions consistent with clinician knowledge of recurrent event behavior in the population.



(B)

i	j	$X^*_{i,j}$	$\delta^*_{i,j}$
1	1	80	1
1	2	203	1
1	3	217	0
2	1	125	0
3	1	48	1
3	2	62	1
3	3	75	1
3	4	147	1
3	5	240	0

(C)

i	t	$X_i(t)$	$\delta_i(t)$
1	0	80	1
1	60	20	1
1	120	83	1
1	180	23	1
2	0	125	0
2	60	65	0
2	120	5	0
3	0	48	1
3	60	2	1
3	120	27	1
3	180	60	0

Figure 2.1: Panel (A) displays a visualization of recurrent events in both traditional format (black) and censored longitudinal format (blue). Start times of follow-up windows ($\mathcal{T} = \{0, 60, 120, 180 \text{ days}\}$, with $a = 60$; this step does not concern τ) are highlighted by vertical dashed lines. Traditional observed recurrent event data for individuals $i = 1, 2, 3$ are observed in Panel (B), while the observed censored longitudinal data for these individuals is displayed in Panel (C). All participants were administratively censored after 8 months (240 days) of follow-up. Participant 2 has no recorded events before C_2. Additionally, while $X^*_{3,3}$ is technically observed, it is not included in the censored longitudinal data set because it is the second observation in the window, and there is no t such that $X_i(t) = X^*_{3,3} - t$. Besides potential correlation between separate event times seen in the same individual, this data structure induces correlation due to the same event contributing to the final censored longitudinal values in more than one follow-up window. For example, $X^*_{1,2}$ produces two measures in the longitudinal data structure, $X_1(120) = 83$ and $X_1(180) = 23$.

For each follow-up window starting at $t \in \mathcal{T}_i$, we define predictors based on the history recurrent events which often can be predictive of future events. Three definitions are presented below based on the motivating application in this paper, though others are possible.

First, consider a dynamic predictor that averages the observed τ^*-restricted event times at the check-in time points prior to time t. Specifically, in the history before t, identify those check-in time points: $\mathcal{T}_i^*(t) = \{t^* \in \mathcal{T}_i \cup \mathcal{T}_i^0 : t^* + \tau^* \leq t\}$, where τ^* may differ from the original τ. Here $\mathcal{T}_i^0$ is a set of check-in time points before time t_0 that are evenly spaced with a^* time units away; it is introduced here solely for constructing predictors based on history information prior to time t_0. Here a^* may differ from a used in the construction of the censored longitudinal data structure at the beginning of Section 2.2. The τ^*-restricted event time is then $\min\{T_i(t^*), \tau^*\}$. We average them over all the t^* in $\mathcal{T}_i^*(t)$ to construct:

$$H_{f,i}^{(1)}(t) = \sum_{t^* \in \mathcal{T}_i^*(t)} \min\{T_i(t^*), \tau^*\} \Big/ |\mathcal{T}_i^*(t)|. \tag{2.1}$$

Note that in Equation (2.1), at $t = t_0$, the definition assumes there exists historical recurrent event information for analysis. Otherwise, patient history prior to t_0 can be supplemented via (multiple) imputations of recurrent event times, $\widetilde{T}_{i,j}$, which are then converted to imputes of the censored longitudinal random variable, denoted as $\widetilde{T}_i(t^*) = \widetilde{T}_{i,j} - t^*$ for $t^* \in \mathcal{T}_i^*(t_0)$. We now can define a partially observed version of Equation (2.1):

$$H_{p,i}^{(1)}(t) = \sum_{t^* \in \mathcal{T}_i^*(t)} I(t^* \leq t_0) \times \min\{\widetilde{T}_i(t^*), \tau^*\} + I(t^* > t_0) \times \min\{T_i(t^*), \tau^*\} \Big/ |\mathcal{T}_i^*(t)|. \tag{2.2}$$

Second, at check-in time t, we may construct a predictor based on how long in the past τ^* period did the most recent event occur, if it exists. With the full history recurrent event information available, the time since the most recent exacerbation at t is

$$H_{f,i}^{\dagger}(t) = \begin{cases} \min\limits_{j: T_{i,j} < t} \{|t - T_{i,j}|\} & \text{if } \{j : T_{i,j} < t\} \neq \emptyset; \\ \tau^* & \text{if } \{j : T_{i,j} < t\} = \emptyset. \end{cases}$$

This information can be coded as a continuous history variable, $\min\{\tau^*, H_{f,i}^{\dagger}(t)\}$, or be coded as a categorical variable, for instance:

$$\begin{aligned} H_{f,i}^{(2)}(t) = I\left\{31 \text{ days} \leq H_{f,i}^{\dagger}(t) < 92 \text{ days}\right\} + 2 \times I\left\{92 \text{ days}^* \leq H_{f,i}^{\dagger}(t) < 183 \text{ days}\right\} + \\ 3 \times I\left\{183 \text{ days} \leq H_{f,i}^{\dagger}(t) < \tau^*\right\} + 4 \times I\left\{\tau^* \leq H_{f,i}^{\dagger}(t)\right\}, \end{aligned} \tag{2.3}$$

where increasing levels of $H^{(2)}_{f,i}(t)$ indicate longer periods without recurrent events in the past. When the full history of recurrent information is not available for analysis at t_0, one may impute $\widetilde{T}_{i,j}$ and identify the most recent imputed event time before t_0 as $\max \widetilde{T}_i$. We then define a partial history version of Equation (2.3) as:

$$H^{\dagger}_{p,i}(t) = \begin{cases} |t - \max \widetilde{T}_i| & \text{if } t \leq t_0; \\ |t - \max \widetilde{T}_i| & \text{if } t > t_0 \text{ and } \{j : T_{i,j} < t\} = \emptyset; \text{ and} \\ \min\limits_{j:T_{i,j}<t} \{|t - T_{i,j}|\} & \text{if } t > t_0 \text{ and, } \{j : T_{i,j} < t\} \neq \emptyset. \end{cases}$$

Again, similar to the $H^{(2)}_{f,i}(t)$ case, we may code this variable as the continuous version $H^{(2)}_{p,i}(t) = \min\{\tau^*, H^{\dagger}_{p,i}(t)\}$ or use a categorical form, as in Equation (2.3), replacing $H^{\dagger}_{f,i}(t)$ with $H^{\dagger}_{p,i}(t)$.

Third, the history recurrent event information predictor can be based on the frequency with which a recurrent event occurred within τ^* time units at each of the past check-in time points $t^* \in \mathcal{T}^*_i(t)$. If all the historical recurrent event information is available for analysis, we define

$$H^{(3)}_{f,i}(t) = \sum_{t^* \in \mathcal{T}^*_i(t)} I\{T_i(t^*) \leq \tau^*\} \Big/ |\mathcal{T}^*_i(t)|. \tag{2.4}$$

In the case where events prior to time t_0 are imputed as described above, we define

$$H^{(3)}_{p,i}(t) = \sum_{t^* \in \mathcal{T}^*_i(t)} [I(t^* \leq t_0) \times I\{\widetilde{T}_i(t^*) < \tau^*\} + I(t^* > t_0) \times I\{T_i(t^*) < \tau^*\}] \Big/ |\mathcal{T}^*_i(t)|. \tag{2.5}$$

Relatedly, analysis using multiply-imputed partial history variables included in the set of dynamic predictors $\mathbf{Z}_i(t)$ requires combination via Rubin's rule, where appropriate. Briefly, Rubin's rule allows an analyst to combine estimates of a parameter of interest, say ν, from separate analyses of imputed datasets, summarizing results with an overall point estimate and corresponding variance estimate [Rubin, 1987]. Let $m = 1, 2, \ldots, M$ denote the number of complete (potentially imputed) datasets with point estimates $\widehat{\nu}_m$, and estimated variance $\widehat{\sigma}^2_m$. The combined point estimate is $\bar{\nu} = \frac{1}{M} \sum_{m=1}^{M} \widehat{\nu}_m$. The pooled variance, V_t, associated with $\bar{\nu}$ has two components: the within- and the between-imputation variance. We define $V_t = \frac{1}{M} \sum_{m=1}^{M} \widehat{\sigma}^2_m + \frac{(M+1)}{M(M-1)} \sum_{m=1}^{M} (\widehat{\nu}_m - \bar{\nu})^2$. Relying on theory of asymptotic tests for large n, $\frac{\bar{\nu}^2}{V_t} \sim \chi^2_{(1)}$. For more details, see Rubin [1996].

2.3 Pseudo-Observations for Dynamic Prediction

In this section, we (1) briefly review the heuristics behind the use of pseudo-observations in censored data settings and (2) develop pseudo-observations appropriate for dynamic prediction based on the censored longitudinal data structure described in Section 2.2.

Pseudo-observations are a variation of jack-knife methodology applicable to censored time-to-event outcomes [Andersen et al., 2003]. This technical device has been used successfully for a single time-to-event and recurrent events in manuscripts such as: Andrei and Murray [2007], Hengelbrock et al. [2016], Xia et al. [2020], Zhao et al. [2020]. Essentially a pseudo-observation of a censored outcome, once created, can be used in regression contexts as if it was an uncensored outcome.

The theoretical justification behind the use of pseudo-observations typically centers on an estimand of interest for each individual i, which in our dynamic prediction setting is $P\{T_i(t) \geq \tau | \mathbf{Z}_i(t)\}$, i.e., the conditional probability of being event free in the follow-up window of length τ starting at t given an individual's covariates, $\mathbf{Z}_i(t)$, known at the beginning of this follow-up window for $i = 1, \ldots, n(t)$, where $n(t)$ is the number of individuals at risk at time t. This conditional probability can be equivalently expressed as a conditional expectation, $E[I\{T_i(t) \geq \tau\} | \mathbf{Z}_i(t)]$, where $I\{T_i(t) \geq \tau\}$ given $\mathbf{Z}_i(t)$ is a Bernoulli$[P\{T_i(t) \geq \tau | \mathbf{Z}_i(t)\}]$ random variable that is not completely observed if $T_i(t)$ is censored prior to τ, and $E[I\{T_i(t) \geq \tau\} | \mathbf{Z}_i(t)]$ is a random variable that depends on $\mathbf{Z}_i(t)$ and its corresponding probability distribution. The law of total expectation states that the mean of $E[I\{T_i(t) \geq \tau\} | \mathbf{Z}_i(t)]$ taken across the distribution of $\mathbf{Z}(t)$ is algebraically equivalent to $E[I\{T(t) \geq \tau\}] = P\{T(t) \geq \tau\}$, which is the (unconditional, marginal) expected value of $I\{T(t) \geq \tau\}$ in the absence of knowledge of $\mathbf{Z}(t)$. That is,

$$P\{T(t) \geq \tau\} = E[P\{T(t) \geq \tau | \mathbf{Z}(t)\}], \tag{2.6}$$

where the outermost expectation on the right hand side of this expression is taken with respect to the population distribution of $\mathbf{Z}(t)$.

The intuition behind creating a pseudo-observation for individual i comes from applying Equation (2.6) to two different empirically observed distributions of $\mathbf{Z}(t)$ based on the cohort of $n(t)$ individuals contributing data during the follow-up window starting at t: (1) the empirically observed distribution of $\mathbf{Z}(t)$ based on the entire cohort, $j = 1, \ldots, n(t)$, and (2) the empirically observed distribution of $\mathbf{Z}(t)$ based on this cohort, excluding individual i.

For cohort (1), Equation(2.6) becomes

$$P\{T(t) \geq \tau\} = \frac{1}{n(t)} \sum_{j=1}^{n(t)} P\{T_j(t) \geq \tau | \mathbf{Z}_j(t)\}. \tag{2.7}$$

For cohort (2), Equation(2.6) becomes

$$P^{(-i)}\{T(t) \geq \tau\} = \frac{1}{n(t)-1} \sum_{j \neq i}^{n(t)} P\{T_j(t) \geq \tau | \mathbf{Z}_j(t)\}, \tag{2.8}$$

where we use the notation $P^{(-i)}\{T(t) \geq \tau\}$ to indicate that individual i being excluded from the empirical distribution of $\mathbf{Z}(t)$ affects quantities on both sides of this expression. Taking Equations (2.7) and (2.8), together and solving for $P\{T_i(t) \geq \tau | \mathbf{Z}_i(t)\}$, we get

$$P\{T_i(t) \geq \tau | \mathbf{Z}_i(t)\} = n(t) P\{T(t) \geq \tau\} - \{n(t)-1\} P^{(-i)}\{T(t) \geq \tau\}, \tag{2.9}$$

where the left hand side of this expression is the dynamic prediction probability for patient i that we would like to obtain and the right-hand side involves marginal population quantities. The right-hand side of this Equation can be estimated completely nonparametrically using Kaplan-Meier estimates $\widehat{P}(T(t) \geq \tau)$ for $P(T(t) \geq \tau)$, and $\widehat{P}^{(-i)}(T(t) \geq \tau)$ for $P^{(-i)}(T(t) \geq \tau)$; the formulae are provided in Supplemental Materials Section A1. An appropriate pseudo-observation for dynamic prediction in the recurrent event setting is one that has approximately the same mean as $I[T_i(t) \geq \tau]$ given $\mathbf{Z}_i(t)$, the uncensored random variable of interest. The pseudo-observation,

$$\widehat{S}_i^{\tau}(t) = n(t)\widehat{P}(T(t) \geq \tau) - [n(t)-1]\widehat{P}^{(-i)}(T(t) \geq \tau), \tag{2.10}$$

has asymptotic mean, $P[T_i(t) \geq \tau | \mathbf{Z}_i(t)]$, which meets this criterion. In fact, in the special case with no censoring, all pseudo-observations, $\widehat{S}_i^{\tau}(t)$, reduce algebraically to $I(T_i(t) \geq \tau), i = 1, \ldots, n(t)$.

2.4 Semiparametric Model for Dynamic Prediction (Modified XMT Model)

Pseudo-observations, $\widehat{S}_i^{\tau}(t), t \in \mathcal{T}_i, i = 1, \ldots, n(t)$, may be used as observed outcomes in a variety of analyses typically reserved for uncensored data, including Generalized Estimating Equation (GEE) models and longitudinal random forest algorithms. Consider the marginal

model (estimated by GEE),

$$\text{logit}[\widehat{S}_i^\tau(t)] = \beta^\top \mathbf{Z}_i(t), \tag{2.11}$$

where outcomes for individual i, $\{\widehat{S}_i^\tau(t),\ \text{for } t \in \mathcal{T}_i\}$ are assumed to have an unstructured correlation matrix. Because the originally published XMT model used pseudo-observations for τ-restricted means and not our custom made pseudo-observations for dynamically predicting recurrent event probabilities, $\widehat{S}_i^\tau(t), i = 1, \ldots, n(t)$, we call model (2.11) "the modified XMT (m-XMT) model" in this chapter. Existing software can be used to fit this model. Estimated dynamic prediction probabilities from this model are $\widehat{P}[T_i(t) \geq \tau | \mathbf{Z}_i(t)] = \text{logit}^{-1}\{\widehat{\beta}^\top \mathbf{Z}_i(t)\}$. This proposed regression model is a semiparametric model for dynamic prediction of recurrent event probabilities, subject to the same limitations described for parametric and semiparametric models in the introduction; that is, difficulty in the regression specification when there are many possibly highly correlated predictors.

2.5 Random Forest for Dynamic Prediction (RFRE.PO algorithm)

Regression trees are a popular approach to incorporating multi-way interactions among predictors by finding groups of observations that are similar. A tree is grown in a few steps where at each step a new branch sorts the data leftover from the preceding step into bins based on one of the predictors. The sequential branching slices the space of predictors into rectangular partitions and approximates the true outcome-predictor relationship with the average outcome within each partition. Therefore, to grow a tree is to find bins that best discriminate among the outcomes. The specific predictor and the value split at each branching are chosen to minimize prediction error.

The number of possible trees is combinatorially large, challenging efficient global optimization. Greedy algorithms have been developed to approximate the optimal global tree by myopically optimizing prediction error at the start of each branch. We will focus on binary trees in this paper for their popularity and effectiveness. The loss associated with the prediction error for a branch is often termed "impurity" which measures how similarly the observations behave on either side of the binary split. The branching procedure halts, for example, when the number of observations in a terminal node or the number of terminal nodes reaches respective thresholds. Advantages of a regression tree include invariance to monotonic transformation of predictors, flexible approximation to potentially severe nonlinearities, and the capacity to approximate $L - 1$ way interactions for a tree with depth L.

To overcome potential overfitting, ensemble methods can be used to combine predictions from many trees into a single prediction. Random forest is such a ensemble regularizer based on Breiman's bootstrap aggregation [Breiman, 2001], or "bagging," which averages over multiple predictions obtained from trees grown from multiple bootstrap samples hence stabilizing the overall prediction. Random forest is a variation of bagging designed to further reduce the correlation among trees grown using different bootstrap samples by "dropout" which considers only a randomly drawn subset of predictors for splitting at each potential branch. Such a strategy ensures that early branches for some trees will not always split on predictors that offer the most gain in prediction accuracy. This reduces the correlation among predictions by multiple trees to further improve the variance reduction relative to the standard bagging along with reduced computational cost.

Among the many variants of random forest, we are particularly interested in random forests for longitudinal data, as recurrent event analysis uses data obtained from subjects over time. Adler et al. [2011] proposed an algorithm for tree-based ensemble methods that takes into consideration the dependence structure inherent in longitudinal data analysis. We provide a primer on this specific variation of random forest.

Let $[\widehat{S}_i^\tau(t),\ \mathbf{Z}_i(t)]$ be the data for individual, i, pertaining to the τ-duration follow-up window starting at t, where $\widehat{S}_i^\tau(t)$ is a pseudo-observation as defined in Section 2.3 and $\mathbf{Z}_i(t)$ is a set of p-dimensional covariates available at time $t \in \mathcal{T}_i$ for $i = 1, 2, ..., n$; here we use $\mathcal{T}_i$'s introduced in Section 2.2 with shared measurement timings to align more closely with our construction in Section 2.3, although the algorithm presented herein is applicable to scenarios with subject-specific measurement timings. When pseudo-observations of probability of time-to-first-recurrent-events are applied to random forests for prediction of the structure described here, we call the algorithm "`RFRE.PO`" (short for "Random Forest for Recurrent Events based on Pseudo-Observations"); we depict the analysis pipeline in Figure 2.2. Also note that the $\mathbf{Z}_i(t)$ may include time t. The following algorithm assumes $\widehat{S}_i^\tau(t) = g\{\mathbf{Z}_i(t)\} + \epsilon_i(t)$ with $\mathbb{E}[\epsilon_i(t)] = 0$ and is designed to estimate $g(\cdot)$ which describes a potentially complex relationship between the outcome and the predictors.

For $b = 1, \ldots, B$, we repeat Step 1 and 2 below ($B = 500$ is used in this chapter):

Step 1 (Bootstrap): Generate a bootstrap sample $\{(\widehat{S}_i^\tau(t), \mathbf{Z}_i(t)), (i,t) \in \mathit{Bootstrap}(b)\}$ from the original data set, where $\mathit{Bootstrap}(b)$ is the b-th bootstrap sample; $\mathit{Bootstrap}(b)$ is obtained by first sampling n subjects with replacement, followed by selecting all check-in time points $t \in \{t_0, t_1, ..., t_b\}$ for each bootstrapped subject i [Adler et al., 2011]. The data corresponding to individuals in $\mathit{Bootstrap}(b)$ is called "bagged", and data not in the sample is called "out-of-bag."

Step 2 (Grow a tree): Initialize the tree stump by grouping all observations together.

2a. (Form split variables) At a potential branching, randomly select m predictors from the p predictors with $m \ll p$; in this paper we use the default choice of $m = \sqrt{p}$ which works well in our simulation and data analysis. Let the randomly selected predictors be $\{Z_{i,s_1}(t), \ldots, Z_{i,s_m}(t)\}$ at a particular potential branching for bagged observation (i,t) where $(s_1, \ldots, s_m)$ are the indices for the subset of m predictors. For ease of presentation, we denote it by $\mathbf{W} = (W_1, \ldots, W_m)$. For a continuous or categorical observed covariate W_k, construct a grid of all observed values of the covariate, $\widetilde{W}_k = \{w_{kl}, l = 1, \ldots, L_k\}$. Consider all possible thresholds in $\widetilde{W}_k$ and their corresponding indicators, $\{I(Z_{i,s_k}(t) < w_{kl}), (i,t) \in Bootstrap(b), w_{kl} \in \widetilde{W_k}\}$ for subject i at check-in time t in the bootstrap sample $Boostrap(b)$. These derived indicator functions for continuous and categorical covariates, along with covariates in $\mathbf{W}$ that were already binary, are "potential split variables".

2b. (Choose branching): Choosing a covariate among $\mathbf{W}$ and the split value w_{kl} will result in two daughter nodes, $\{N_{w_{kl},1}, N_{w_{kl},0}\}$. Criteria for selecting the best variable and split vary by outcome type and metric of prediction accuracy. For continuous outcomes, the split with minimal sum of squared errors (SSE), is selected, though for other responses types, loss functions such as the Gini Index may be selected. For any particular node split under consideration using covariate w, that partitions individuals into two groups ($\{N_{w,0}, N_{w,1}\}$) of sizes $n_{w,0}$ and $n_{w,1}$, we minimize the criterion $SSE(w) = \sum_{c=0,1} \sum_{q=1}^{n_{w,c}} (\widehat{S}_q^{\tau}(t) - \overline{S_c})^2$, where $\overline{S_c} = \frac{1}{n_{w,c}} \sum_{q=1}^{n_{w,c}} \widehat{S}_{q,c}^{\tau}(t)$, where $c = 0, 1$.

2c. (Stopping criteria and prediction from the grown tree) Repeat Step 2a and 2b until one of the following hyperparameters has been met: minimum node size ($\min\{n_{w,c}, c = 0, 1\}$) is set to 40 in this paper) is reached, or $SSE(w)$ does not decrease any further with more branches. This results in a single binary decision tree. For each terminal node ("leaf") of the tree C, predict $P\{T_i(t) \geq \tau\}$ by the average of the outcome for observations in leaf C. Let the prediction be $\widetilde{P}_b(\mathbf{Z}_i(t); \widehat{\boldsymbol{\theta}}_b)$, where $\widetilde{\boldsymbol{\theta}}_b$ is the collection of terminal node predictions in the b-th tree, and $\mathbf{Z}_i(t)$ is the covariate profile for subject i.

Step 3 (Ensemble prediction): The final prediction for an individual i' with covariates $\mathbf{Z}_{i'}(t)$ is given by the average of all the outputs from B binary decision trees above: $\widetilde{P}_{i'}\{T_{i'}(t) \geq \tau | Z_{i'}(t)\} = \frac{1}{B} \sum_{b=1}^{B} \widetilde{P}_b(\mathbf{Z}_{i'}(t); \widehat{\boldsymbol{\theta}}_b)$.

Remark 1 *The two-stage bootstrapping scheme has a few statistical and computational ad-*

vantages consistent with results from Adler et al. [2011]: (1) subjects with many measurements are as likely to be included in the bootstrap sample as subjects with few measurements, making it unlikely that observations from subjects with longer follow-up time overpower subjects with shorter follow-up times in the tree-growing stage and (2) out-of-bag data for each tree remains independent of the bagged data, making predictions from such a forest remarkably simple.

Permutation tests are applied to the out-of-bag sample to generate p-values assessing statistical significance for predictors used within the random forest algorithm while controlling for the remaining predictors. Each variable, W, is selected and the values of W in the out-of-bag sample are permuted d times, to obtain $\mathscr{W}_s$, $s = 1, 2, ...d$. This breaks their association with $\widehat{S}_i^{\tau}(t)$. The out-of-bag samples corresponding to both $\mathscr{W}_s$ and W are then passed through the forest to obtain new predictions, $\mathcal{P}_{\mathscr{W},s}$ and $\mathcal{P}_W$ respectively. Exploiting the fact that $p_{w,\text{test}} = \dfrac{\frac{1}{d}\sum_{s=1}^{d}(\mathcal{P}_W - \mathcal{P}_{\mathscr{W},s})}{\sigma_{\mathscr{W},W}}$ has an asymptotically standard normal distribution, one can then obtain a Wald-type test of variable significance ($\sigma_{\mathscr{W},W}$ is estimated from the permutation distribution of prediction errors).

2.6 Evaluation Metrics

Both the m-XMT models and `RFRE.PO` may be evaluated using Harrell's C-statistic (also called C-index or the index of concordance), a non-parametric measure of concordance [Harrell Jr et al., 1982]. Briefly, the C-statistic compares the predicted risk scores or fit values for any two observations with the actual time-to-event for those two observations. The proportion of observations that are correctly ordered is reported as the C-statistic. Event pairs where censoring obscures the correct ordering of events within the pair are not included in the C-statistic calculation. Any model that perfectly orders patients based on some risk score has a C-statistic of 1, while a model with no predictive power has C-statistic equal to 0.50. C-statistics for events in the censored longitudinal data structure given in Section 2.2 can be reported across all events or within τ-duration follow-up windows starting at t. In addition to C-statistics, overall and time-dependent mean squared error (MSE) estimates are used to assess model calibration as described in Supplemental Materials Section A.2.

2.7 Simulation Study

As mentioned in Section 2.1, random forest algorithms are known for nonparametrically capturing complex relationships between a large number of predictors and the outcome of

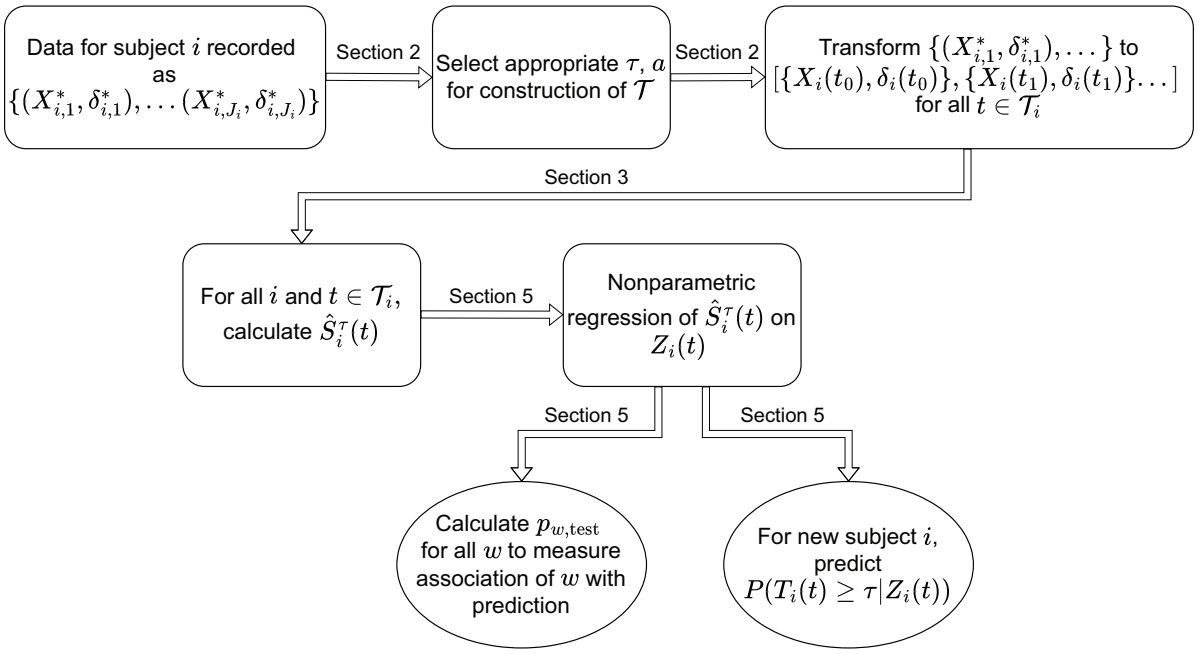


Figure 2.2: A flowchart depicting the analysis pipeline of `RFRE.PO`. Labels above arrows correspond to the section which describes the corresponding step in greater detail. An analyst begins with data recorded as recurrent event times for subject i, $j = 1, 2, \dots, J_i$ before constructing an appropriate $\mathcal{T}$, and transforming data into the censored longitudinal framework. Upon creating such a structure, the analyst calculates pseudo-observations, which are ultimately fed to a historical random forest. Recall that $p_{w,\text{test}}$ refers to the permutation test statistic described in Section 2.5, and w refers to a variable selected for permutation

interest, where these complex relationships might involve higher-order interactions, nonlinearity, and collinearity between predictors. In this section, we evaluate the ability of `RFRE.PO` to predict $P\{T_i(t) \geq \tau \mid \mathbf{Z}_i(t)\}$ in a scenario where the true relationship between $\mathbf{Z}_i(t)$ and $I\{T_i(t) \geq \tau\}$ is many times more complicated than a modeler who tends to use main effect terms would anticipate. For comparison, we also describe results from using a perfectly specified modified XMT model, labeled m-XMT (True), and a modified XMT model that only uses main effect terms for $\mathbf{Z}_i(t)$ in equation (2.11), labeled m-XMT (Main Effects).

The goals of our simulation study are (1) to verify high performance of our random forest algorithm for prediction of $P\{T_i(t) \geq \tau|\mathbf{Z}_i(t)\}$ in the recurrent event setting and (2) to demonstrate the importance of event rate history in modeling $P\{T_i(t) \geq \tau|\mathbf{Z}_i(t)\}$. As part of goal (2), for each modeling strategy [`RFRE.PO`, m-XMT (True), m-XMT (Main Effects)], we consider three different settings corresponding to availability of historical event rate information as described in Section 2.2: (1) *none*, (2) $H_p^{(1)}(t)$, or (3) $H_f^{(1)}(t)$.

We now describe the details of how data were simulated. The true relationship between recurrent events ($T_{i,1} < T_{i,2} < \ldots$) and patient characteristics (continuous covariates Z_1, Z_2, Z_4, and categorical covariates Z_3, Z_5, Z_6, Z_7) is designed to be intentionally complex. Recurrent event gap-times marginally follow an exponential distribution with hazard: $\lambda_i = \exp\{Z_{i2}\sin(Z_{i1}/Z_{i6}) + (-1)^{I(Z_{i2}>2)}Z_{i3} + Z_{i1}Z_{i6} + Z_{i2}^2 Z_{i4}\}$. We simulate $Z_1, Z_2 \sim \text{MVN}(\mu, \Sigma)$, where $\mu = (0,2)^\top$ and $\Sigma = \begin{pmatrix} 5 & 0.7 \\ 0.7 & .8 \end{pmatrix}$, $Z_3 + 2 \sim \text{Poisson}(4)$, $Z_4 - 0.1 \sim \text{Beta}(7,1)$, $Z_5 \in \{0,1,2\}$ with probability $\left(\frac{2}{3}, \frac{1}{6}, \frac{1}{6}\right)$, $Z_6 \in \{-5,-2,2,3\}$ with probability $\left(\frac{1}{10}, \frac{1}{3}, \frac{11}{30}, \frac{1}{5}\right)$, and $Z_7 \in \{0,1,2,3\}$ with equal probability. The sign in front of Z_3 depends on the value of Z_2 which is correlated with Z_1. Z_5 and Z_7 are noise covariates. Individuals with simulated hazards corresponding to mean times-to-recurrent event outside of the range of $\frac{1}{15}$ and $\frac{15}{8}$ years are removed from the sample and replaced. We induce correlation among exponentially distributed gap times for each subject i by utilizing the Gaussian copula method as described by Xia et al. [2020], for $\rho = 0, 0.3, 0.6$ or 0.9. The resultant recurrent event times are converted to a censored longitudinal data format as described in Section 2.2. Censoring is independently generated from an exponential distribution so that 0%, 23%, 45%, or 63% of individuals are censored.

The censored longitudinal data structure assumes $\mathcal{T} = \{0, \frac{1}{12}, \frac{1}{6}, \ldots 2\}$ years, and $\tau = \frac{1}{6}$ year. For individual i, $\mathcal{T}_i$ is the subset of follow-up window start-times in $\mathcal{T}$ where individual i remains at risk for recurrent events. Both the m-XMT (True) model and the m-XMT (Main Effects) estimate $\text{logit}\{E(S_i^\tau(t))\}$ assuming the relationship $\text{logit}\{E(S_i^\tau(t))\} = \beta^\top \mathbf{Z}_i^*(t)$. For m-XMT (True), the ad hoc nonlinear model, $\mathbf{Z}_i^*(t) = \{Z_{i2}\sin(Z_{i1}/Z_{i6}),\ (-1)^{I(Z_{i2}>2)}Z_{i3},\ Z_{i1}Z_{i6},\ Z_{i2}^2 Z_{i4}\}^\top$. The m-XMT (Main Effects) model has covariate vector $\mathbf{Z}_i^*(t) = \{Z_{i1},$

$Z_{i2},\ Z_{i3},\ Z_{i4},\ Z_{i5},\ Z_{i6},\ Z_{i7}, t\}^\top$. Covariate histories are included as additional patient characteristics in $\mathbf{Z}_i^*(t)$ for full and partial history availability settings. For full history availability settings this predictor takes the form $H_{f,i}^{(1)}(t)$ as given in Equation (2.1) with $\mathcal{T}_i^*(t) = \{\frac{-1}{12}\} \cup \{t_k \in \mathcal{T}_i : t_k + \tau^* < t\}$, and $\tau^* = \frac{1}{12}$. For partial history availability settings, imputes for missing event histories prior to $t = 0$ follow recommendations in Section 2.2 with $\widetilde{T}_{i,j} \sim$Exp(λ_i), where $\lambda_i \sim$Unif$(\frac{8}{15}, 15)$. Partial history covariates are added to $\mathbf{Z}_i^*(t)$ based on Equation (2.2), with the same $\mathcal{T}_i^*(t)$ and τ^* as $H_f^{(1)}(t)$. All finite sample performance results are based on 500 simulated datasets of n=500 participants per scenario.

Figures 2.3A and 2.3B display simulated model fit for (1) the modified XMT model using main effects terms only {m-XMT (Main Effects), shown in purple}, (2) our random forest algorithm for recurrent events (RFRE.PO, shown in green) and (3) the true modified XMT model {m-XMT (True), shown in yellow}. Panel (A) displays violin plots of C-statistics (indices of concordance) seen in simulation with corresponding measures of center and spread shown in Table 2.1. For partial history information, average C-statistics from $m = 10$ imputed data sets are shown. Within a box, historical covariate information increases from left to right. Panel (B) expands on the scenario boxed in red from panel (A), i.e., the low censoring setting with 0.3 correlation. Here violin plot C-statistic distributions are time-dependent, reflecting model fit for $\tau = \frac{1}{6}$ duration follow-up windows starting at times, t, marked on the horizontal axis. The quality of historical information increases from top to bottom. The m-XMT (True) model performance provides a benchmark for assessing performance of the proposed (truth-blind) dynamic prediction methods, the m-XMT (Main Effects) model and the `RFRE.PO` algorithm.

Panel A results show that when there is no historical covariate used for dynamic prediction, the m-XMT (True) model has the best C-statistic performance, as expected, for all levels of censoring and correlation, followed by the `RFRE.PO` algorithm and then the m-XMT (Main Effects) model. As correlation between recurrent event times increases, use of either partial or full history covariates in these latter two dynamic prediction algorithms brings C-statistic performance closer to that of the m-XMT (True) model. When using full history covariates, `RFRE.PO` outperforms the m-XMT (Main Effects) model for $\rho = 0.3$ and performs similarly for $\rho = 0.6$. When using partial history covariates, `RFRE.PO` outperforms m-XMT (Main Effects) for both $\rho = 0.3$ and $\rho = 0.6$. For $\rho = 0.9$, use of either partial or full history covariates in the main effects model slightly outperforms both the `RFRE.PO` algorithm and the m-XMT (True) model.

Generally, the inclusion of full history covariates is more effective than that of partial history covariates, though the benefit of having full history covariates available is most strongly seen in the lowest two correlation settings when applying the m-XMT (Main Effects) model.

In the highest correlation setting, partial and full covariate history algorithms have very similar performance. In the absence of correlation between recurrent event times, use of history covariates shows modest, if any, improvement for the `RFRE.PO` algorithm C-statistic whereas C-statistic performance for the m-XMT (Main Effects) model deteriorates noticeably.

Time-dependent C-statistic performance highlighted in Panel B shows how quickly algorithms using partial and full covariate history information improve and stabilize over the course of follow-up. Stabilization happens quickly, with little change in time-dependent follow-up windows after $t = 5$. Because covariate history variables improve with increasing t, there is an induced interaction between these predictors and t, particularly when the algorithms are using the partial history covariates. This type of interaction is naturally assessed by the `RFRE.PO` approach.

Together, Figures 2.3A and 2.3B indicate that in highly correlated settings, (1) patient history is an essential predictor to include in dynamic prediction algorithms and (2) the `RFRE.PO` algorithm seems to outperform the m-XMT (Main Effects) algorithm in all but the very highest correlation setting where history covariates are used for dynamic prediction. Additional simulation results are available in Supplemental Materials. Section A.3 summarizes mean squared error results with corresponding Table 2.2, while Section A.4 describes the ability of these algorithms to distinguish signal covariates (Z_1, Z_2, Z_3, Z_4, Z_6) versus noise covariates (Z_5, Z_7) in Tables A.3 and A.4.

2.8 Prediction of COPD Exacerbations

Chronic obstructive pulmonary disease (COPD) is most often attributed to smoking history, although cases may also be caused by breathing polluted air such as biomass fuel or industrial chemicals [Safiri et al., 2022]. COPD is characterized by periods of relative stability punctuated with episodic exacerbations of coughing, struggling for breath, and/or wheezing that may lead to hospitalization, in which case exacerbations are called "severe." Such exacerbations increase overall disease progression, eventually leading to lung transplantation or death [Seemungal et al., 1998, Donaldson et al., 2002], indicating some correlation between events within a subject. It is of clinical interest to dynamically predict the chances of being exacerbation-free over time.

Azithromycin for the Prevention of COPD [Albert et al., 2011] was a multi-center clinical trial evaluating the effect of daily oral azithromycin (250mg) in preventing severe COPD exacerbations during a year of follow-up. We apply `RFRE.PO` to predict the probability of remaining exacerbation-free over a subsequent $\tau = 180$ day follow-up window in this cohort ($n = 1035$ study participants with complete baseline data and longitudinal sleep study

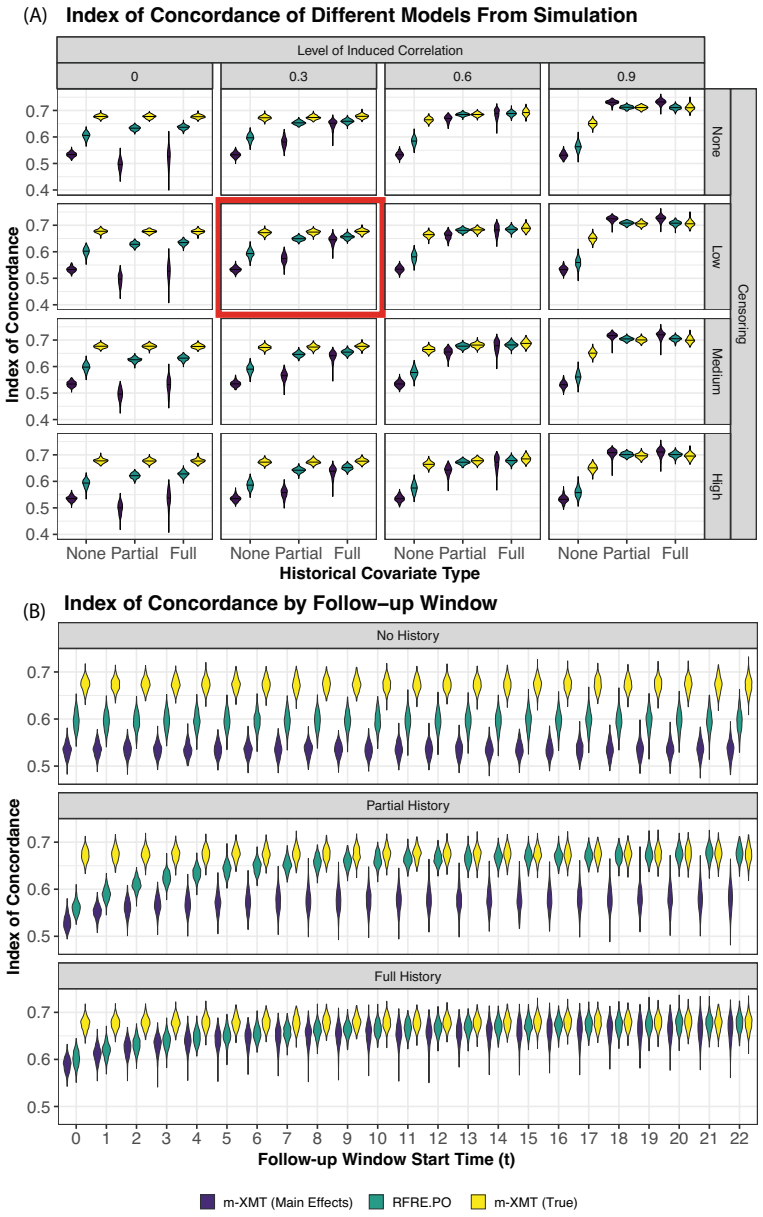


Figure 2.3: Simulated model fit for (1) the modified XMT model using main effects terms only {m-XMT (Main Effects), shown in purple}, (2) our random forest algorithm for recurrent events (RFRE.PO, shown in green) and (3) the true modified XMT model {m-XMT (True), shown in yellow}. Results are based on 500 simulated datasets of n=500 participants per scenario. Panel (A) displays violin plots of C-statistics (indices of concordance) seen in simulation. For partial history information, average C-statistics from $m = 10$ imputed data sets are shown. Measures of center and spread are in Table 2.1. Within a box, historical covariate information increases from left to right. Panel (B) expands on the scenario boxed in red from panel (A), i.e., the low censoring setting with 0.3 correlation. Here violin plot C-statistic distributions are time-dependent, reflecting model fit for $\tau = \frac{1}{6}$ duration follow-up windows starting at times, t, marked on the horizontal axis. The quality of historical information increases from top to bottom.

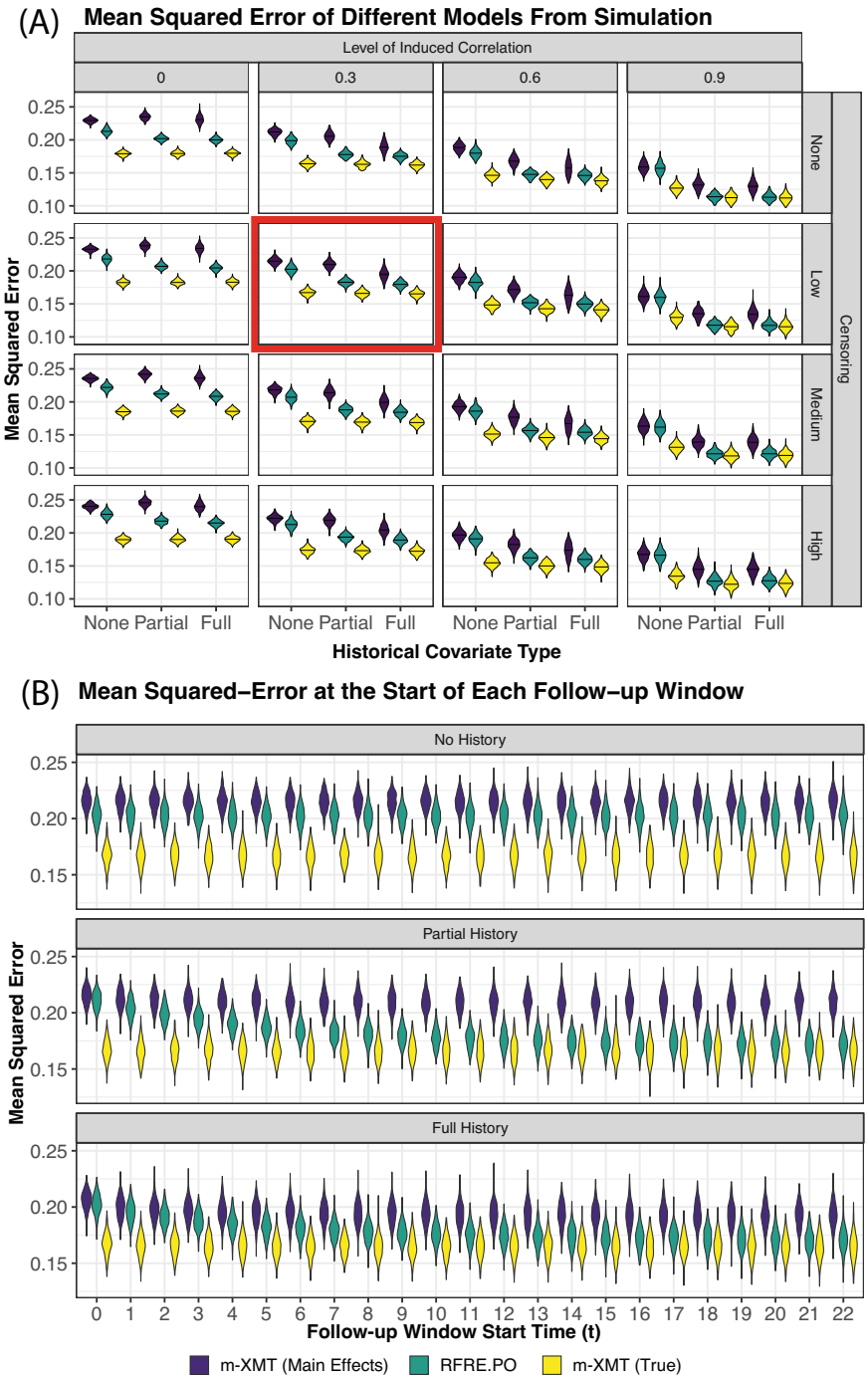


Figure 2.4: Panel (A) displays the mean squared error (MSE) for simulation settings displayed in Figure 2.3. Censoring increases from top to bottom, and the amount of induced correlation increases from left to right. Within a box, the quality of historical information increases from left to right. Panel (B) shows the MSE at the start of each follow-up window for the setting boxed in red in Panel (A). From top to bottom, more historical information is added in Panel (B), and the MSE corresponding to `RFRE.PO` in $H_p^{(1)}(t)$ settings. 500 replicates were performed, each with $n = 500$ subjects.

variables). There are 962 available covariates, including demographic variables, social well-being surveys, sleep quality metrics and clinical characteristics such as forced vital capacity (FVC), and medication use. There is strong correlation between variables as some are derived from each other, making `RFRE.PO` an attractive alternative to model-based predictions (see Figure 2.5). Censored longitudinal data is constructed with $\mathcal{T} = \{0, 30, 60, ..., 180\}$, and used for all analyses, with $T_i = \{0, 30, \dots t_{b_i}\}$ for individual i.

Baseline information on recurrent event times prior to $t = 0$ were not available. We defined a full history covariate, $H_{f,i}(t)$, as I(Severe exacerbation recorded prior to t), with $H_{f,i}(0)$=0. We also considered two partial covariate history variables from Section 2.2, the categorical version of $H^{(2)}_{p,i}(t)$ given in Equation 2.3, with $H^{\dagger}_{p,i}(t)$ replacing $H^{\dagger}_{f,i}(t)$, and $H^{(3)}_{p,i}(t)$. For categorical $H^{(2)}_{p,i}(t)$, we assume $\tau^* = 1$ year and $\mathcal{T}^*_i(t) = \{t^* \in \mathcal{T}_i \cup -1 \text{ year} : t^* + 1 \text{ year} \leq t\}$. This variable categorizes $H^{\dagger}_{p,i}(t)$ in the same manner Equation 2.3 categorizes $H^{\dagger}_{f,i}(t)$ with higher values indicating longer exacerbation-free periods leading up to t. For $H^{(3)}_{p,i}(t)$, we assume $\tau^* = 30$ days and $\mathcal{T}^*_i(t) = \{\mathcal{T}_i \cup -30 \text{ days} : t^* + 30 \text{ days } \leq t\}$, so that this variable estimates the probability that prior follow-up windows had an exacerbation within 30 days. Imputed event times prior to time 0 were sampled from an exponential(λ_i) distribution with $1/\lambda_i$ generated from a uniform(100 day, 180 day) distribution, which was comparable to placebo group outcomes. Rubin's rule [Rubin, 1987] was used to combine inference across multiply imputed datasets as described in Section 2.2.

For comparison purposes, we present `RFRE.PO` results alongside results from two m-XMT models. The first m-XMT model (Clinical Input) includes baseline variables used by Xia et al. [2020] for this cohort; these predictors are commonly used to assess confounder-adjusted treatment effects in COPD journals. The second m-XMT (Forward Selection) model is based on Wald forward selection applied to the first multiply imputed data set with $p < 0.05$ required for entry; covariates that introduce model instability are discarded from consideration. Prediction algorithms for `RFRE.PO` and the forward selection model were applied to a training cohort (70% of original cohort) and all algorithms were evaluated via the C-statistic in a validation cohort (remaining 30% of original cohort). For any particular multiply imputed validation dataset, standard errors of the C-statistic are calculated via the bootstrap with $b = 100$ bootstrap samples. C-statistics for models involving multiply imputed history variables are combined across multiply imputed datasets using Rubin's rule. Parameter estimates and standard errors for m-XMT (Forward Selection), which involves the imputed partial history covariates, are combined across multiply imputed validation datasets via Rubin's Rule. Permutation tests associated with the `RFRE.PO` algorithm predictors are calculated within the training dataset, using out-of-bag samples as described in Section 2.5.

For m-XMT (Clinical Input), where estimation of treatment effect, and not necessarily

prediction was the goal, predictors are baseline forced expiratory volume in one second in liters (Baseline FEV_1), age in decades (Age10), male gender $\{I(\text{Male})\}$, and baseline smoking status $\{I(\text{Smoker})\}$:

$$\text{logit}\{E(S_i^{\tau}(t))\} = \beta_0 + \beta_1 I(\text{Treatment})_i + \beta_2(\text{Baseline FEV}_1)_i + \beta_3(\text{Age10})_i + \beta_4 I(\text{Male})_i + \beta_5 I(\text{Smoker})_i,$$

where parameter estimates for the m-XMT model are displayed in Supplemental Table A.5.

The m-XMT model based on Wald forward selection, is:

$$\begin{aligned}\text{logit}\{E(S_i^{\tau}(t))\} = {} & \beta_0 + \beta_1 I\{\text{Categorical } H_{p,i}^{(2)}(t) = 0\} + \beta_2 I(\text{Hospitalization before } t = 0)_i \\ & + \beta_3 I(\text{Corticosteroid use in year before } t = 0)_i + \beta_4 H_{f,i}(t) + \beta_5 t + \beta_6 I(\text{Male})_i \\ & + \beta_7 I(\text{Race = Black})_i + \beta_8 I(\text{Sudden feelings of panic at time } t)_i \\ & + \beta_9 I(\text{Sudden feelings of "butterfly" fright at time } t)_i \\ & + \beta_{10} I(\text{Seen mental health provider at time } t)_i + \beta_{11} I(\text{Has pain at time } t)_i \\ & + \beta_{12}(\text{St. George Respiratory Questionnaire symptom score at time } t)_i \\ & + \beta_{13} I(\text{Antiarrhythmic use at time } t)_i + \beta_{14}(\text{FEV}_1/\text{FVC \% predicted at time } t)_i \\ & + \beta_{15} I(\text{Azithromycin treatment group})_i \\ & + \beta_{16} I(\text{Inhaled corticosteriod or long-acting beta-agonist use at time } t)_i + \\ & \beta_{17} I(\text{Other baseline medications})_i + \beta_{18}(\text{Supplemental oxygen use at time } t)_i + \\ & \beta_{19} I(\text{Mixed beta-blocker use at time } t)_i + \beta_{20} I(\text{Bronchiectasis diagnosis at time } t)_i\end{aligned}$$

with parameter estimates displayed in Supplemental Table A.6.

For the `RFRE.PO` algorithm, Supplemental Table A.7 highlights predictors that had a statistically significant permutation test Z-score for at least one of the multiply imputed training datasets. Additional predictors that did not achieve statistical significance via the permutation test, but were used in the `RFRE.PO` algorithm in some form, are not shown.

Figure 2.6 displays validation dataset Z-scores for predictors included in the m-XMT (Clinical Input) and m-XMT (Forward Selection) algorithms as well as out-of-bag sample Z-scores for `RFRE.PO` predictors highlighted in Supplemental Table A.7. In algorithms that included history covariates {m-XMT (Forward Selection) and `RFRE.PO`}, these predictors showed strong validation results in Figure 2.6. When prediction algorithms were applied to individuals in the validation cohort, the `RFRE.PO` algorithm outperformed both the m-XMT (Clinical Input) and m-XMT (Forward Selection) algorithms in terms of both the C-statistic (C-statistic 0.613 compared to 0.564 and 0.571, respectively) and estimated mean

squared-error (MSE 0.242 compared to 0.252 and 0.264); See Table A.8. Time-dependent C-statistics and $MSE(t)$ for follow-up windows starting at time, t, in the validation dataset are displayed in Figure A.2 for the various prediction algorithms. Compared to the other algorithms, `RFRE.PO` has fairly stable and attractive performance metrics across the various follow-up windows, likely because interactions with the various predictors over time are naturally incorporated if useful for prediction.

As can be seen in Figure 2.6, variables identified via m-XMT models are for the most part identified as important variables in the `RFRE.PO` algorithm, although the version of the predictors may vary. For example, variables related to the long-acting muscarinic or beta-blockers, and/or steroid treatment are incorporated in different forms into these models (highlighted on the y-axis in purple). Similarly, different variables relating to patient well-being (highlighted in orange) are featured in the m-XMT (Forward Selection) model (has pain at time t) versus the `RFRE.PO` algorithm (pain score at time t and standardized pain score at time t), as well as different measures of pulmonary function (highlighted in pink). The `RFRE.PO` algorithm highlights a few variables that were not used in the forward selection model, including general health score at time t and leukocytosis present at time t. The `RFRE.PO` algorithm has no problem including information from correlated predictors; for instance `RFRE.PO` included all history variables, whereas this introduced model instability in the m-XMT (Forward Selection) model. See Figure 2.7 for predictions and history covariates for the Azithromycin for the Prevention of COPD cohort.

The `RFRE.PO` algorithm includes predictors, with corresponding split decisions, that minimize squared error of the predictions relative to the pseudo-observations. This algorithm is completely agnostic to statistical significance and scientific experience, so long as the squared error within nodes along cut-points decreases, so validation checks are quite important. With variable power to detect statistical significance of useful predictors, methods that require statistical significance for model inclusion may be disadvantaged in terms of prediction when compared to the `RFRE.PO` algorithm that does not have this requirement. And although the `RFRE.PO` algorithm may occasionally include a predictor that is not useful in the validation set (type I error analog), it has a much better chance of including predictors that are useful but would not have met statistical significance (akin to a type II error) in traditional statistical model building environments. If the scientific goal of analysis is prediction, the `RFRE.PO` approach seems remarkably simple and effective.

2.9 Discussion

We presented two novel approaches, `RFRE.PO` and the m-XMT model, for dynamically estimating probabilities of being event-free across a sequence of follow-up windows. An additional novelty is our advice for incorporating covariate history into prediction algorithms, with a multiple imputation approach when covariate history is not known prior to time $t = 0$. For purposes of prediction, `RFRE.PO` tends to outperform the m-XMT (Main Effects) model that a traditional modeler might first attempt. However, the m-XMT model has the advantage of interpretable parameters and a clear formula for prediction that offers more transparency to users, and may accommodate further interactions than explored in the m-XMT (Main Effects) model.

Our presentation did not touch on how causal relationships might be evaluated using these methods. In the Azithromycin for Prevention of COPD study, assessment of the treatment effect disallows using information on time-dependent predictors that could have been changed by treatment after time $t = 0$. To do so would adjust away part of the treatment effect under study. Similarly, clinical trial analyses may require adjustment for baseline confounders even if they do not improve prediction metrics. The m-XMT model allows for more intentional use of covariates to match scientific goals of analysis when compared to the `RFRE.PO` algorithm. The m-XMT (Clinical Input) model was the only appropriate model for assessing the effect of azithromycin presented in this chapter.

The growing popularity of machine learning tools is likely to dominate analyses of high-dimensional data, and the `RFRE.PO` algorithm contributes to this literature as one of the first machine learning methods to be able to make dynamic predictions from recurrent event data. There are tuning parameters in random forest algorithms that may potentially be tuned for better performance, warranting further research. Some authors have studied parameter tuning [e.g., Scornet, 2018], but we have not seen work for longitudinal random forests to date. The development of (1) the theory of tuning parameters for longitudinal random forests and (2) the potential application of the censored longitudinal data structure for dynamic prediction to other machine learning algorithms are all important future research areas.

Table 2.1: Simulated C-statistics listed as Mean (ESD) for varying prediction algorithms when Z_1, Z_2 are correlated with each other, censoring percentages (none is 0%, light is 23%, moderate is 45%, and heavy is 63%), and gap-time correlations ($\rho = 0$, 0.3, 0.6, 0.9). 500 replicates were performed with $n = 500$ for each replicate, and the same data is reported in Figure 2.3.

	$\rho = 0$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.9$
		No Censoring		
m-XMT (Main Effects)	0.534 (0.009)	0.533 (0.010)	0.533 (0.010)	0.531 (0.011)
`RFRE.PO`	0.605 (0.012)	0.596 (0.014)	0.584 (0.015)	0.563 (0.016)
m-XMT (True)	0.677 (0.006)	0.673 (0.007)	0.665 (0.008)	0.651 (0.010)
m-XMT (Main Effects) with $H_p^{(1)}(t)$	0.495 (0.021)	0.581 (0.018)	0.668 (0.012)	0.730 (0.008)
`RFRE.PO` with $H_p^{(1)}(t)$	0.633 (0.007)	0.653 (0.007)	0.685 (0.006)	0.712 (0.006)
m-XMT (True) with $H_p^{(1)}(t)$	0.677 (0.007)	0.674 (0.007)	0.685 (0.007)	0.711 (0.007)
m-XMT (Main Effects) with $H_f^{(1)}(t)$	0.522 (0.035)	0.648 (0.017)	0.687 (0.017)	0.732 (0.010)
`RFRE.PO` with $H_f^{(1)}(t)$	0.636 (0.008)	0.659 (0.008)	0.688 (0.008)	0.711 (0.007)
m-XMT (True) with $H_f^{(1)}(t)$	0.677 (0.006)	0.678 (0.007)	0.692 (0.010)	0.711 (0.010)
		Light Censoring		
m-XMT (Main Effects)	0.533 (0.009)	0.533 (0.009)	0.534 (0.010)	0.532 (0.011)
`RFRE.PO`	0.601 (0.015)	0.593 (0.014)	0.580 (0.017)	0.562 (0.017)
m-XMT (True)	0.677 (0.007)	0.672 (0.008)	0.664 (0.009)	0.650 (0.011)
m-XMT (Main Effects) with $H_p^{(1)}(t)$	0.496 (0.023)	0.574 (0.018)	0.668 (0.012)	0.724 (0.009)
`RFRE.PO` with $H_p^{(1)}(t)$	0.628 (0.008)	0.649 (0.007)	0.681 (0.007)	0.708 (0.006)
m-XMT (True) with $H_p^{(1)}(t)$	0.676 (0.006)	0.674 (0.007)	0.682 (0.008)	0.705 (0.008)
m-XMT (Main Effects) with $H_f^{(1)}(t)$	0.526 (0.032)	0.643 (0.017)	0.681 (0.019)	0.725 (0.011)
`RFRE.PO` with $H_f^{(1)}(t)$	0.634 (0.008)	0.656 (0.008)	0.685 (0.008)	0.708 (0.008)
m-XMT (True) with $H_f^{(1)}(t)$	0.677 (0.007)	0.676 (0.008)	0.689 (0.011)	0.706 (0.011)
		Moderate Censoring		
m-XMT (Main Effects)	0.534 (0.010)	0.535 (0.010)	0.535 (0.011)	0.532 (0.012)
`RFRE.PO`	0.597 (0.015)	0.590 (0.015)	0.577 (0.016)	0.558 (0.017)
m-XMT (True)	0.677 (0.007)	0.672 (0.008)	0.664 (0.009)	0.649 (0.012)
m-XMT (Main Effects) with $H_p^{(1)}(t)$	0.495 (0.022)	0.564 (0.018)	0.661 (0.013)	0.716 (0.011)
`RFRE.PO` with $H_p^{(1)}(t)$	0.625 (0.009)	0.646 (0.009)	0.677 (0.008)	0.704 (0.007)
m-XMT (True) with $H_p^{(1)}(t)$	0.677 (0.007)	0.674 (0.008)	0.681 (0.008)	0.700 (0.008)
m-XMT (Main Effects) with $H_f^{(1)}(t)$	0.529 (0.030)	0.638 (0.017)	0.676 (0.022)	0.719 (0.013)
`RFRE.PO` with $H_f^{(1)}(t)$	0.632 (0.009)	0.654 (0.008)	0.682 (0.009)	0.705 (0.008)
m-XMT (True) with $H_f^{(1)}(t)$	0.676 (0.007)	0.676 (0.008)	0.688 (0.011)	0.700 (0.010)
		Heavy Censoring		
m-XMT (Main Effects)	0.535 (0.010)	0.535 (0.011)	0.535 (0.012)	0.532 (0.012)
`RFRE.PO`	0.592 (0.015)	0.585 (0.016)	0.569 (0.017)	0.551 (0.017)
m-XMT (True)	0.677 (0.008)	0.673 (0.008)	0.664 (0.010)	0.649 (0.012)
m-XMT (Main Effects) with $H_p^{(1)}(t)$	0.498 (0.022)	0.557 (0.018)	0.641 (0.015)	0.707 (0.014)
`RFRE.PO` with $H_p^{(1)}(t)$	0.621 (0.009)	0.642 (0.008)	0.672 (0.008)	0.701 (0.007)
m-XMT (True) with $H_p^{(1)}(t)$	0.677 (0.008)	0.673 (0.008)	0.678 (0.009)	0.695 (0.009)
m-XMT (Main Effects) with $H_f^{(1)}(t)$	0.535 (0.031)	0.635 (0.019)	0.669 (0.025)	0.709 (0.017)
`RFRE.PO` with $H_f^{(1)}(t)$	0.628 (0.010)	0.652 (0.009)	0.678 (0.010)	0.701 (0.009)
m-XMT (True) with $H_f^{(1)}(t)$	0.676 (0.008)	0.676 (0.008)	0.685 (0.011)	0.695 (0.012)

Table 2.2: Simulated MSE listed as Mean (ESD) for varying prediction algorithms considered in Section 2.7, with censoring percentages (None, Light, Moderate, and Heavy), and gap-time correlations ($\rho = 0$, 0.3, 0.6, 0.9). 500 replicates were performed with $n = 500$ for each replicate.

	$\rho = 0$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.9$
		No Censoring		
m-XMT (Main Effects)	0.229 (0.003)	0.212 (0.005)	0.188 (0.006)	0.159 (0.008)
`RFRE.PO`	0.213 (0.005)	0.199 (0.005)	0.180 (0.006)	0.157 (0.008)
m-XMT (True)	0.179 (0.003)	0.164 (0.004)	0.146 (0.005)	0.127 (0.006)
m-XMT (Main Effects) with $H_p^{(1)}(t)$	0.235 (0.005)	0.206 (0.006)	0.168 (0.007)	0.131 (0.008)
`RFRE.PO` with $H_p^{(1)}(t)$	0.202 (0.003)	0.178 (0.005)	0.147 (0.005)	0.114 (0.005)
m-XMT (True) with $H_p^{(1)}(t)$	0.179 (0.004)	0.163 (0.005)	0.140 (0.005)	0.112 (0.006)
m-XMT (Main Effects) with $H_f^{(1)}(t)$	0.231 (0.007)	0.189 (0.009)	0.157 (0.011)	0.130 (0.009)
`RFRE.PO` with $H_f^{(1)}(t)$	0.200 (0.004)	0.175 (0.004)	0.146 (0.005)	0.113 (0.005)
m-XMT (True) with $H_f^{(1)}(t)$	0.180 (0.004)	0.162 (0.004)	0.138 (0.005)	0.112 (0.006)
		Light Censoring		
m-XMT (Main Effects)	0.233 (0.004)	0.215 (0.005)	0.190 (0.006)	0.161 (0.008)
`RFRE.PO`	0.218 (0.005)	0.203 (0.006)	0.183 (0.007)	0.160 (0.008)
m-XMT (True)	0.182 (0.004)	0.167 (0.005)	0.148 (0.006)	0.129 (0.007)
m-XMT (Main Effects) with $H_p^{(1)}(t)$	0.238 (0.005)	0.210 (0.006)	0.172 (0.007)	0.135 (0.008)
`RFRE.PO` with $H_p^{(1)}(t)$	0.207 (0.004)	0.183 (0.005)	0.152 (0.006)	0.118 (0.006)
m-XMT (True) with $H_p^{(1)}(t)$	0.182 (0.004)	0.166 (0.005)	0.143 (0.006)	0.115 (0.006)
m-XMT (Main Effects) with $H_f^{(1)}(t)$	0.234 (0.006)	0.195 (0.008)	0.163 (0.011)	0.134 (0.009)
`RFRE.PO` with $H_f^{(1)}(t)$	0.204 (0.004)	0.180 (0.005)	0.150 (0.006)	0.118 (0.006)
m-XMT (True) with $H_f^{(1)}(t)$	0.183 (0.004)	0.166 (0.005)	0.141 (0.006)	0.116 (0.007)
		Moderate Censoring		
m-XMT (Main Effects)	0.236 (0.004)	0.218 (0.005)	0.193 (0.007)	0.163 (0.008)
`RFRE.PO`	0.222 (0.005)	0.207 (0.006)	0.186 (0.007)	0.162 (0.008)
m-XMT (True)	0.185 (0.004)	0.170 (0.005)	0.151 (0.006)	0.131 (0.007)
m-XMT (Main Effects) with $H_p^{(1)}(t)$	0.242 (0.005)	0.214 (0.006)	0.172 (0.007)	0.139 (0.008)
`RFRE.PO` with $H_p^{(1)}(t)$	0.212 (0.004)	0.188 (0.005)	0.156 (0.006)	0.122 (0.006)
m-XMT (True) with $H_p^{(1)}(t)$	0.186 (0.004)	0.169 (0.005)	0.146 (0.006)	0.119 (0.006)
m-XMT (Main Effects) with $H_f^{(1)}(t)$	0.237 (0.006)	0.200 (0.008)	0.169 (0.012)	0.139 (0.009)
`RFRE.PO` with $H_f^{(1)}(t)$	0.209 (0.004)	0.184 (0.005)	0.154 (0.006)	0.122 (0.006)
m-XMT (True) with $H_f^{(1)}(t)$	0.185 (0.004)	0.168 (0.005)	0.144 (0.006)	0.119 (0.007)
		Heavy Censoring		
m-XMT (Main Effects)	0.240 (0.004)	0.221 (0.005)	0.197 (0.008)	0.167 (0.009)
`RFRE.PO`	0.228 (0.005)	0.212 (0.006)	0.191 (0.008)	0.166 (0.009)
m-XMT (True)	0.190 (0.004)	0.173 (0.005)	0.155 (0.006)	0.135 (0.007)
m-XMT (Main Effects) with $H_p^{(1)}(t)$	0.246 (0.005)	0.219 (0.006)	0.182 (0.008)	0.145 (0.009)
`RFRE.PO` with $H_p^{(1)}(t)$	0.218 (0.004)	0.194 (0.005)	0.162 (0.006)	0.127 (0.007)
m-XMT (True) with $H_p^{(1)}(t)$	0.190 (0.004)	0.174 (0.005)	0.151 (0.006)	0.123 (0.007)
m-XMT (Main Effects) with $H_f^{(1)}(t)$	0.240 (0.006)	0.205 (0.008)	0.173 (0.011)	0.145 (0.01)
`RFRE.PO` with $H_f^{(1)}(t)$	0.215 (0.004)	0.189 (0.006)	0.160 (0.007)	0.128 (0.007)
m-XMT (True) with $H_f^{(1)}(t)$	0.190 (0.004)	0.172 (0.006)	0.148 (0.007)	0.124 (0.008)

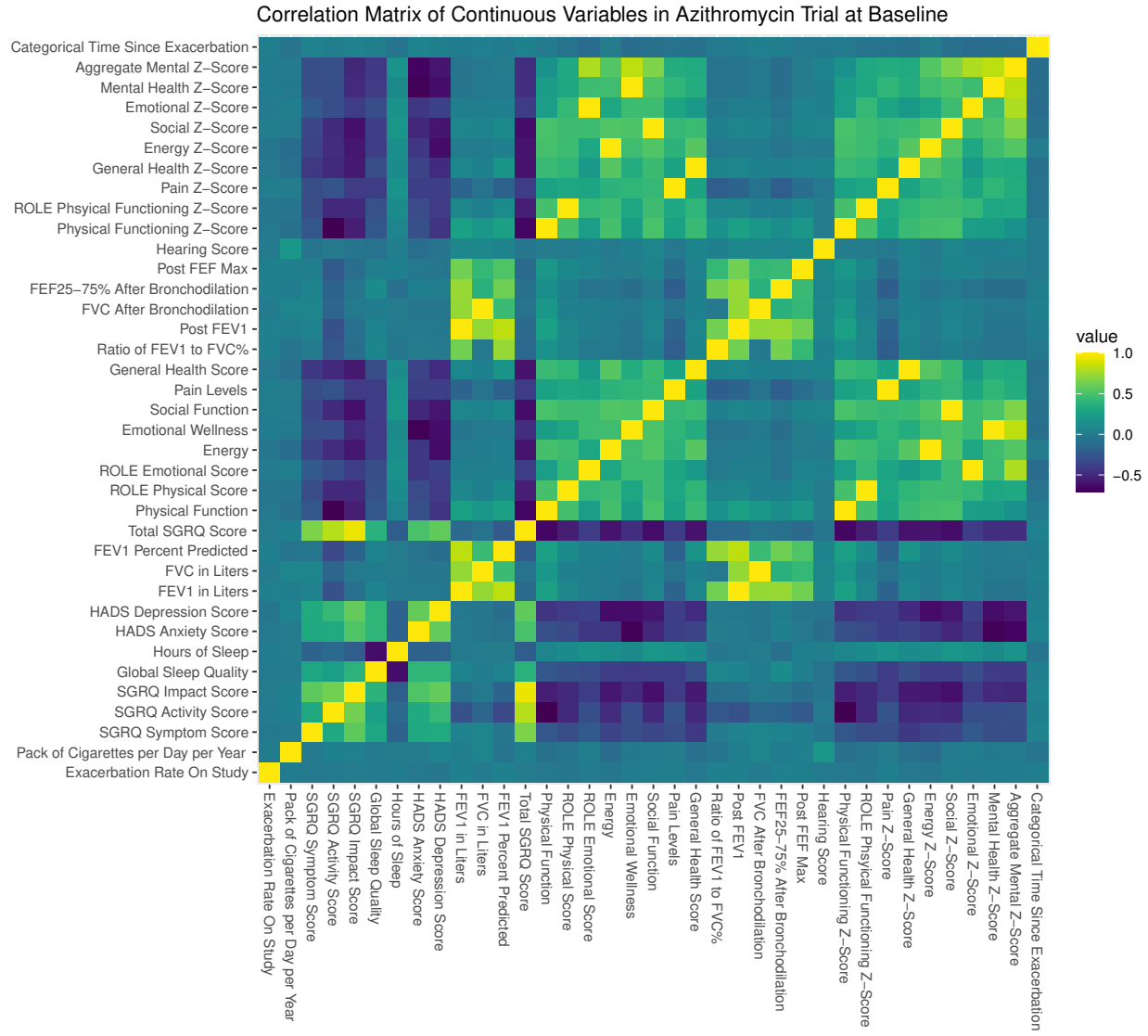


Figure 2.5: A correlation heat map of continuous and ordinal variables in the Azithromycin for the Prevention of COPD cohort, measured at baseline. Variables derived from each other have instances of particularly high positive or negative correlation. These correlation patterns make random forest a more attractive modeling alternative compared to semi-parametric and parametric methods.

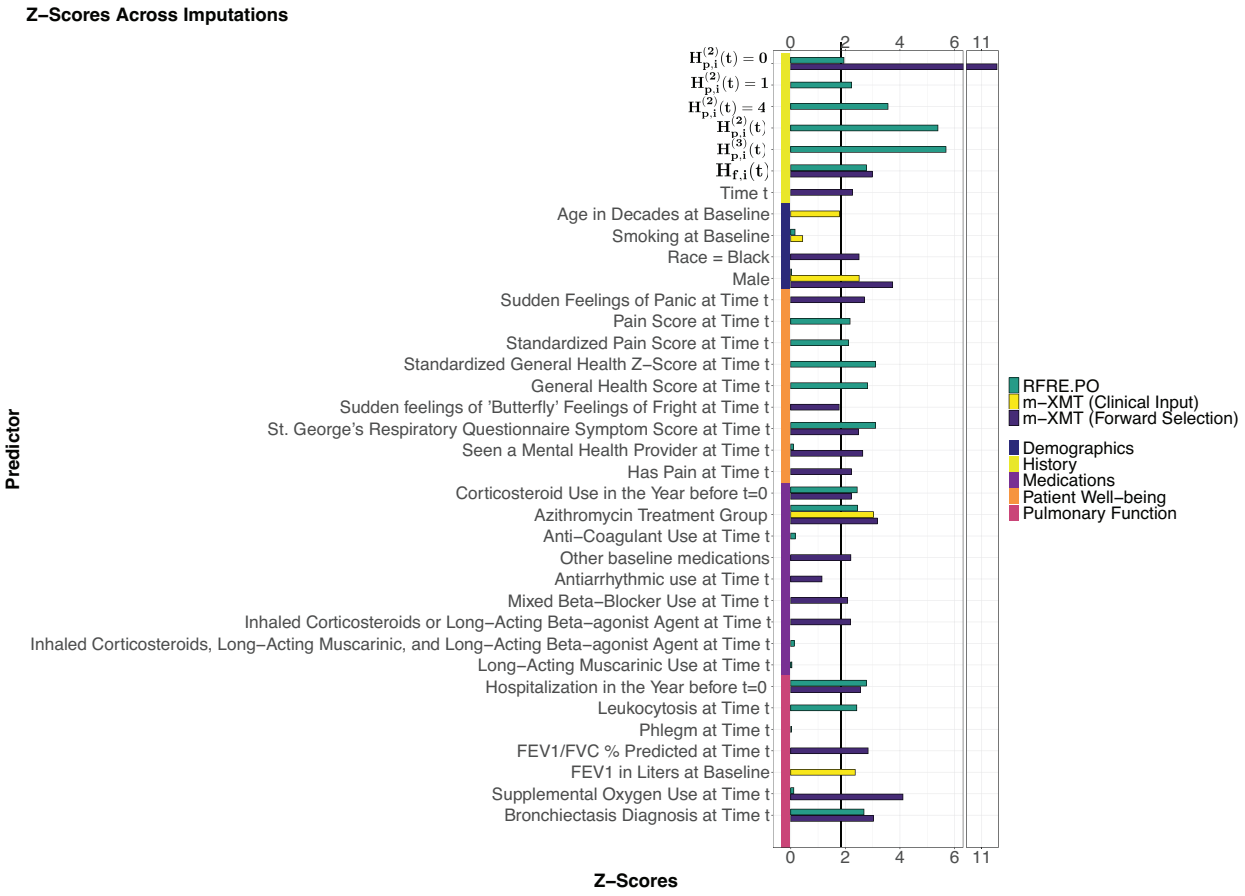


Figure 2.6: Wald Statistic Z-scores derived from the validation set for predictors that were included in the m-XMT models (Clinical Input model from Xia, Murray, Tayob 2020) and the m-XMT selected based on forward selection of the Wald statistic, after adjusting for multiple imputation using Rubin's Rule, where appropriate, and permutation test Z-scores derived from out-of-bag samples as part of the `RFRE.PO` algorithm, adjusting for multiple imputation using Rubin's Rule where appropriate. For the `RFRE.PO` method, 100 permutations were used in calculating the permutation Z-scores. Note that some Wald forward selection variables lost significance as additional terms were added. Colors along the y-axis characterize the type of predictor displayed. Predictors are grouped into one of five categories, displayed in the legend on the right. For the analysis, we created three history variables, $H_f(t), H_p^{(2)}(t)$ and $H_p^{(3)}(t)$. $H_f(t)$ was an indicator if the subject had experienced a severe exacerbation on study time prior to time t. $H_p^{(2)}(t)$ was assumed to take a categorical form based on whether the participant had an exacerbation in the past zero to 31 days ($H_p^{(2)}(t) = 4$), 31-92 days ($H_p^{(2)}(t) = 3$), 93-182 days ($H_p^{(2)}(t) = 2$), 183 days to 365 days ($H_p^{(2)}(t) = 1$), or finally, if that time was more than 365 days or never ($H_p^{(2)}(t) = 0$). $H_p^{(3)}(t)$ is the estimated 30-day event rate using all follow-up prior to time t. Both of these partial history variables are updated at later follow-up windows starting at $t > 0$ based on observed exacerbation data for each individual.

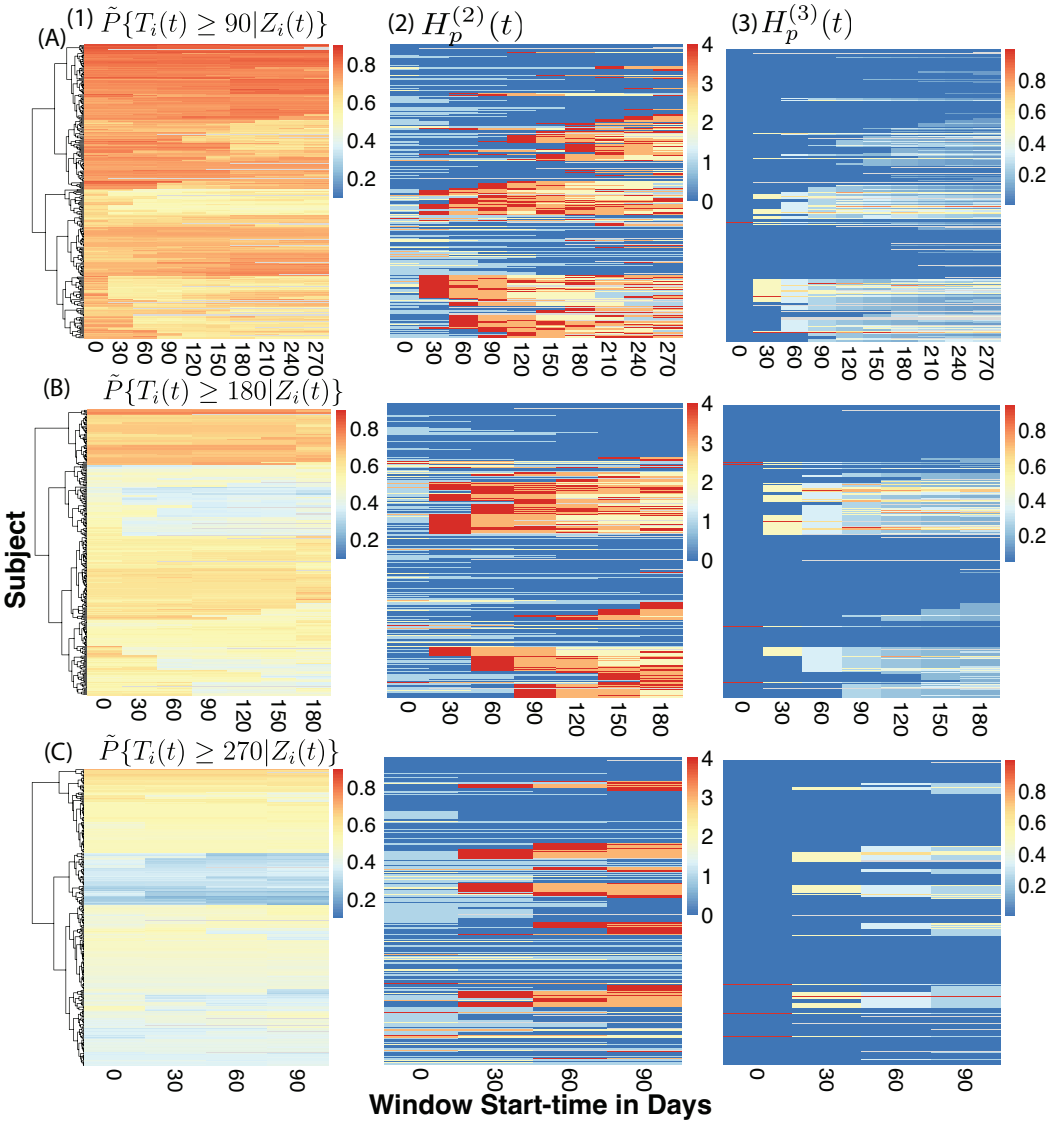


Figure 2.7: `RFRE.PO` predictions and history covariates from the Azithromycin for the Prevention of COPD validation cohort via a 3×3 panel heatmap display. Within each row of heatmaps, subjects are ordered vertically, and clustered according to the default settings of the `pheatmap` command in R applied to the fitted dynamic prediction values for that row of heatmaps. Heatmap rows marked (A), (B) and (C) vary values of follow-up window duration, τ, assuming $\tau = 90$, $\tau = 180$ and $\tau = 270$, respectively. Heatmap columns marked (1), (2) and (3) display: (1) dynamic predictions, $\widetilde{P}\{T_i(t) \geq \tau|Z_i(t)\}$, shown for follow up windows starting at t, with t spaced every 30 days until $t+\tau$ exceeds 360, (2) corresponding categorical $H_p^{(2)}(t)$ and (3) corresponding continuous $H_p^{(3)}(t)$. Both partial history variables are described in detail in Section 2.2, and the specification of $\mathcal{T}_i^*(t)$ and τ^* are described in Section 2.8. Although individuals are not clustered identically in the heatmaps shown in rows (A), (B) and (C), the heatmap color schemes (as expected) reflect higher probabilities of remaining event-free over shorter follow up windows, i.e., $\widetilde{P}\{T_i(t) \geq 90|Z_i(t)\} > \widetilde{P}\{T_i(t) \geq 180|Z_i(t)\} > \widetilde{P}\{T_i(t) \geq 270|Z_i(t)\}$. The display of the history variables seems to reflect in the fitted values from the `RFRE.PO` method, as patterns in column (1) are roughly represented in columns (2) and (3).

CHAPTER 3

A Random Forest Inverse Probability Weighted Pseudo-Observation Framework for Alternating Recurrent Events*

3.1 Introduction

Recurrent events are common across multiple disciplines of research, such as economic, medical, behavioral, political, and biological. For example, in the biomedical context, patients with chronic illnesses often experience periods of relative health, punctuated by episodic instances of illness. Specifically, those with chronic obstructive pulmonary disease (COPD) experience breathing emergencies requiring hospital stays of varying duration. Many statistical challenges for this data structure have been identified and addressed, including dependent censoring when using a gap-time analysis, and addressing correlation between recurrent event times [e.g., Lin et al., 2000, Li et al., 2010, 2017, Xia and Murray, 2019, Xia et al., 2020, Wang et al., 2020, Zhao et al., 2020, Castro-Pearson et al., 2022, Loe et al., 2025a]. In the COPD example, if times-to-release from hospitalization are long, they can be viewed as alternating times-to-event that occur after each hospitalization. Alternating recurrent events of short duration do not have much impact on analyses of the primary recurrent event of interest (such as time-to-hospitalization), and have been largely ignored in the literature. However, longer alternating event durations can have profound impacts on the analysis of the time to the primary recurrent event of interest [see, for detailed explanation Lee and Cook, 2019b], which is the focus of our current manuscript.

More formally, we define alternating recurrent events as a time-to-event of primary interest followed by a time-to-event indicating when the individual returns to the at-risk state for the primary event of interest. For example, in the CareQOL study [Carlozzi et al., 2025],

[0]Under revision at *Biometrics* as: Abigail Loe, Susan Murray and Zhenke Wu. "A Random Forest Inverse Probability Weighted Pseudo-observation Framework for Alternating Recurrent Events." A pre-print may be found here.

caregivers of patients with traumatic brain injuries (TBIs) experienced repeated periods of depression before returning to baseline. To truly understand the primary recurrent event of interest, i.e. time-to-depressed state, one needs to account for the time-to recovery from depression, i.e. the secondary (alternating) recurrent event type. A key concept that motivates our work is that subjects cannot be at-risk for entering a depressed state while currently in one. This missingness pattern related to the primary recurrent event time results in a biased analysis sample, similar to that seen in Rao [1965], Patil and Rao [1978], and Vardi [1982]. That is, since depression-prone patients are more likely to be removed from the risk set for the primary endpoint at regular intervals they are not adequately represented in the analysis.

Alternating recurrent events have been the subject of much recent study. There are intensity based models [Hu et al., 2011], which extend the well-known Andersen and Gill [1982] model, frailty models, where dependence between events within a subject is linked through an individual random effect [Li et al., 2010, Mazroui et al., 2013a, Wang et al., 2020, Li et al., 2017], or copula-based multi-state models [Lee and Cook, 2019a]. Each of these models focuses on modeling the hazard in a semiparametric manner, or relying on likelihood based methods for parameter estimation. To our knowledge, there are no methods that model restricted mean event times in prespecified follow-up windows, with the use of a random forest estimated inverse probability weight to correct for bias due to differential at-risk status. Such a tool would increase both interpretability of model results and robustness of bias correction due to the nonparametric nature of random forest algorithms.

This manuscript develops an alternating recurrent event analysis approach where accounting for alternating (secondary) event times is crucial in avoiding bias. Our approach first transforms the recurrent event data structure to a censored longitudinal dataset comparable to that of Tayob and Murray [2015] and second uses random forest algorithm-derived inverse probability weights (IPW) to correct for sampling bias due to the secondary alternating recurrent events. As to first step in our approach (the censored longitudinal dataset and corresponding analyses), several advantages have been proposed in the literature including: (1) These approaches allow for a better understanding of short-term clinical outcomes where the user may define various short-term horizons. (2) Efficiency gains in restricted mean estimation have been noted when borrowing information across follow-up windows, and in the recurrent-event setting, [Xia and Murray, 2019]. (3) By having prespecified starting times to the follow-up windows where times-to-first-event are recorded, the data structure avoids dependent censoring that can arise with dependent event times in gap-time analyses [Xia et al., 2020, Zhao et al., 2020, Loe et al., 2025a]. (4) When used with pseudo-observations [Andersen et al., 2003, 2004] to address the additional independent censoring mechanism in these settings, data from this structure may be analyzed using traditional longitudinal

data methods for which there is an abundance of software familiar to readers learning our approach. One such modeling choice is the use of random forest IPW weights to account for potentially informative missingness due to the secondary alternating recurrent event. Our random forest IPW weights allow for complex relationships between many potentially collinear predictors and at-risk status for the primary recurrent event of interest.

The rest of this paper is organized as follows. In Section 3.2, we transform alternating recurrent event data into a censored longitudinal data format, based on which we construct pseudo-observations (PO) in Section 3.3. Although the PO construction unlocks access to popular regression methods, it is subject to selection into the at-risk sets, which we correct using IPW in Section 3.4. In Section 3.5, we describe the random forest approach for obtaining probability weights and useful history covariates that substantially improve the performance of our method. Performance of the proposed method is evaluated in Section 3.6. The model is applied to mobile health data from the CareQOL study in Section 3.7, and we conclude with a discussion in Section 3.8.

3.2 Censored Longitudinal Data Structure

For individual i, alternating recurrent event times $T_{i,0}^{(1)} < T_{i,0}^{(2)} < T_{i,1}^{(1)} < T_{i,1}^{(2)} < \dots$ are potentially censored by the random variable C_i, where C_i is independent of $T_{i,j}^{(1)}, T_{i,j}^{(2)}$, $i = 1, ..., n$, and $j \geq 0$. Here, j indexes the event pairings, i indexes subjects, and the superscript k, $k \in \{1, 2\}$ indexes the alternating primary ($k = 1$) events and secondary ($k = 2$) events. Upon experiencing an event at $T_{i,j}^{(1)}$, individual i is not in the set of individuals at-risk for the primary recurrent event time for $t \in [T_{i,j}^{(1)}, T_{i,j}^{(2)})$ until the secondary alternating recurrent event occurs. That is, $T_{i,j}^{(2)}, j \geq 0$ are times at which individual i re-enters the population of individuals at-risk for the primary alternating recurrent event type. For instance, in the setting of TBI caregivers, $T_{i,j}^{(1)}, j \geq 0$ would be times when caregivers enter a depressive state, and $T_{i,j}^{(2)}, j \geq 0$ are the times when subjects leave the depressive state to be at-risk for re-entering a depressive state.

Due to potential censoring, observed data for each individual is a set of event times and indicator functions. In this manuscript, we denote potentially censored event times using $X_{i,j}^{(k)} = \min(T_{i,j}^{(k)}, C_i)$ and the corresponding censoring indicator $\delta_{i,j}^{(k)} = I(X_{i,j}^{(k)} < C_i)$. To avoid the well-known dependent censoring problem due to gap-time analysis, we follow the censored longitudinal data approach of Tayob and Murray [2015], Xia et al. [2020], Zhao et al. [2020], and Loe et al. [2025a]. They transform correlated recurrent event data into a series of times-to-first-event in follow-up windows of length τ, starting at $t \in \mathcal{T}_i = \{t_0, t_1, ...t_{b_i}\}$. When data are uncensored, t_{b_i} is chosen so that $t_{b_i} + \tau$ does not exceed available follow-up

time. When observation i is censored, $t_{b_i} = \max\{t_l \in \mathcal{T}_i : t_l < C_i\}$. For the follow-up window starting at time t, $n(t)$ is the number of uncensored subjects available for follow-up at t. See Figure 3.1, Panel (A) for a visualization of traditionally recorded alternating recurrent events with corresponding data in Panel (B).

To create the censored longitudinal dataset, for each $t \in \mathcal{T}_i$, and each individual $i = 1, 2, ..., n(t)$ who remains uncensored at t, it is helpful to map the relationship between the original alternating recurrent events denoted by $T_{i,j}^{(1)} < T_{i,j}^{(2)}$ for $j \geq 0$ with those that appear first in the follow-up window starting at t. The algorithm then becomes, for each individual $i = 1, 2, \ldots n(t)$ uncensored at time t:

1) Identify the j index of the first alternating recurrent event of type $k = 1, 2$ after t by $\eta_i^{(k)}(t) = \min\{j | T_{i,j}^{(k)} > t, j = 0, 1, \ldots\}$.

2) Define residual censoring time from time t (the start of the follow-up-window) as $C_i(t) = C_i - t$. Let $T_i^{(k)}(t) = T_{i,\eta_i^{(k)}(t)}^{(k)} - t$. Let $X_i(t) = \min\{T_i^{(1)}(t), T_i^{(2)}(t), C_i(t)\}$ be the first observed event time measured from time t regardless of its event type, with corresponding censoring indicator, $\delta_i(t) = I\{X_i(t) < C_i(t)\}$.

3) Define $R_i(t) = I[T_i^{(1)}(t) < T_i^{(2)}(t)]$, to indicate that individual i was at-risk for the primary alternating recurrent event as of time t.

4) Construct the censored longitudinal dataset at t by $\Big\{X_i(t), R_i(t), \delta_i(t), Z_i(t), i = 1, 2, \ldots n(t)\Big\}$ where $Z_i(t)$ are covariates that are known as of time t. The conversion of alternating recurrent events into a censored longitudinal framework is shown in Figure 3.1, Panel (C). Data may then be analyzed using appropriate longitudinal and survival analysis methods.

The censored longitudinal dataset described above is characterized by choice of window spacing, $a = t_k - t_{k-1}$, and the window of clinical interest, τ. In the simple recurrent events setting, setting a to $\frac{1}{3}$ the recurrent event rate captures roughly 90% of event times [Xia and Murray, 2019]. However, a may be selected to be as small as one unit of time apart, a choice that increases computational burden, with diminishing returns in terms of event capture. We recommend selecting τ based on clinical expertise, in order for τ to reflect the period during which covariate measurements have an effect on the outcome. For example, blood oxygen levels on day t may be a strong predictor of health status over the subsequent week (τ), but is likely less informative for predicting health status over the next twenty years.

In this manuscript, we define $T_i(t)$ as the time to the primary recurrent event of interest if patients were at-risk for this event at all start-times, $t \in \mathcal{T}_i$. In the setting with alternating recurrent events, we only partially observe $T_i(t) = R_i(t) \times T_i^{(1)}(t)$ when $R_i(t) = 1$. When

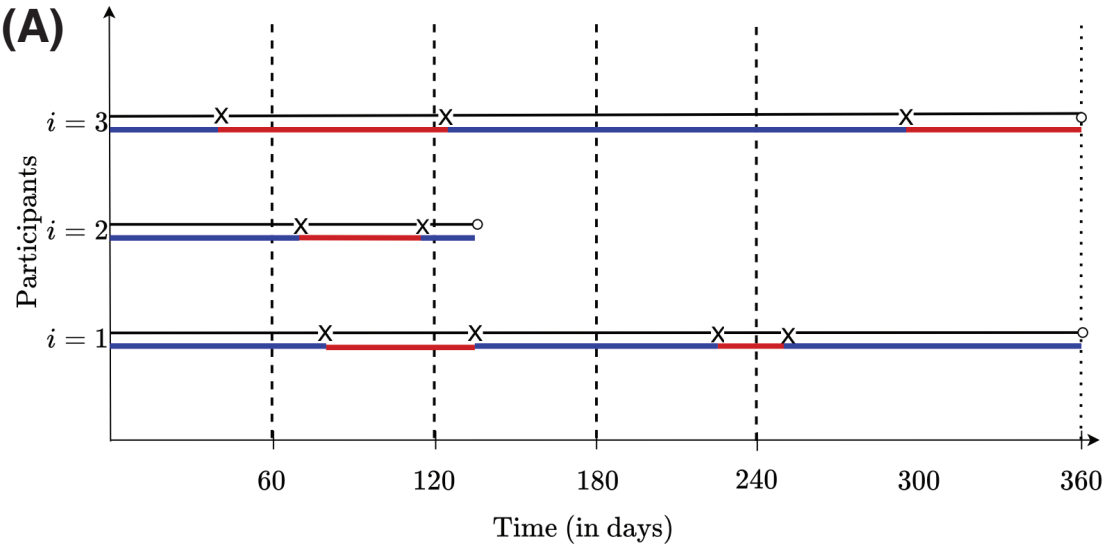
$R_i(t) = 0$, individual i is not at-risk at time t for this event, so that $T_i(t)$ is missing. For further discussion on the clinical relevance of $T_i(t)$ and consideration of $T_i(t)$ rather than $T_i^{(1)}(t)$, see Appendix B.1. We will hereafter refer to $\mathcal{CC}(t) = \{i = 1, 2, \ldots, n_R(t) : R_i(t) = 1\}$ as the set of uncensored individuals at time t with complete data on the primary recurrent event endpoint, with $n_R(t) \leq n(t)$ being the number of individuals in the set $\mathcal{CC}(t)$. We define $\{X_i(t), \delta_i(t) : i \in \mathcal{CC}(t)\}$ as the complete case data for the primary recurrent event of interest at t. In Section 3.4, we express our interest in understanding $T_i(t)$ for $i = 1, 2, \ldots, n(t)$, rather than $T_i^{(1)}(t)$ for $i \in \mathcal{CC}(t)$, and define models with this goal in mind. In particular, we introduce longitudinal restricted mean regression models to estimate $E[\min\{T_i(t), \tau\}|Z_i(t)]$, where τ is the follow-up window length from t. The τ-restricted mean time-to-event uses τ to focus inference on clinically relevant horizons, e.g., 30-day readmission in hospital outcomes or 5-year survival in oncology, matching clinical or policy relevance [e.g., Royston and Parmar, 2013]. But before details for model development are given, we summarize a pseudo-observation approach to address the censoring by $C_i(t)$ in Section 3.3.

3.3 Construction of Pseudo-Observations

Andersen et al. [2003] proposed a technical device now employed in many longitudinal survival contexts called "pseudo-observations," a variation of jack-knife methods that provide asymptotically unbiased estimates of regression parameters in the presence of censoring. We provide a primer on the use of pseudo-observations, calculated separately for a particular follow-up window at time t, based on the $n_R(t)$ individuals in $\mathcal{CC}(t)$. The issue of selection into the risk set at t caused by analyzing only individuals in $\mathcal{CC}(t)$ will be corrected in Section 3.4. For each $t \in \mathcal{T}$, we define $S_t^{(1)}(u) = P\{T_i^{(1)}(t) > u\}$ as the primary-event-free survival function based on the $n_R(t)$ individuals in $\mathcal{CC}(t)$, and $\theta(t) = E[\min\{T_i^{(1)}(t), \tau\}|R_i(t) = 1] = \int_0^\tau S_t^{(1)}(u)du$, i.e., the complete case τ-restricted mean event time at follow-up window t. A consistent nonparametric estimator for $\theta(t)$ is $\widehat{\theta}(t) = \int_0^\tau \widehat{S}_t^{(1)}(u)du$, where $\widehat{S}_t^{(1)}(u)$ is the Kaplan-Meier survival curve estimated from data pairs $\{X_i(t), \delta_i(t) : i \in \mathcal{CC}(t)\}$. Define leave-one-out estimator $\widehat{\theta}^{-i}(t)$ as the estimator where individual i is excluded from the cohort. Then a pseudo-observation for subject $i, i \in \mathcal{CC}(t)$ at time $t \in \mathcal{T}_i$ is:

$$PO_i^\tau(t) = n_R(t)\widehat{\theta}_i(t) - \{n_R(t) - 1\}\widehat{\theta}^{-i}(t),$$

which can be easily obtained using the `pseudomean` function from the `survival` package in the programming language `R`. Andersen et al. [2004] showed that pseudo-observations



(B)

i	j	k	$X_{i,j}^{(k)}$	$\delta_{i,j}$
1	0	1	80	1
1	1	2	127	1
1	2	1	221	1
1	3	2	249	1
1	4	1	360	0
2	0	1	75	1
2	1	2	118	1
2	2	1	130	0
3	0	1	48	1
3	1	2	122	1
3	2	1	290	1
3	3	2	360	0

(C)

i	t	$X_i(t)$	$R_i(t)$	$\delta_i(t)$
1	0	80	1	1
1	60	20	1	1
1	120	7	0	1
1	180	41	1	1
1	240	9	0	1
2	0	75	1	1
2	60	15	1	1
2	120	10	1	0
3	0	48	1	1
3	60	62	0	1
3	120	2	0	1
3	180	110	1	1
3	240	50	1	1

Figure 3.1: A visualization of alternating recurrent events in both the longitudinal sense and the alternating censored longitudinal format. In Panel (A), events within a subject are marked with an "X." The primary and secondary event types are the start and the end of the red segments, respectively. Blue segments represent at-risk periods for the primary event type. Start times of potential follow-up windows of length $\tau = 120$ corresponding to $\mathcal{T} = \{0, 60, 180, 240 \text{ days}\}$ are represented with dashed vertical lines. Traditional observed alternating recurrent events are recorded in Panel (B), while Panel (C) corresponds to those same event converted to an alternating censored longitudinal format. All participants were administratively censored after 360 days of follow-up, thus $t = 300$ is not included in $\mathcal{T}$. Note that $X_{1,0}^{(1)}$ (the first 'X' in subject 1) produces multiple measures in the alternating censored longitudinal data format. Additionally, subject 2 produces three measures where $R_2(t) = 1$; subject 3 produces the same number of measures where $R_3(t) = 1$, but was removed from the at-risk population at $t = 60, 120$.

such as $PO_i^\tau(t)$, have the same conditional expectation, $E[PO_i^\tau(t)|R_i(t) = 1, Z_i(t)]$, as $E[\min\{T_i^{(1)}(t), \tau\}|R_i(t) = 1, Z_i(t)]$.

Essentially, once calculated, a pseudo-observation can be used in a regression analysis as an uncensored outcome as if one were actually modeling $E[\min\{T_i^{(1)}(t), \tau\}|R_i(t) = 1, Z_i(t)]$. When calculating $PO_i^\tau(t)$ for $t \in \mathcal{T}$ and $i = 1, 2, \ldots, n_R(t)$, one obtains a longitudinal dataset that may be used with traditional longitudinal analysis methods, with all the benefits of large sample behavior and asymptotic guarantees attributed to those methods [Robins et al., 1995].

3.4 Inverse Probability Weighted Pseudo-Observations

In Section 3.3, we constructed pseudo-observations, $PO_i^\tau(t)$, that have conditional expectation, $E[\min\{T_i^{(1)}(t), \tau\}|R_i(t) = 1, Z_i(t)]$, so that an uncensored outcome is available for analysis, but it is still subject to selection into the risk set at time t, i.e., this does not necessarily correspond to $E[\min\{T_i(t), \tau\} \mid Z_i(t)]$. Now using pseudo-observations, $PO_i^\tau(t)$, we show how inverse probability weights (IPW) may be used to correct this issue to construct models for estimating $E[\min\{T_i(t), \tau\}|Z_i(t)]$. We need two assumptions for establishing the correction in Theorem 1.

Assumption 2 (Ignorability) *There exist covariates $W_i(t)$ such that $R_i(t) \perp\!\!\!\perp T_i(t) \mid W_i(t)$, where $R_i(t)$ tracks at-risk status for the primary alternating recurrent event, $T_i(t)$ is the residual primary alternating recurrent event free survival time, and $W_i(t)$ is a set of covariates that contains $Z_i(t)$, the regressors of interest.*

Assumption 3 (Sufficient regressors) *The restricted mean time-to-primary-event in the at-risk-population only depends on the covariates $Z_i(t)$:*

$$\mathbb{E}\left[\min\{T_i^{(1)}(t), \tau\}|Z_i(t), W_i(t), R_i(t) = 1\right] = \mathbb{E}\left[\min\{T_i^{(1)}(t), \tau\}|Z_i(t), R_i(t) = 1\right].$$

Ignorability (Assumption 2) is particularly important, but also unverifiable. That is, one can identify covariates associated with both $R_i(t)$ an $T_i(t)$ and incorporate these into IPW estimates. But it is impossible to know if there are additional unmeasured covariates that induce dependence between $R_i(t)$ and $T_i(t)$ if unaccounted for in IPW estimates. An advantage of using a random forest approach for IPW estimation (described in Section 3.5) is the increased plausibility of the ignorability assumption. That is, the functional form and interactions between covariates navigated by random forest estimation techniques may be beyond a typical analyst's capabilities to capture in IPW estimation. Furthermore, random forest algorithms are able to incorporate information from colinear predictors in a manner

that is limited by standard regression model approaches. Section 3.5 highlights history (and reflective history) covariates that are useful for practitioners to include in model and IPW estimation.

Assumption 3 indicates that covariates included in $Z_i(t)$ contain all needed information from $W_i(t)$ that might aid in modeling $\min\{T_i^{(1)}(t), \tau\}$. If $Z_i(t)$ includes all covariates in $W_i(t)$, then Assumption 3 automatically holds. However, one need not include covariates from $W_i(t)$ that have no association with $\min\{T_i^{(1)}(t), \tau\}$ for Assumption 3 to hold, and over-populating $Z_i(t)$ with unnecessary predictors may decrease efficiency of parameter estimates substantially.

Theorem 4 (IPW corrects the selection into the at-risk set) *Under Assumptions 2, 3, and* $0 < \pi_i(t) < 1$*, the following holds:*

$$\mathbb{E}\left\{ \frac{R_i(t)}{\pi_i(t)} PO_i^{\tau}(t) \middle| Z_i(t) \right\} = \mathbb{E}\Big[\min(T_i(t), \tau) | Z_i(t) \Big]. \tag{3.1}$$

See Section B.2 in the Appendix for a detailed proof. Hence, for each individual, if we can consistently estimate $\dfrac{R_i(t) PO_i^{\tau}(t)}{P\{R_i(t) = 1 | W_i(t)\}}$, this term will have the same limiting expectation as the random variable $\min\{T_i(t), \tau\}$ given $Z_i(t)$, $i = 1, 2, \ldots n(t)$. In Section 3.5, we describe a random forest estimate, $\widehat{\pi}(t)$ for $P\{R_i(t) = 1 | W_i(t)\}$ and conditions where these estimates are consistent. Under a linear regression model specification of the target $\mathbb{E}[\min\{T_i(t), \tau\} \mid Z_i(t)] = \beta^{\top} Z_i(t)$, the estimates $\widehat{\beta}$ may then be taken as the solution to the following PO-based generalized estimating equation:

$$\mathbf{0} = \sum_{i=1}^{n} \bar{\mathbf{Z}}_i \mathbf{W}_i \mathbf{\Lambda}_i^{-1} (\widehat{\mathbf{PO}}_i^{\tau} - \bar{\mathbf{Z}}_i^{\top} \beta), \tag{3.2}$$

where Λ_i is the usual matrix product of variance functions and working correlation structure, and $\mathbf{W}_i$ is a diagonal matrix with entries $\frac{R_i(t)}{P\{R_i(t)=1|W_i(t)\}}$. Large sample results follow similarly to Robins et al. [1995]. In this paper we use an unstructured variance covariance matrix, with sandwich standard errors to account for unmodeled correlation, though depending on the form of the censored longitudinal data structure selected, an analyst may select other variance-covariance structures, such as the independence or Toeplitz working correlation models [e.g., Xia et al., 2020]. Hereafter, we refer to this model as the pseudo-observation alternating recurrent event IPW and random forest GEE model, or PAIR-GEE for short; non-identity link functions may be used as appropriate depending on the intended target regression form.

3.5 Estimation of Weights via Random Forest

We observe pairs $\{R_i(t), W_i(t)\}$, where $R_i(t)$ is at-risk indicator for the primary event type and $W_i(t)$ is a set of p-many potentially time-varying and correlated covariates measured at time t, as well as t itself. Here we describe how to obtain estimates, $\widehat{\pi}_i(t)$, for $P\{R_i(t) = 1|W_i(t)\}$ using a random forest algorithm. These algorithms are pervasive in machine learning literature, with opportunities for custom implementation in the programming language `R`, although very few algorithms have been described in a way that is applicable for the longitudinal $R_i(t)$, $t \in \mathcal{T}_i$. Our general approach follows that of Adler et al. [2011] with algorithmic details specific to constructing $\widehat{\pi}_i(t)$ given below. As such, we benefit from similar advantages of being less parametric than competitor GEE regression models for dependent binary responses, while capturing complex, nonlinear associations between predictors and outcomes. In the below, we lean on random forest jargon, which we explain in the context of our problem as we go through the algorithm.

The predictors $W_i(t)$ are allowed to be highly collinear, and should contain $Z_i(t)$ as well as a large number of event history variables observed prior to t that are some of the most helpful predictors when event times are correlated [Loe et al., 2025a]. Some examples of event history variables known at t include time since last event, average times-to-event for primary and secondary event types in earlier follow-up windows, and numbers of these event types prior to t, as well as t itself. If event history data is not available at $t = 0$, the random forest algorithm is able to proceed, broadening the predictor pool as t moves to later follow-up windows. We will also explore the use of reflective history predictors similar to those described in missing data literature for longitudinal analyses with non-monotone dropout [Sun and Tchetgen Tchetgen, 2018], where these predictors use either future or past information relevant to t, but not both, depending on which temporal direction has more longitudinal measurements available. Such longitudinal missingness patterns are fully observed when analyzing retrospective data; thus covariates that may perfectly predict $R_i(t)$ after t, are available for analysis. For further discussion, see Supplemental Materials Section B.3. Once these predictors are constructed, the random forest algorithm starts by repeating Steps 1 and 2 B-many times:

Step 1) Two-stage bootstrap sample: Sample with replacement from the set of unique subjects, $i = 1, 2, \ldots n$. For each sampled subject (indexed by ℓ) in the bootstrap dataset, randomly select one follow-up window $t_{j_\ell} \in \mathcal{T}_\ell$. Label the corresponding selected data pairs, $\{R_\ell(t_{j_\ell}), W_\ell(t_{j_\ell})\}$. Subjects in the bootstrap sample are called "bagged, " with the remaining subjects referred to as "out-of-bag."

Step 2) Grow a tree:

a.) From the p possible covariates, sample $m < p$ covariates. m is oftentimes a user-selected value, but if no value is specified, many software programs default to $\sqrt{p}$ or $\frac{p}{3}$. In this paper, while the random forest implementation is custom, we use the default choices from the `ranger` package, which recommends $m = \text{floor}(\sqrt{p})$, a choice that works well in simulation and data analysis. For each $s \in \{1, \dots m\}$, consider a grid of all possible dichotomizations (or indicator functions) of selected covariates, denoted $\{\boldsymbol{D}_1, \boldsymbol{D}_2, ...\boldsymbol{D}_m\}$, where $\boldsymbol{D}_s$ is a grid of all observed values of selected covariate D_s.

b.) Consider the values of the objective function created by splitting the data into two groups based on each potential dichotomization. In our case, the outcome of interest is a Bernoulli random variable; we consider the objective function known as the Gini cost function (for further details see Chapter 4 of Breiman [2001]). Bin the data into two "daughter" nodes that reflect the split minimizing the objective function.

c.) Recursively make splits in the daughter nodes until one of three hyperparameters is met: (1) a minimum number of observations is found in a daughter node, (2) the maximum number of splits as specified by the analyst has been made, or (3) no further reduction in the objective function can be made with more splits. For the purposes of this paper, we select minimum number of nodes to be 20 in simulation, and 8 in data application, which balances a reasonable size for terminal nodes, while ensuring enough splits to approximate complicated relationships.

Predictions from tree b for individual i' in the associated out-of-bag sample are made by following the split decisions that correspond to covariate pattern $W_{i'}$, and averaging the at-risk indicators found in the terminal node that i' belongs to. When all B trees are grown, random forest predictions for individual i' averages the predictions arising from trees where i' is out-of-bag, allowing fitted values to remain independent from the tree-growing stage.

There are two main advantages to this forest. First, due to the two-stage bootstrapping scheme, subjects with many measurements are as likely to be included in the sampling step as subjects with few measurements. This feature contributes to model stability in the presence of censoring, as individuals who are censored early have their observed data frequently included. Second, out-of-bag data remain independent of the bagged data, yielding simple fitted values where proofs involving consistency of independently distributed outcomes from Breiman [2001] may still be applied.

Random forests in general have many hyperparameters to tune: number of covariates to sample, m, the minimum number of observations in terminal nodes, the maximum number

of splits allowed, and the loss function evaluated at each split. Hyperparameter selection for the longitudinal random forest described is still an area of open research.

3.6 Simulation Study

The goals of our simulation study are two fold: (1) to compare the performance of PAIR-GEE models to a complete case (unweighted) analysis described in Section 3.3 and (2) to demonstrate the importance of including history covariates in estimating weights via random forest. For goal (1), we report coverage probabilities of 95% confidence intervals, average standard error, empirical standard error, bias and percent bias of estimates. For goal (2), we define seven useful history variables, discussed in Supplemental Materials Section B.3.

We generate correlated piece-wise exponential alternating recurrent event times ($\rho \approx 0.80$) using a Gaussian copula approach similar to that described by Xia et al. [2020] for a binary predictor, $Z = 0, 1$. Independent censoring was generated by $\min\{12, L_i\}$, where $L_i \sim \text{Laplace}(\mu = 12, b = 1.5)$, which corresponds to roughly 30% of observations censored before the final time point. The censored longitudinal data structure takes the form $t \in \{0, 1, 2, \dots 11\}$ months with $\tau = 1$ month. Piece-wise hazards at times $[0, 6), [6, 12]$ are $\{1/2, 1/4\}$ for $Z = 0$ and $\{1/3, 1/6\}$ for $Z = 1$. Our data follow:

$$\begin{aligned} E[\min\{T_i(t), \tau = 1\}] =& 0.714 \times I(Z_i = 1)I(t < 6) + 0.631 \times I(Z_i = 0)I(t < 6) + \\ &+ 0.826 \times I(Z_i = 1)I(t \geq 6) + 0.764 \times I(Z_i = 0)I(t \geq 6). \end{aligned}$$

so that the 1-month restricted means for $t < 6$ are $\{0.631, 0.714\}$ for $Z = 0$ and $Z = 1$, respectively, and for $t \geq 6$, the 1-month restricted means are $\{0.764, 0.826\}$ for $Z = 0$ and $Z = 1$, respectively. Figure B.1 in Supplemental Materials displays six sample subjects that follow the various hazards of alternating recurrent events in the correlated case. Subjects with $Z_i = 0$ have both more primary events and less availability at window start times $t \in \mathcal{T}$ in the complete case analysis. Figure B.2 in Supplemental Materials shows a representative simulated dataset and the empirical restricted means over time using an unweighted GEE approach that ignores the issue of selection into the risk set. An additional simulation, considering considering independent gap-times and continuous covariates, is included in Appendix Section B.4 with results displayed in Table B.2 and Figure B.3.

The seven history covariates described in Section 3.5 were included in the random forest algorithm for estimating PAIR-GEE weights. The following notation is used to indicate the degree of of historical information available in the results shown in Figure 3.2, Table 3.1 and Supplemental Table B.1. $\boldsymbol{H}_f$ indicates a full year of patient history was available prior to

$t = 0$. $\boldsymbol{H}_{p_1}$ and $\boldsymbol{H}_{p_2}$ indicate that the patient history included information on the last two or one alternating recurrent events, respectively, if these occurred within the prior year. $\boldsymbol{H}_{p*}$ indicates that no historical information prior to $t = 0$ was available and that reflexive history covariates (as described in Section 3.5) were used. $\boldsymbol{H}_{p_3}$ indicates that complete history was available on 50% of individuals, but was otherwise unavailable. $\boldsymbol{H}_{p_4}$ indicates no availability of history covariates prior to $t = 0$. $\boldsymbol{H}_n$ indicates no history covariates were included in the random forest algorithm.

The empirical distribution of coefficient estimates across simulation replicates for the censored setting is displayed in Figure 3.2, with numeric results displayed in Table 3.1. Uncensored results are displayed in Appendix Table B.1 and Figure B.4. As expected, the unweighted GEE analysis gives the poorest performance in terms of bias and coverage. Any use of covariate information in estimating IPW weights reduces bias. However, success varies greatly depending on the amount of historical information. Thus, results emphasize the importance of seeking out as much historical covariate information as possible in estimating these weights.

Whereas methods with high degree of historical covariate information ($\boldsymbol{H}_f$, $\boldsymbol{H}_{p_1}$, $\boldsymbol{H}_{p_2}$ and $\boldsymbol{H}_{p*}$) gave excellent results in terms of bias and 95% confidence interval coverage, as historical information becomes less available ($\boldsymbol{H}_{p_3}$, $\boldsymbol{H}_{p_4}$, $\boldsymbol{H}_n$) performance deteriorates. One exception was in estimating β_4, which was influenced by an abrupt change in hazard after $t \geq 6$. All methods struggled in terms of bias and 95% coverage probability, though the magnitude of under coverage varies drastically (coverage probability of 0.928 for PAIR-GEE with $\boldsymbol{H}_f$ weights to 0.008 for the unweighted GEE method). This phenomenon is further evaluated via Figure B.5, which shows the window-by-window bias for all PAIR-GEE models and the unweighted model; for t near 6, PAIR-GEE models that rely on history covariates have varying, non-zero levels of bias, while somewhat stabilizing near zero by the final window start time, $t = 11$. For the PAIR-GEE model with $\boldsymbol{H}_n$ and the unweighted model, this does not happen, but rather after $t = 0$, bias consistently remains around 0.05. The variation in bias overtime seems to indicate error in weight estimation near the hazard change-point; weights are not able to adapt quickly enough to changes in recurrent event risk profiles.

Table 3.1 also displays the ratio of empirical standard errors to standard errors obtained via the sandwich estimator. To some degree, all models underestimated variation in $\widehat{\beta}_1, \widehat{\beta}_2$, and $\widehat{\beta}_4$, while over estimating variation of $\widehat{\beta}_3$. This is likely due to unaccounted for variance in the estimation of the random forest IPW weights. Variance may be estimated by a bootstrap method concisely summarized in Williams and Savitsky [2021], though this greatly increases computational burden. Table B.3 displays PAIR-GEE results from simulation when weighted parameter estimates are bootstrapped. A significant increase in coverage is seen, due to the

much wider confidence intervals and an overestimation of the variance of $\widehat{\beta}$. As such, we recommend using unstructured variance covariance models with sandwich standard errors over the bootstrap.

A highlight of the simulation results is the strong performance of the reflective covariate $\boldsymbol{H}_{p*}$ that uses either past or present information only on study time relative to time t, but not both, depending on which direction had more measurements available for construction of history covariates. This predictor greatly outperformed the historical covariates of type $\boldsymbol{H}_{p_4}$, that had the same amount of historical information available, but only used information prior to t in estimating the IPW weights.

3.7 Application to the CareQOL Study

The 6-month CareQOL study [Carlozzi et al., 2025] randomized $n = 257$ caregivers of traumatic brain injury patients at baseline to receive, or not receive self-care push notifications from the CareQOL app, which tailored the message content by previous levels of physical activity, and sleep minutes measured by FitBit devices along with previous levels of anxiety, strain, and depression based on ecological momentary assessments (health-related quality of life, HRQOL). Other demographic information (ethnicity, age, race, sex) and the caregiver's patient level of independence were recorded by study personnel. All study participants had weekly self-reported HRQOL measures of depression (t-score). Subjects with a weekly t-score two points above baseline are labeled as in a depressive state. Hence our alternating recurrent events are entering (primary) and exiting (secondary) a depressed state, respectively. The findings from this study suggest that self-care push notifications did not enhance outcomes (somewhat worse main effect), with benefits observed among participants with active engagement with the app [Carlozzi et al., 2025]. Here we take a deeper dive and analyze the weekly data over the 6-month period. Using our proposed method, we are able to analyze transitions in and out of the depressed state to interpret what may have contributed to the non-improvement of the end-of-study HRQOL outcomes.

We apply our proposed method to 12 weeks of follow-up of the CareQOL cohort with $\tau = 2$ weeks and $t \in \mathcal{T} = \{0, 2, 4, 6, 8, 10\}$ weeks. For the $n = 257$ subjects, computation of the pseudo-observations via the package `pseudo` took less than 0.05 seconds, while random forest IPW weights for all $t \in \mathcal{T}$ were calculated in under 73 seconds. We fit the following marginal model using GEE, with and without random forest IPW weights, supplying the random forest algorithm with $\boldsymbol{H}_p*$ variables as described in Section B.3 of the Appendix as

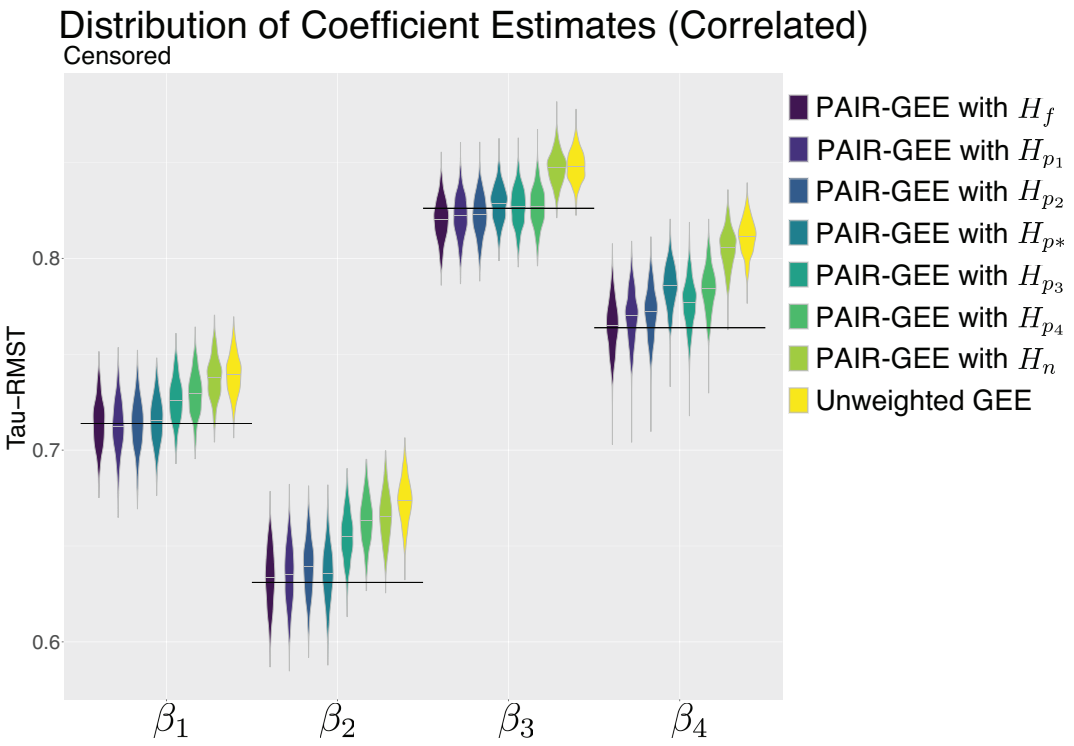


Figure 3.2: Violin plots of censored point estimates seen in simulation with censoring applied (500 replicates, $n = 750$). Coefficients β_1 and β_2 reflect those with coefficients $Z_i = 1$ and $Z_i = 0$ before $t = 6$ and coefficients β_3 and β_4 after and including $t = 6$, respectively. True coefficient values are indicated via horizontal lines. Because τ= 1 (month) in simulation, bias away from true parameter values multiplied by 100% can be interpreted as absolute percentage bias from the true paramter value. Features included in the random forest algorithm for estimating PAIR-GEE weights involve the seven history covariates described in Section 3.5 with varying degrees of availability that decreases from left to right for each displayed coefficient. Here, $\boldsymbol{H}_f$ indicates the full year of patient history was available prior to $t = 0$. $\boldsymbol{H}_{p_1}$ and $\boldsymbol{H}_{p_2}$ indicate that the patient history included information on the last two or one alternating recurrent events, respectively, if these occurred within the prior year. $\boldsymbol{H}_{p*}$ indicates that no historical information prior to $t = 0$ was available and that reflexive history covariates (as described in Section 3.5) were used. $\boldsymbol{H}_{p_3}$ indicates that complete history was available on 50% of individuals, but was otherwise unavailable. $\boldsymbol{H}_{p_4}$ indicates no availability of history covariates prior to $t = 0$. $\boldsymbol{H}_n$ indicates no history covariates were included in the random forest algorithm.

Table 3.1: A table displaying results from simulation in the case where primary recurrent events have correlation of $\rho = 0.8$, and secondary recurrent events are independent of each other. The hazard of primary alternating recurrent event completely determines the hazard of the alternating secondary recurrent event, approximately 30% of subjects are censored before the final follow-up window, and results come from 500 replicates of $n = 750$ each.

Model	Covariate	$\widehat{\beta} - \beta$	$\lvert\widehat{\beta} - \beta\rvert/\beta$	Coverage	$\widehat{SE}$	ESD	$\widehat{SE}$/ESD
PAIR-GEE, with $\boldsymbol{H}_f$	$\widehat{\beta}_1$	0.000	0.000	0.934	0.013	0.014	0.963
	$\widehat{\beta}_2$	-0.002	0.003	0.934	0.015	0.016	0.919
	$\widehat{\beta}_3$	0.006	0.007	0.946	0.011	0.011	1.046
	$\widehat{\beta}_4$	-0.001	0.001	0.928	0.013	0.014	0.929
PAIR-GEE, with $\boldsymbol{H}_{p_1}$	$\widehat{\beta}_1$	0.001	0.002	0.932	0.014	0.015	0.925
	$\widehat{\beta}_2$	-0.003	0.005	0.918	0.015	0.016	0.895
	$\widehat{\beta}_3$	0.004	0.005	0.950	0.011	0.011	1.018
	$\widehat{\beta}_4$	-0.006	0.008	0.884	0.013	0.014	0.921
PAIR-GEE, with $\boldsymbol{H}_{p_2}$	$\widehat{\beta}_1$	0.000	0.000	0.924	0.013	0.014	0.928
	$\widehat{\beta}_2$	-0.007	0.012	0.890	0.014	0.015	0.933
	$\widehat{\beta}_3$	0.004	0.004	0.954	0.011	0.011	1.018
	$\widehat{\beta}_4$	-0.008	0.011	0.856	0.013	0.014	0.933
PAIR-GEE, with $\boldsymbol{H}_{p*}$	$\widehat{\beta}_1$	-0.002	0.002	0.952	0.013	0.013	1.051
	$\widehat{\beta}_2$	-0.004	0.006	0.934	0.015	0.015	1.009
	$\widehat{\beta}_3$	-0.002	0.003	0.958	0.010	0.010	1.081
	$\widehat{\beta}_4$	-0.022	0.029	0.512	0.011	0.012	0.986
PAIR-GEE, with $\boldsymbol{H}_{p_3}$	$\widehat{\beta}_1$	-0.013	0.018	0.806	0.012	0.012	0.983
	$\widehat{\beta}_2$	-0.024	0.038	0.556	0.013	0.013	0.985
	$\widehat{\beta}_3$	-0.001	0.001	0.948	0.011	0.010	1.019
	$\widehat{\beta}_4$	-0.013	0.017	0.790	0.012	0.013	0.952
PAIR-GEE, with $\boldsymbol{H}_{p_4}$	$\widehat{\beta}_1$	-0.016	0.022	0.712	0.012	0.012	0.983
	$\widehat{\beta}_2$	-0.032	0.051	0.292	0.013	0.013	0.992
	$\widehat{\beta}_3$	-0.001	0.001	0.944	0.011	0.010	1.019
	$\widehat{\beta}_4$	-0.020	0.026	0.574	0.011	0.012	0.950
PAIR-GEE, with $\boldsymbol{H}_n$	$\widehat{\beta}_1$	-0.024	0.034	0.414	0.011	0.011	0.973
	$\widehat{\beta}_2$	-0.034	0.054	0.224	0.012	0.013	0.953
	$\widehat{\beta}_3$	-0.021	0.025	0.328	0.009	0.009	1.024
	$\widehat{\beta}_4$	-0.041	0.054	0.026	0.010	0.010	0.962
Unweighted GEE	$\widehat{\beta}_1$	-0.026	0.036	0.328	0.011	0.011	1.019
	$\widehat{\beta}_2$	-0.043	0.067	0.058	0.012	0.012	1.000
	$\widehat{\beta}_3$	-0.022	0.027	0.296	0.009	0.008	1.089
	$\widehat{\beta}_4$	-0.047	0.062	0.008	0.010	0.009	1.011

well as other patient characteristics:

$$\begin{aligned} E[\min\{T_i(t), \tau\}] = \beta_0 &+ \beta_1 I(\text{Ethnicity}_i = \text{Non-Hispanic}) + \beta_2 \text{Age}_i + \beta_3 I(\text{Race}_i = \text{White}) \\ &+ \beta_4 I(\text{Male}_i) + \beta_5 I(\text{Caregiver's TBI Patient}_i = \text{Independent}) \\ &+ \beta_6 I(\text{Caregivers' TBI Patient}_i = \text{Partially Independent}) \\ &+ \beta_9 I(\text{Push Notifications}_i). \end{aligned}$$

For both weighted and unweighted methods, time-to depression shortens for those on treatment while adjusting for other covariates, though the PAIR-GEE model accounts for additional uncertainty in point estimates, resulting in wider confidence intervals (Figure 3.3, Panel A). Additionally, the PAIR-GEE model detects a much stronger association between caregivers of independent TBI patients and time-to depression recovery (approximately 5.28 days fewer than caregivers for fully dependent TBI patients) than the complete case analysis (roughly 3.98 fewer days).

3.8 Discussion

We presented a novel IPW pseudo-observation based approach to estimating τ-restricted mean time-to-event for the primary event type using alternating recurrent event data, and discussed different covariates to be used in the estimation of these weights. In settings with high quality patient history information and correlated alternating recurrent events, our model is able to estimate τ-restricted means while correcting for the bias caused by implicit selection of subjects into the at-risk population that a complete case analysis suffers.

Of note is the weakness of the proposed model that does not use any history coefficients ($\boldsymbol{H}_n$); it performs almost as poorly as the unweighted model. We make three general recommendations on the use of history variables: (1) whenever possible, full history covariates should be obtained, unless when the hazard of the primary event type is constant over time; (2) if the full history is not available, at least the timings of the last alternating recurrent event pair within a subject is sufficient; and (3) in the absence of any historical information, reflective history that accounts for missingness pattern at $t > t_0$ may be substituted.

One of the benefits of our method is that the analysis of alternating recurrent events is agnostic to different primary event definitions. For example, in analyzing depression free time, the definition of $T_i(t)$ may be time-to depressive state after time t, but we also defined it as time-to exiting the depressive state after time t, that is, a reversal of primary and secondary event definitions. Our analysis method, then, brings to light the dynamics of mental health status, as in Section 3.7, where treatment decreased time-to depression and

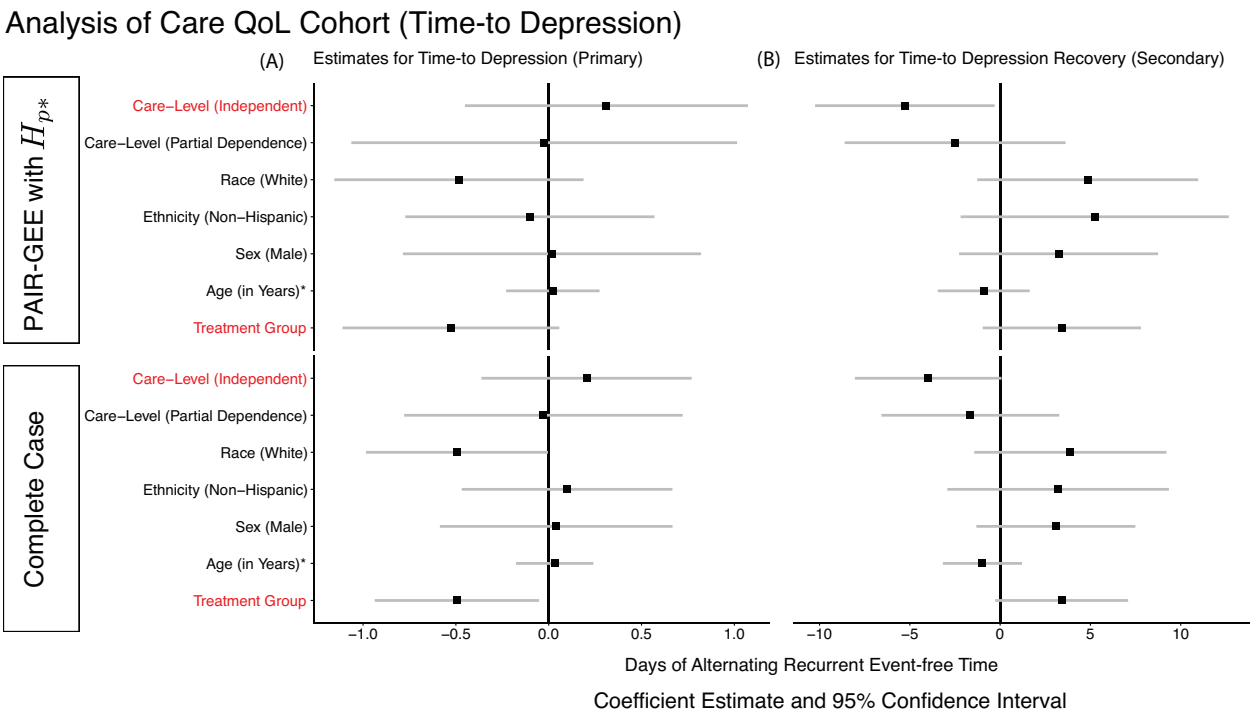


Figure 3.3: Point estimates and corresponding 95% confidence intervals from the CareQOL data analysis. The left and right panels correspond to the estimates of 2-week restricted mean time-to-entering and exiting depressed states, respectively. Significant coefficients, or coefficients that change significance between methods are displayed in a red font. Covariates that were standardized prior to inclusion in the model are denoted with an asterisk. Our analysis finds that after adjustment for common covariates, being in the intervention group was no longer significantly associated with a decrease in the number of days till entering a depressive state.

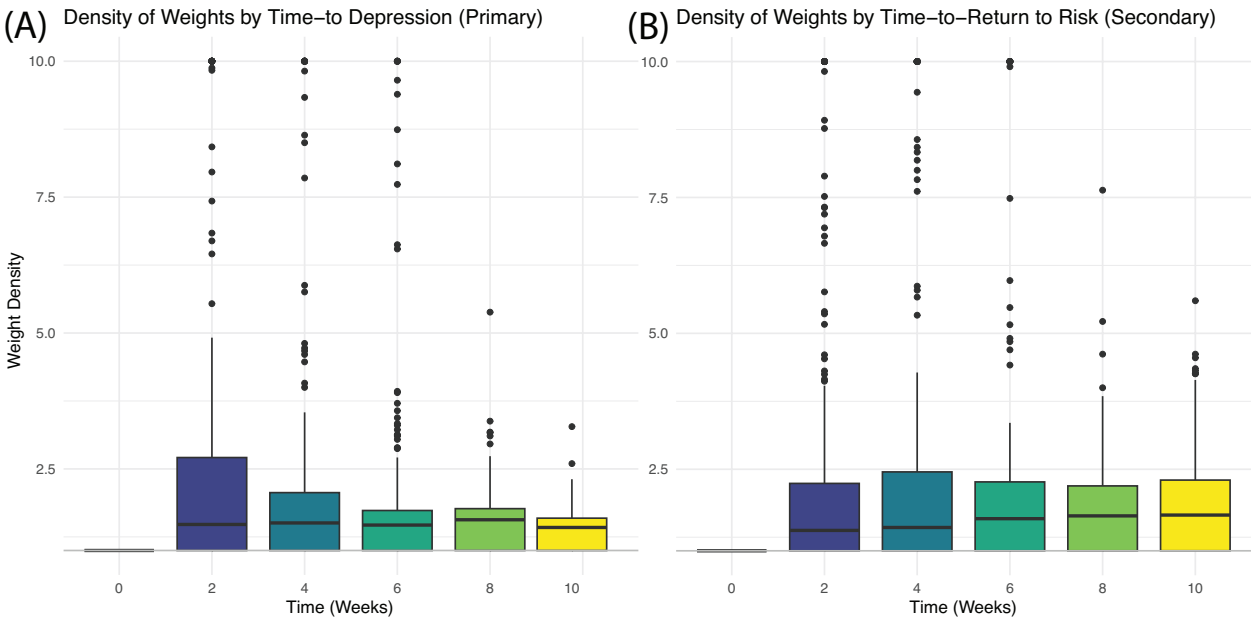


Figure 3.4: Distribution of inverse probability weights, based on random forest estimation of $\widehat{P}\{R_i(t)|\mathbf{Z}_i(t), \boldsymbol{H}_{p*}\}$ for subjects that are at-risk at each time point (Panel A), and subjects that are in a secondary state at time t (Panel B). Subjects are assumed to be at-risk at time $t = 0$. Weights were capped at 10.

increased time in depression recovery.

In simulations, we noted the consistent under coverage of the τ-restricted mean after a change in hazard was observed. This leads to interesting questions about how one might adjust for changes in the health state when calculating patient-specific history covariates. Important work in the future may consider kernel density-type calculations of history covariates to adjust for more recent patient specific histories, rather than treating all historical events as equally important at time t. As health data become richer and more temporally granular due to electronic health records and fitness trackers, these methodological refinements will be essential.

CHAPTER 4

A Neural Network-Based Dynamic Prediction Model for Alternating Recurrent Events*

4.1 Introduction

Alternating recurrent events are common outcomes in biomedical, sociological, political and behavioral disciplines. For example, periods of hospitalization alternate with periods of discharge. Broadly, an alternating recurrent event may be defined as a time-to-event outcome of interest (or *primary event*), which immediately triggers a refractory period (or *secondary event*), wherein an individual is not at-risk for the primary event. Examples may include hospitalization and discharge times as in Hu et al. [2011], periods of depressive mood and baseline mood as in Loe et al. [2025b], or recurrent dengue infections, as in Van Wyk et al. [2025].

Accurate and well-calibrated predictions of the next primary alternating recurrent event provide subjects and practitioners with a tool to understand the current health trajectory, as well as informing daily decision making. However, analysis is complicated by the two mutually exclusive event types; for example, when considering time-to next period of depressed mood for medical residents, time spent in a depressed episode competes with time not in a depressed mood. Analyses that fail to account for the time spent in the depressed state, risk biasing the analysis of the time-to the initialization of the next depressed period.

There are broadly two schools of thought when approaching recurrent event analysis: (1) state-based modeling using partially observed Markov processes, such as the S(E)IR models in time-series analysis and dynamical systems, where the transition probabilities between states are of interest [e.g., Ionides et al., 2006, Bretó et al., 2009], and (2) semiparametric intensity-based models, where an analyst obtains estimates of covariate effect on relative hazard. This paper focuses on extensions of the semiparametric alternating recurrent events

[0]First author work, with co-authors Dr. Susan Murray and Dr. Zhenke Wu.

modeling framework to dynamic prediction. To that end, Hu et al. [2011] extended the Cox-based partial likelihood by modeling hazards of events to obtain interpretable parameter estimates of the intensity of recurrent events. However, their approach does not address potential dependence between recurrent event times, and the resultant, well-known dependent gap-time issue. Lee and Cook [2019a] and Lee and Cook [2019b] chose to model dependent events with a copula approach. Loe et al. [2025b] explored a random forest inverse probability weight to de-bias a semiparametric model for alternating recurrent events, handling potential censoring via pseudo-observations. Each of the latter type of model provides, on some level, interpretable parameter estimates, and can inform an analyst of population level dynamics, but requires the specification of a moment-based relationship between covariates and an outcome.

Simultaneously, with the advent of modern computing methods, there has been an increased interest in black-box methods for individual dynamic predictions, which do not require an analyst to specify the any parametric relationship between an outcome and covariates. In the recurrent events case, Yokota and Matsuyama [2019] relied on the presence of smoothers within a semiparametric context to provide updated survival probabilities, while Zhao and Feng [2020] and Zhao et al. [2020] constructed a neural network and a random forest, respectively, for dynamic predictions of a single time-to-event. To date, we are unaware of any authors who have proposed dynamic prediction models for alternating recurrent events, particularly in the case of "online" or real-time predictions, where predictions for an individual are updated based on newly observed experience.

In this paper, we present a neural network assisted approach to dynamically predict the time to the primary recurrent event over a τ-length period of follow-up at prespecified check-in time points, borrowing from Tayob and Murray [2015] to transform the alternating recurrent events into a structure known as a censored longitudinal dataset. This data structure then allows for construction of pseudo-observations to address independent censoring mechanisms [Andersen et al., 2003, 2004]. To model these pseudo-observations, we utilize a long short-term memory neural network (LSTM). LSTMs were introduced by Hochreiter and Schmidhuber [1997] as a solution to the well-documented vanishing gradient problem that plagues recurrent neural networks that use previous outcomes for predictions (for more detail on vanishing gradients, see Hochreiter and Schmidhuber [1997, 1996] or Bengio et al. [1994]). LSTMs use a gating structure to elegantly incorporate previous observed behavior into fitted values at a time (or times) of interest, and can accommodate many different input and output data types. As a subtype of deep learning, LSTMs are widely recognized to excel in settings where there are many covariates with a possibly unknown relationship with an outcome, and prediction, rather than interpretation of coefficients, is of interest.

By including an inverse probability weight (IPW) in the LSTM loss function, our proposed network accounts for the alternating nature of the data. Rather than providing retrospective (offline), de-biased analysis of a cohort, this paper proposes an online dynamic prediction tool for (primary) alternating recurrent events. We use inverse probability weights (IPW) to correct potential bias, as in Loe et al. [2025b], but compare results in cases where the IPW weight is constructed based on a random forest or an LSTM. In constructing our LSTM weights and LSTM outcome model, we de-mystify LSTMs for a statistical reader. Our weight is then applied to a novel dynamic prediction framework using LSTM; however, the field of machine learning and prediction modeling evolves at a fast pace. While our framework is demonstrated by LSTM, it is appropriate for use with the more advanced ML methods plugged-in.

The rest of this paper is organized as follows. In Section 4.2 we describe the censored longitudinal data structure designed for analyzing alternating recurrent events. In Section 4.3, we create pseudo-observations for addressing independent censoring mechanisms. Sections 4.4 and 4.6 propose two different dynamic prediction models, the first based on a neural network and the second based on a semiparametric model. The weights used to fit both dynamic prediction models are described in Section 4.5, while the performance of all models is examined via simulations in Section 4.7. The proposed methods are applied to data arising from the Intern Health Study (IHS) in Section 4.8 to predict the start times of periods of low mood. The paper concludes with a discussion of limitations and future directions in Section 4.9.

4.2 Censored Longitudinal Data Structure

For individual i, let the alternating recurrent event times $T_{i,0}^{(1)} < T_{i,0}^{(2)} < T_{i,1}^{(1)} < T_{i,1}^{(2)} < \dots$ be potentially censored by the random variable C_i, where C_i is independent of $T_{i,j}^{(1)}, T_{i,j}^{(2)}$, $i = 1, ..., n$, and $j \geq 0$. Here, j indexes the event pairings, i indexes subjects, and the superscript $k \in \{1, 2\}$ indexes the alternating primary ($k = 1$) and secondary ($k = 2$) events. In the example from the Intern Health Study (IHS), $T_{i,j}^{(1)}$ marks times when an intern enters low-mood periods. Upon entering the low-mood state at $T_{i,j}^{(1)}$, intern i is no longer at-risk for a low-mood state at $t \in [T_{i,j}^{(1)}, T_{i,j}^{(2)})$, until they recover from the low-mood period at time $T_{i,j}^{(2)}$, and re-enter the population of individuals at-risk for periods of low mood. While for our purposes, we have defined $T_{i,j}^{(k)}$ in terms of periods of low and baseline mood, any such alternating structure may be described using these random variables.

For each individual, potential censoring results in observed data $X_{i,j}^{(k)} = \min(T_{i,j}^{(k)}, C_i)$, and censoring indicator, $\delta_{i,j}^{(k)} = I(X_{i,j}^{(k)} < C_i)$, $j = 1, \dots, J_i, k = 1, 2$. See Figure 4.1(A)

for a depiction of alternating recurrent events, subject to censoring. Rather than analyzing potentially dependent gap times, we follow the approach of Tayob and Murray [2015], Xia et al. [2020], Zhao et al. [2020], Loe et al. [2025a] and Loe et al. [2025b], and transform longitudinal data into a set of pre-specified follow-up windows of length τ, that start at $t \in \mathcal{T}_i = \{t_0, t_1, \ldots t_{b_i}\}$. For each $t \in \mathcal{T}_i$, and uncensored individual $i = 1, 2, ..., n(t)$ at time t, the censored longitudinal data structure maps the original alternating recurrent events, $T_{i,j}^{(1)} < T_{i,j}^{(2)}$ for $j \geq 0$, to the first alternating recurrent event after time t. Specifically, we construct the censored longitudinal data for each individual $i = 1, 2, \ldots n(t)$ is:

1) Identify the j index of the first alternating recurrent event of type $k = 1, 2$ after t by $\eta_i^{(k)}(t) = \min\{j | T_{i,j}^{(k)} > t, j = 0, 1, \ldots\}$;

2) Define residual censoring time from time t (the start of the follow-up window) as $C_i(t) = C_i - t$. Let $T_i^{(k)}(t) = T_{i,\eta_i^{(k)}(t)}^{(k)} - t$, with $X_i(t) = \min\{T_i^{(1)}(t), T_i^{(2)}(t), C_i(t)\}$ measuring time-to-first-observed event time after time t (regardless of its event type), and corresponding censoring indicator, $\delta_i(t) = I\{X_i(t) < C_i(t)\}$;

3) Use $R_i(t) = I\{T_i^{(1)}(t) < T_i^{(2)}(t)\}$, to indicate that individual i was at-risk for the primary alternating recurrent event as of time t. Note that we may modify the definition of $R_i(t)$ by switching the superscripts to shift focus from analyzing the primary event type to analyzing the secondary event type;

4) Construct the censored longitudinal dataset at t by $\left\{X_i(t), R_i(t), \delta_i(t), \mathbf{Z}_i(t), i = 1, 2, \ldots n(t)\right\}$ where $\mathbf{Z}_i(t)$ are observed covariates as of time t. Figure 4.1(B) displays the censored longitudinal data tuples (except the covariates), $\{R_i(t), X_i(t), \delta_i(t)\}$, that correspond to the censored alternating recurrent events depicted in Figure 4.1(A).

In this paper, let $T_i(t)$ be the time to primary type of recurrent event after time t. However, for any subject with $R_i(t) = 0$, $T_i(t)$ is unobserved, because at time t this subject has recently experienced the primary event, and has entered into a period anticipating the secondary, not the primary, event type. For example, in the IHS, $T_i(t)$ may be defined as time-to-low-mood after time t, but for subjects who are already depressed in mood at time t and waiting to recover, we do not observe $T_i(t)$. We will hereafter refer to $\mathcal{O}(t) = \{i = 1, 2, \ldots, n_R(t) : R_i(t) = 1\}$ as the set of individuals at time t with observed $T_i^{(1)}(t)$ and who have not reached a censoring event, with $|\mathcal{O}(t)| = n_O(t) \leq n(t)$. We define $\{X_i(t), \delta_i(t) : i \in \mathcal{O}(t)\}$ as the complete case (non-missing) data for the primary recurrent event of interest at t. In Sections 4.4 and 4.6, we propose models for dynamically predicting τ-restricted time-to-the-primary-event $\min\{T_i(t), \tau\}$ using $E[\min\{T_i(t), \tau\}]$. In particular, we introduce (weighted)

LSTM networks and weighted semiparametric models fit via generalized estimating equations to ensure that primary event analyses account for missing subjects due to the secondary state. Prior to providing model specifications in Sections 4.4 and 4.6, we discuss a pseudo-observation approach to account for censoring by $C_i(t)$.

4.3 Pseudo-Observation

Here, we provide a primer on the construction of weighted pseudo-observations, a variation of jack-knife methodology originally proposed by Andersen et al. [2003] and Andersen et al. [2004]. Pseudo-observations are typically constructed by selecting an unbiased estimator for an estimand of interest, i.e., for our dynamic prediction model, τ-restricted mean survival after time t, or $E[\min\{T_i(t), \tau\}]$. The estimator is then constructed from two different data subsets, one with all observations in $\mathcal{O}(t)$, and another of $\mathcal{O}(t)$ with the ith observation omitted.

Let $S^{\mathcal{O}(t)}(u) = P(T_i^{(1)}(t) > u)$ be the survival curve as a function of u corresponding for subjects at-risk for the primary event at time t, and $\theta(t) = E[\min\{T_i^{(1)}(t), \tau\}|R_i(t) = 1] = \int_0^\tau S^{\mathcal{O}(t)}(u)du$. $\theta(t)$ has a consistent nonparametric estimator $\widehat{\theta}(t) = \int_0^\tau \widehat{S}^{\mathcal{O}(t)}(u)du$, where $\widehat{S}^{\mathcal{O}(t)}$ denotes the Kaplan-Meier estimator for the survival curve using subjects in $\mathcal{O}(t)$. Similarly, one can construct $\widehat{\theta}^{(-i)}(t)$, by using the Kaplan-Meier survival function estimated from $\mathcal{O}(t)$ without subject i. Taken together, a pseudo-observation for $E[\min\{T_i^{(1)}(t), \tau\}]$ for subject i at time t is defined as

$$PO_i^\tau(t) = n_O(t)\widehat{\theta}(t) - \{n_O(t) - 1\}\widehat{\theta}^{(-i)}(t).$$

Andersen et al. [2003] showed that in regression settings, the univariate pseudo-observation may replace the bivariate censored survival outcome; for longitudinal settings, results pertaining to estimating equations from Robins et al. [1995] may be applied. These pseudo-observations are calculated using the cohort of individuals in $\mathcal{O}(t)$; details on how to correct the pseudo-observations during estimation to account for selection into $\mathcal{O}(t)$ are given in Section 4.5.

4.4 Statistical Review of Neural Networks

Neural networks are a popular way to model potentially complex functions with many variations and adaptations [see, for example Zazo et al., 2016, Radha et al., 2019, Peng et al., 2023]. Although many features of neural networks have a clear statistical basis, jargon for

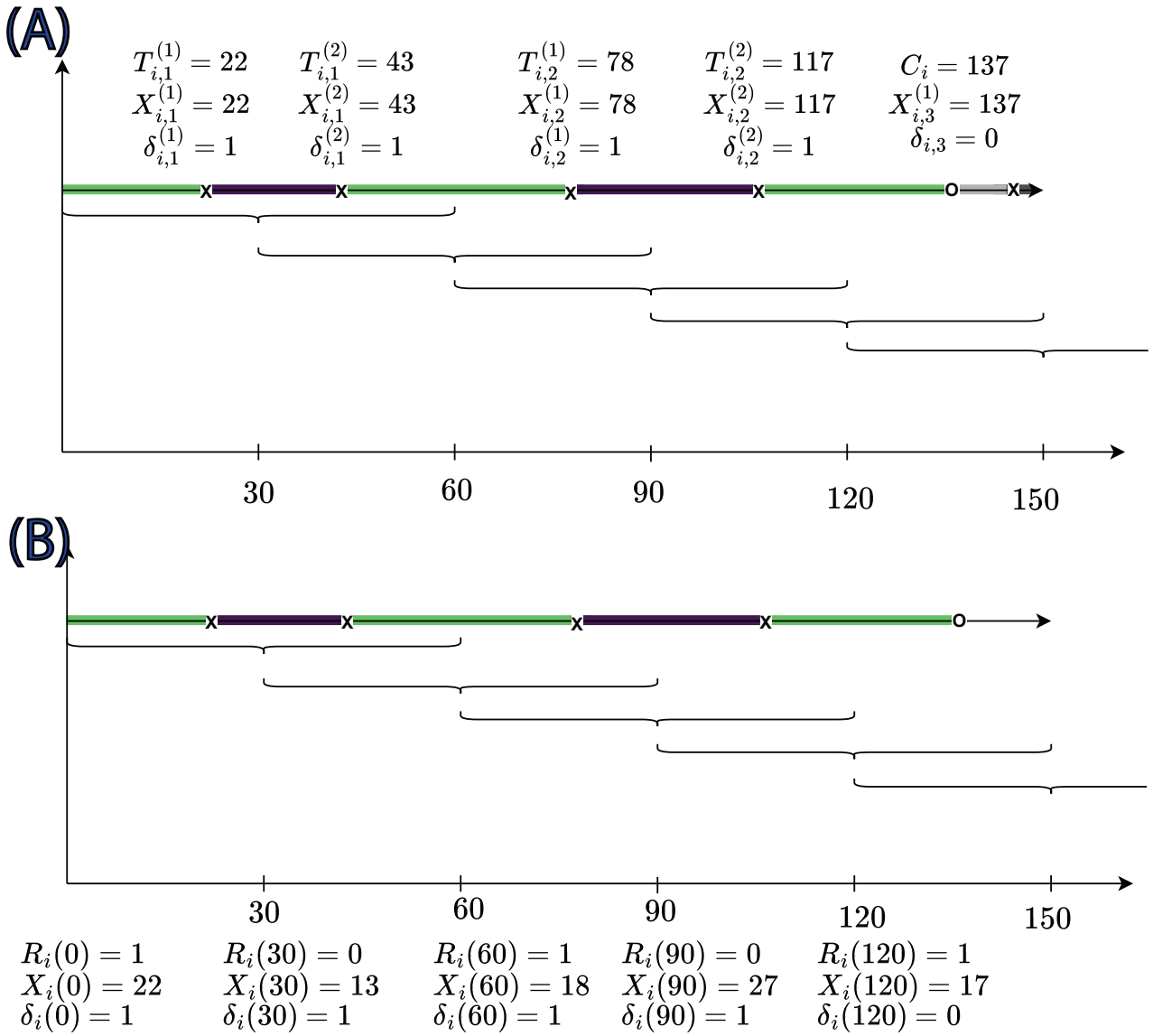


Figure 4.1: A figure depicting the transformation from alternating recurrent events into the appropriate random variables of the alternating censored longitudinal data structure corresponding to $\mathcal{T} = \{0, 30, 60, 90, 120, 150\}$ and $\tau = 60$ days (represented by the curly braces). Panel (A) depicts the traditionally observed alternating recurrent events, with uncensored event times, $T_{i,j}^{(k)}$ listed, with corresponding observed event times $X_{i,j}^{(k)}$ and indicator functions, $\delta_{i,j}^{(k)}$. Periods where subject i is at-risk for a primary alternating recurrent event are depicted in light green, while secondary or recovery states are shown in purple. The τ-length follow-up window corresponding to $\tau = 60$ days and $\mathcal{T}_i = \{0, 30, 60, 90, 120, 150\}$ is shown with curly braces. Panel (B) depicts the same $\mathcal{T}_i$, τ, and subject, but records the random variables appropriate for use with the alternating censored longitudinal data structure described in Section 4.2. Potentially time-varying covariates $Z_i(t)$ may be recorded at each $t = \{0, 30, 60, 90, 120, 150\}$.

describing these features are often quite different for computer science and statistical researchers, causing much unnecessary confusion. To familiarize statistical audiences with the neural network algorithm used in our research, we first introduce a neural network algorithm that corresponds to linear regression in Section 4.4.1. Then in Section 4.4.2 we describe a neural network algorithm appropriate for use with the censored longitudinal dataset of pseudo-observations. Additional practical considerations of applying this neural network algorithm in our setting are given in Section 4.4.3.

4.4.1 Linear Regression as a Neural Network

In the language of computer scientists, neural networks are a user-specified architecture (model) with hidden layers of nodes, consisting of weights on the edges and biases (similar to intercept; different from the bias typically used in statistics) applied to model inputs for the purpose of predicting an outcome. Consider the regression model $g(E[\boldsymbol{Y}|\mathbf{Z}_1,\ldots,\mathbf{Z}_p]) = \mathbf{Z}^\top\beta$, where $\boldsymbol{Y}, \mathbf{Z}_j, j = 1\ldots p$ are n-dimensional vectors, β is of dimension $p \times 1$, and $\mathbf{Z}$ is the $(p{+}1){\times}n$-dimensional design matrix, which includes the intercept term. Here and throughout this manuscript, we use bold face type to denote matrices or vectors. The $g(\cdot)$ function, or link function in statistical language, is typically the identity function for linear regression settings. Generally, we will use the term $g^{-1}(\eta)$ to denote the term that is used to minimize squared-error loss, i.e. $\{\mathbf{Y} - g^{-1}(\eta)\}^\top\{\mathbf{Y} - g^{-1}(\eta)\}$, where $\eta = \mathbf{Z}^\top\beta$ and η is of dimension $n \times 1$.

Translating neural network to statistical terminology: (1) model inputs are covariates, $\mathbf{Z}_j,\ j = 1,\ldots,p$; (2) weights are linear regression slope parameters $\beta_j, j = 1,\ldots,p$; (3) bias corresponds to the linear regression model intercept, β_0; (4) a node is the functional $g^{-1}(\mathbf{Z}^\top\beta)$, which reduces to $\mathbf{Z}^\top\beta$ in the setting mirroring linear regression with an identity link. Individual nodes make up the hidden layers of the neural network, termed hidden because neural network algorithms do not report estimates, $\widehat{\beta}$, just the vector of outcome predictions, $\widehat{\boldsymbol{Y}} = g^{-1}(\mathbf{Z}^\top\widehat{\beta})$. Computer scientists refer to $g^{-1}(\cdot)$ as an activation function that estimates the desired outcome, often in several different layers within the algorithm. Hereafter, we use the terms "slope" to refer to a weight, "intercept" to refer to bias, and "parameter estimates" to refer collectively to weights and biases used in the neural net contexts.

Figure 4.2 displays three increasingly more sophisticated neural networks in Panels (A), (B) and (C), where we include an additional subscript on η to reflect changes across panels. Panel (A) displays a linear regression model in a manner familiar to neural network researchers with a single hidden layer composed of one node; the corresponding statistical

optimization function is displayed below the network. Covariates are represented by circles on the left of the diagram, while the linear predictor is displayed in the rectangular box.

Researchers have found that an increase in the number of nodes in the hidden layer may improve prediction, particularly in the case where the relationship between covariates and outcome is more complex than a linear model. Figure 4.2, Panel (B) depicts a neural network with a single hidden layer composed of three nodes stacked on top of each other. The term, η_B, from Panel (B) is no longer a simple linear predictor as in panel (A), and involves parameter vectors $\alpha_1, \ldots, \alpha_3$ in addition to potentially different parameters, $\beta_1, \ldots, \beta_3$ applied to the same set of predictors, $\mathbf{Z}$. The form of g_ℓ^{-1}, $\ell = 1, 2, 3$ may also vary across nodes.

A network with two hidden layers is depicted in Figure 4.2, Panel (C); hidden layers 1 and 2 have three and two nodes, respectively. The term, η_C, in Panel (C) has increased in complexity from Panel (B). Within η_C, readers can identify terms resembling η_B fed into two different activation functions $h_1^{-1}(\cdot)$ and $h_2^{-1}(\cdot)$. Two issues emerge at this stage of neural network complexity that require comment.

First, the number of parameters involved in minimizing squared error loss based on the η in panel C has more than doubled. While the networks in Panels (B) and (C) have been found to potentially improve prediction compared to the network in Panel (A), parameter estimates are no longer guaranteed to be uniquely identifiable; it may be the case that there are multiple local minima of the loss function that are not the global minimum [Goodfellow et al., 2016], as well as multiple different matrices of parameter estimates that may provide the same predictions. This is an open area of research, with mathematicians, statisticians and computer scientists attempting to find conditions that guarantee network architecture and parameter estimate uniqueness [see, for example, Sussmann, 1992, Fefferman et al., 1994, Vlačić and Bölcskei, 2022]. In practice, however, most theorists recognize that it is generally sufficient if a local minimum is found; i.e. even local minima have sufficiently good predictive performance [Goodfellow et al., 2016].

The second issue that arises in Panel C is the potential for scaling differences to emerge between predicted outcomes generated within the hidden layers as the result of using so many different activation functions, $g_{\ell_1}^{-1}$ and $h_{\ell_2}^{-1}, \ell_1 = 1, 2, 3$ and $\ell_2 = 1, 2$, within the neural network architecture. Computer scientists have approached this issue pragmatically by using the tanh function to rescale elements within the neural net architecture to be between a common -1 to 1 scale to reduce problematic differences. Some practitioners refer to the tanh operation as "normalizing," though we should note that this is not normalizing in a statistical sense.

Under certain assumptions, including non-colinearity of covariates, a linear regression squared error loss function has a closed form solution for parameter estimates. This is not

typically the case for neural networks. Instead, iterative algorithms moving in smaller steps towards loss function (local) minima, via a hyperparameter known as the learning rate; each iteration is referred to as an epoch in neural network literature. Popular algorithms are stochastic gradient descent, batch gradient descent, "Adam" [Kingma and Lei, 2014], and adaptive gradient descent [Duchi et al., 2011], each with mini-batch variations that fall under the framework proposed by Robbins and Monro [1951]. For a comprehensive review, see Bottou et al. [2018].

Because of the large number of parameters built into neural network architectures, these algorithms are particularly susceptible to overfitting. A popular method used by both statisticians and computer scientists to prevent predictions being overfit to the data is to use different data cohorts for model training, validating, and testing. These terms have been used differently depending on the background of an individual researcher, so a quick summary of terms and techniques follows. First, data are split into model building and model testing cohorts. The model building cohort determines the final architecture of the neural network using a k-fold cross-validation algorithm to evaluate architecture hyperparameters including, for instance, the number of hidden layers and the number of iterations used in the optimization algorithm. For a particular architecture defined using a set of hyperparameters, the k-fold cross-validation algorithm (1) splits the model building cohort into training and validation cohorts along a $(k-1):1$ ratio, (2) minimizes the loss function (squared-error loss for linear regression) in the training cohort to establish a working neural network, (3) uses the working neural network applied to the validation cohort to obtain predictions, $\widehat{Y}$, and estimates of the loss function. Steps 1-3 are repeated for each of the k data splits with loss function estimates across the k validation cohorts averaged. This process is repeated for different sets of hyperparameters with a final architecture selection made based on smallest average estimated loss function across the validation cohorts.

Once the architecture hyperparameters are determined via k-fold cross-validation, the full model building cohort is used to rebuild the neural network (via re-minimizing the loss function in this cohort). The testing cohort is then applied to the final neural network to determine overall performance. For a detailed description of hyperparameter tuning and neural networks, see Goodfellow et al. [2016], Chapter 11, where decisions on optimization algorithm learning rate and number of epochs are discussed at length. We will later summarize how these three groups may be more flexibly defined in the context of longitudinal outcomes in Section 4.4.3.

To illustrate a typical process, we now describe the process of fitting the linear regression neural network depicted in Figure 4.2, Panel (A) on training data, though the process generalizes to the architectures displayed in Panels (B) and (C). For each epoch, $a = 1, 2, \ldots n_a$

(where n_a is often between 10 and 200):

1. Calculate fitted values of $\widehat{\mathbf{Y}}^{(a-1)}$ for the linear regression loss function $\widehat{\mathbf{Y}}^{(a-1)} = \mathbf{Z}^\top \widetilde{\beta}^{(a-1)}$, where (a) tracks the epoch. For the first epoch, traditionally $\widetilde{\beta}^{(0)} = 0$, though some authors note that sufficient convergence to a local minima can depend on the initial starting values [Wang et al., 1993].

2. Calculate the updated loss function values. For linear regression, this would be squared error loss with the formula $\mathcal{L}(\mathbf{Y}, \widehat{\mathbf{Y}}) = (\mathbf{Y} - \widehat{\mathbf{Y}}^{(a-1)})^\top (\mathbf{Y} - \widehat{\mathbf{Y}}^{(a-1)}) = \{\mathbf{Y} - g^{-1}(\mathbf{Z}^\top \widetilde{\beta}^{(a-1)}\}^\top \{\mathbf{Y} - g^{-1}(\mathbf{Z}^\top \widetilde{\beta}^{(a-1)}\}$, where g^{-1} is typically taken to be the identity function. More generally, loss functions mimic those seen in generalized linear models, reflecting the assumed distribution of $\mathbf{Y}$.

3. Take partial derivatives of the loss function with respect to each parameter $\beta_0, \ldots, \beta_p$ to determine the gradient of the loss function. Update $\widetilde{\beta}^{(a-1)}$ to $\widetilde{\beta}^{(a)}$ in a step size corresponding to the learning rate and optimization algorithm of your choice. For example, with stochastic gradient descent, $\widetilde{\beta}^{(a)} = \widetilde{\beta}^{(a-1)} - \bigtriangledown \frac{\partial}{\partial \widetilde{\beta}^{(a-1)}} \mathcal{L}$, where $\bigtriangledown$ is the learning rate, and $\frac{\partial}{\partial \widetilde{\beta}^{(a-1)}} \mathcal{L}$ is the partial derivatives of the loss function with respect to each parameter.

4. Repeat steps 1-3 until n_a is reached.

4.4.2 Long Short-term Memory Neural Networks

There are many variations of (recurrent) neural network architectures intended to accommodate longitudinal data. Here we review LSTMs as the method we favor in our proposed modeling approach, largely for three reasons: (1) implementable software is available, (2) LSTMs are known to handle long-range time-dependencies [Hochreiter and Schmidhuber, 1996], and (3) much effort has been made by computer scientists to de-mystify their performance. As the field progresses, this stage of our analysis approach can be replaced with more modern versions of deep learners, such as transformers or gated recurrent units.

A bare-bones LSTM for a single time-point, t_0, is depicted in Figure 4.3 A, with two hidden layers, an intercept (circled **1**) and a single covariate (circled $\mathbf{Z}_{t_0}$). Figure 4.3 B depicts a generalization of the LSTM network at time t_k; taken together, Panels A and B recursively define the architecture termed LSTM. Unlike the networks depicted in Figure 4.2 where each covariate and node is connected to subsequent layers, the o_{t_k} node skips from the first hidden layer to the calculation of η_{t_0}. This style of neural network architecture is designated as not fully connected in computer science literature.

LSTMs have the advantage of connecting architecture designed for particular times, $t \in \{t_0, \ldots, t_b\}$, into an overall super-architecture. Together, panels 4.3 A and 4.3 B recursively define a full LSTM super-architecture, where nodes for $C_{t_{k-1}}$ and $\eta_{t_{k-1}}$ connect the architecture for t_{k-1} to that for $t_k, k = 0, \ldots, b$. As in Section 4.4.1, $\mathbf{Z}_{t_0}^{\top}\beta$ indicates the use of a design matrix, where β is assumed to contain the intercept term β_0. The contributions of $C_{t_{k-1}}$ and $\eta_{t_{k-1}}$ are traditionally weighted by other nodes in the t_k-level architecture (a strategy in deep learning known as "gating"). Gating connects longitudinal information between architectures, while avoiding a well-known "vanishing gradient" problem (see, for further details, Bengio et al. [1994]).

Fo a dynamic online prediction of the $\tau-$restricted mean for the follow-up window that starts at $t_k, t_k \in \mathcal{T}$, or $E[\min\{T_i(t_k), \tau\} \mid \bar{\mathbf{Z}}_i(t_k)]$, where $\bar{\mathbf{Z}}_i(t)$ is the vector of covariates values $\mathbf{Z}_i(t_k), \mathbf{Z}_i(t_{k-1}) \ldots \mathbf{Z}_i(t_0)$, we follow the neural network architecture laid out in Figure 4.3B, noting that for dynamic prediction of $E[\min\{T_i(t_k), \tau\} \mid \bar{\mathbf{Z}}_i(t_k)]$, the super-architecture only connects architectures across previous follow-up windows. That is, we select $\beta_{\cdot,t}, \alpha_{\cdot,t}, t < t_k - \tau$ that minimize:

$$\arg\min_{\beta_{\cdot,t},\alpha_{\cdot,t}} \sum_{t<t_k-\tau} \{\widehat{\mathbf{PO}}^{\tau}(t) - g^{-1}(\eta_t)\}^{\top}\mathbf{W}(t)\{\widehat{\mathbf{PO}}^{\tau}(t) - g^{-1}(\eta_t)\},$$

where $\mathbf{W}(t)$ is a $n(t) \times n(t)$ diagonal matrix, with diagonal entries $\frac{R_i(t)}{\widehat{P}\{R_i(t)=1|\bar{\mathbf{Z}}_i(t)\}}$, and zeros off the diagonal. Components of η_{t_k} are laid out in Figure 4.3B. For estimation of $E[\min\{T_i(t_k), \tau\} \mid \bar{\mathbf{Z}}_i(t_k)]$, the final activation link, g, is taken to be the identity link. Estimation of $\widehat{P}\{R_i(t) = 1 \mid \bar{\mathbf{Z}}_i(t)\}$ is covered in Section 4.5.

4.4.3 Useful Extensions to Bare-bones LSTM

Neural networks and the bare-bones LSTM described in Section 4.4.2 have many proposed extensions. In this section, we selectively describe (1) common parameters available from the off-the-shelf Python module `torch.nn`, and (2) a few additional neural network design choices in common use that we selected for use with `torch.nn`, tailored for our analysis. Options in `torch.nn` include: η (computer scientists call this hidden state), parameter pruning (computer scientists call this the drop-out proportion), number of LSTM layers, and the encoding of similarities within individuals via embeddings. Additional options we specified for our algorithm included a grid search to simultaneously optimize the number of training epochs and the learning rate for the stochastic optimization algorithm.

A hyperparameter favored by computer scientists is allowing η_t to be extended from a scalar to a matrix in $\mathbb{R}^{n(t)\times h}$, with columns $\eta_t^{(1)}, \eta_t^{(2)}, \ldots \eta_t^{(h)}$, and similarly $\mathbf{Z}_i(t)$ to be

extended from a vector in $p \times 1$ to a matrix of dimension $\mathbb{R}^{p \times n(t)}$. Computer scientists call h the hidden dimension. For dimensions $h > 1$, the inverse link is applied to a linear combination of the h columns of η_t, i.e. $g^{-1}(\gamma_0 + \gamma_1 \eta_t^{(1)} + \ldots \eta_t^{(h)})$.

Beyond increasing h, computer scientists have also found that vertically stacking additional architecture at time t increases prediction performance [Graves, 2014]. General recommendations are for at most 4 layers of architecture to be "stacked" on top of each other [as used in Sutskever et al., 2014]. For instance, the entire architecture at time t seen in Figure 3B might be stacked on top of a similar architecture at time t, where the (η_{t_k}) elements taken from the right side of the top stacked architecture play the role of the covariate inputs (Z_{t_k} terms) on the left of the bottom stacked architecture. Additional subscripts on parameters and link functions in the bottom stacked architecture would complete this version of stacking architecture within time t.

The final two off-the-shelf options include (1) regularization via drop-out and (2) embeddings. Regularization via drop-out refers to randomly forcing a proportion of parameters, p_d, to 0 during each epoch of training [Srivastava et al., 2014, Zaremba et al., 2014]. Subject and time point embeddings are a concept borrowed from text mining [Mikolov and Zweig, 2012, Mikolov et al., 2013, Sutskever et al., 2014]. For each individual, a dimension b vector of embedding values may be allocated to quantify similarity of outcomes experienced by individuals over time. These embedding values are treated as parameters to be estimated in addition to other neural network outputs, and are available after training is complete.

Training, validation and testing cohorts in the real-time LSTM prediction setting require a bit more explanation than was given in the simpler linear regression-oriented neural network setting described in Section 4.4.1. Here we also wish to use grid search methods to optimize, for instance, the number of training iterations selected from the set, $\{1, \ldots, n_a\}$, and the learning rate, $\triangledown$ selected from a size m grid of candidate learning rates, $\{\triangledown_1, \triangledown_2 \ldots, \triangledown_m\}$, where the optimal values of these hyperparameters may differ for dynamic predictions at the different follow-up window start times, $t_k \in \mathcal{T}$. While in the following, we describe a grid search for two hyperparameters, all hyperparameters may be selected via such an algorithm, though a large number of points in a grid increases significantly increases computation time. Suppose our current focus is finding optimal hyperparameters for the follow-up window starting at t_k. At all $t \in \mathcal{T}$ the network described in Figure 4.3 may be fit. However, in constructing a training and validation set for hyperparameter selection, we require at least one window with complete follow-up for each of the training and validation data cohorts for use in the more sophisticated rolling-origin algorithm (proposed by Armstrong and Grohman [1972] and more concisely summarized by Tashman [2000].) This results in data from (at least) t_0 and t_1 forming a model warm-up period, where predictions are less trustworthy

due to a lack of hyperparameter validation. In utilizing the rolling-origin algorithm for the optimal learning rate and number of training epochs, for each candidate learning rate, $\nabla_\ell, \ell = 1, \ldots, m$, we first train on data observed up to t_{k-2}, using observed data at t_{k-1} as a validation set, assuming $t_{k-1} < t_k - \tau$ (if $t_{k-1} > t_k - \tau$, t_{k-1} and t_{k-2} revert to the two most recent window start times with complete follow-up). The testing cohort becomes observed data at t_k. Behavior of the loss function in the training set is recorded after each training iteration completes for each learning rate, so that the (joint) optimal number of training iterations, $n_{opt}(t_k)$, and the optimal learning rate, $\nabla_{opt}(t_k)$, can be selected. Corresponding loss function performance in the validation cohort using the same set of candidate values for these hyperparameters are then recorded.

Loss function decreases for each additional training iteration are expected in the training set, but not necessarily in the validation set. When increasing the number of training iterations from 1 to $n_a(t_k)$ in the validation set for a particular learning rate, ∇_ℓ, if $\max_p$ consecutive increases in the loss function are seen, $n_{opt}(t_k)$ is set to the number of training iterations, prior to the start of when these increases were seen (called early stopping in computer science literature); otherwise $n_{opt}(t_k) = n_a$. The value of $\max_p$ is a user chosen value traditionally between 5 and 10 that is called the tolerance or patience in computer science literature. Further details on early stopping may be found in Prechelt [1998]. Validation loss and recommended number of epochs are recorded for each candidate learning rate, with final learning rate and recommended epochs, $\nabla_{opt}(t_k)$ and $n_{opt}(t_k)$ respectively, selected based on the pairing that yields minimal loss in the validation set. A second network is then trained with $\nabla_{opt}(t_k)$ and $n_{opt}(t_k)$ on data from $\{t_0, t_1, \ldots t_{k-2}, t_{k-1}\}$. Network performance for dynamic prediction is tested using data observed at t_k. For a depiction of this algorithm, see Figure 4.4.

This custom grid-search has three attractive properties: (1) learning rate is generally considered one of the most important hyperparameters to tune [Goodfellow et al., 2016], (2) the splitting algorithm maps neatly onto the censored longitudinal data, and (3) this algorithm allows for the use of learned subject specific embeddings to model correlation within individuals.

4.5 Weight Estimation

The pseudo-observations from Section 4.3 are constructed only using the set of individuals in $\mathcal{O}(t)$, yet Loe et al. [2025b] showed that by weighting pseudo-observations with inverse probabilities, an analyst may account for non-random selection into the at-risk set. The inverse probability weights may either come from an LSTM, as described in Section 4.4, or

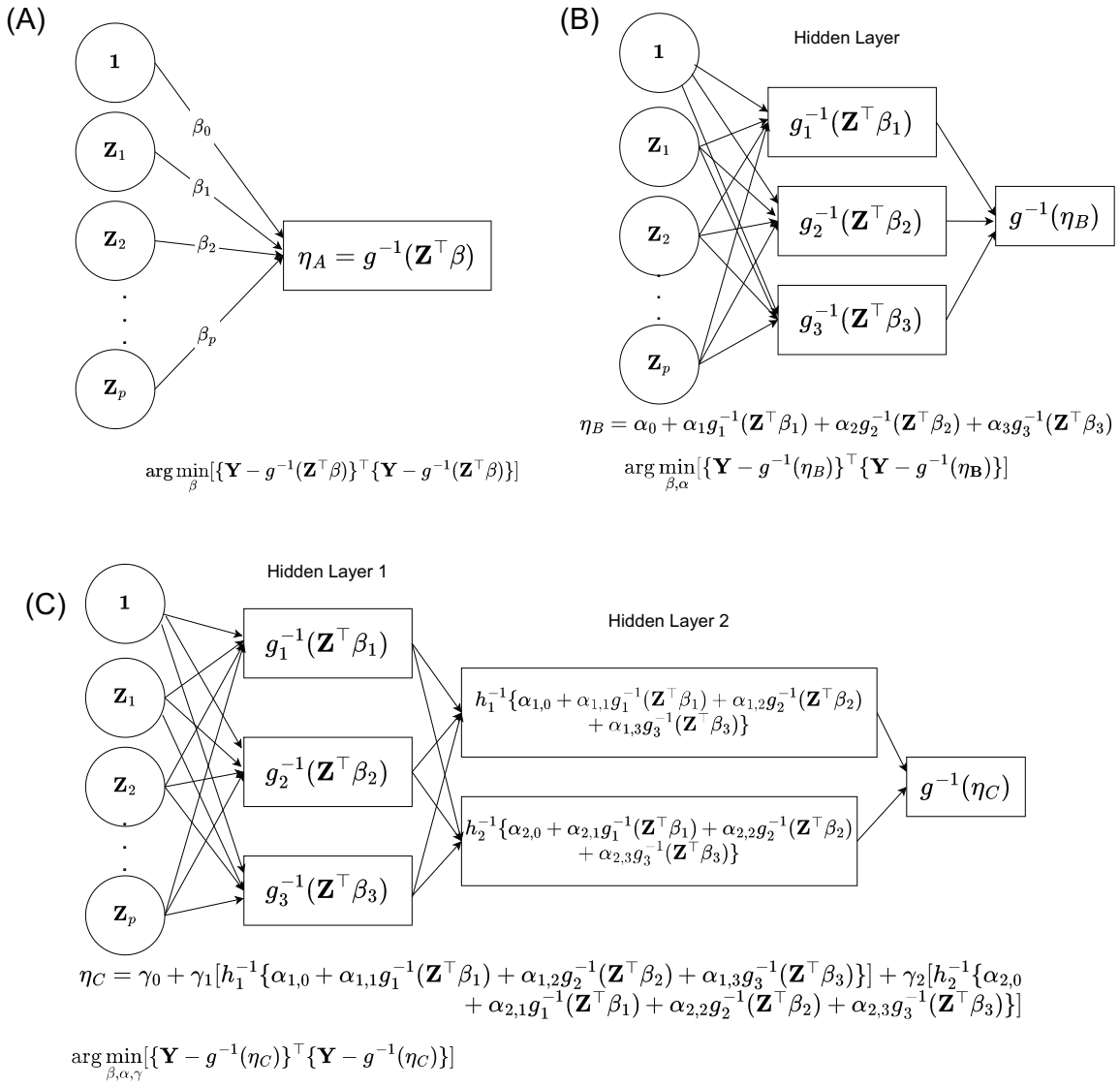


Figure 4.2: Three fully connected neural network architectures of increasing complexity, with squared error loss. Vectors of covariates are depicted in circles on the left, while nodes containing estimable parameters are displayed in rectangular boxes. Networks are called "fully connected" because each covariate is passed to each node in the first layer, and for networks with more than one layer, outputs from previous layers are passed to all subsequent layers. Panel (A) displays an architecture with one hidden layer containing one node, akin to linear regression in the special case where g is the identity link. Panel (B) displays a network architecture with three nodes in one hidden layer, that are ultimately combined to obtain the linear component, η_B. As Panel (C) contains multiple hidden layers, Panel (C) proposes a network that may also be termed a "deep learner."

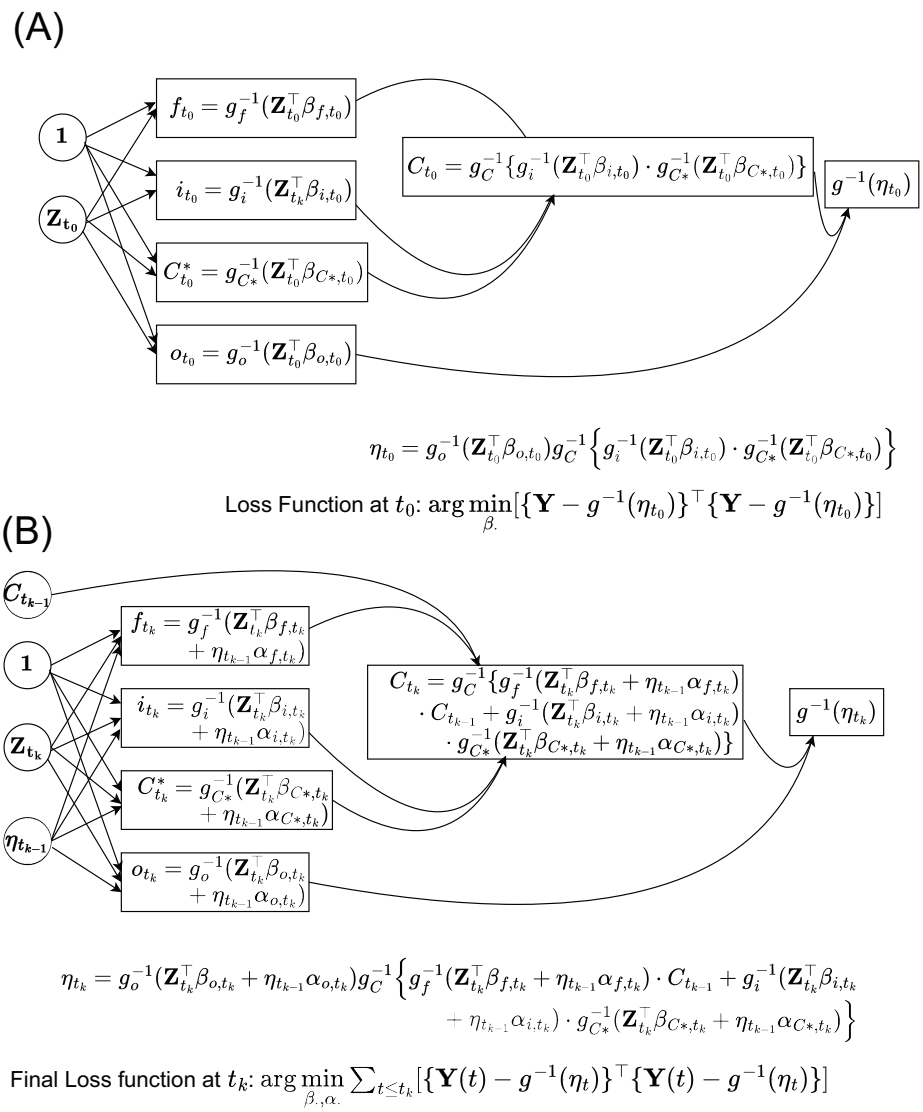


Figure 4.3: A panel graphic displaying the network architecture termed "long short-term memory," or LSTM at time t_k. Panel (A) displays the functionals used in training an LSTM at time t_0 (base case), while Panel (B) displays the self-referential architecture for later $t \in \mathcal{T}$ (recursive). When connected recursively across time points, Panel A and Panel B create an LSTM at multiple $t \in \mathcal{T}$. Software package `torch` allows user-specification of links g_f, g_i, g_{C*}, g_o and g_C but defaults to g_f, g_i, g_{C*} and g_o as the log-link and g_C and g_{C*} as the inverse hyperbolic tangent; thus $g_{f,i,C*,o}^{-1}(x) = \frac{\exp(x)}{1+\exp(x)}$, and $g_{C,C*}^{-1}(x) = \tanh(x)$. In Section 4.7, we used defaults from the `torch` package in Python. In contrast to Figure 4.2, where outputs from one layer are combined via another generalized linear model functional, node outputs from $C_{t_k}^*$ and i_{t_k} are combined together via a dot product, as are f_{t_k} with $C_{t_{k-1}}$, and C_{t_k} with o_{t_k}. Not all covariates are used in all nodes in the first layer; covariate $C_{t_{k-1}}$ is only used in the second layer, while o_{t_k} is calculated in the first layer, before final combination in the third. At each time, inputs that depend on the algorithm at the prior time point itself are included ($C_{t_{k-1}}$ and $\eta_{t_{k-1}}$). At t_0, these quantities default to 0. In Panel (B), while ultimate output of interest is $g^{-1}(\eta_{t_k})$, loss from the previous time point factors in to the modeling of η_{t_k}.

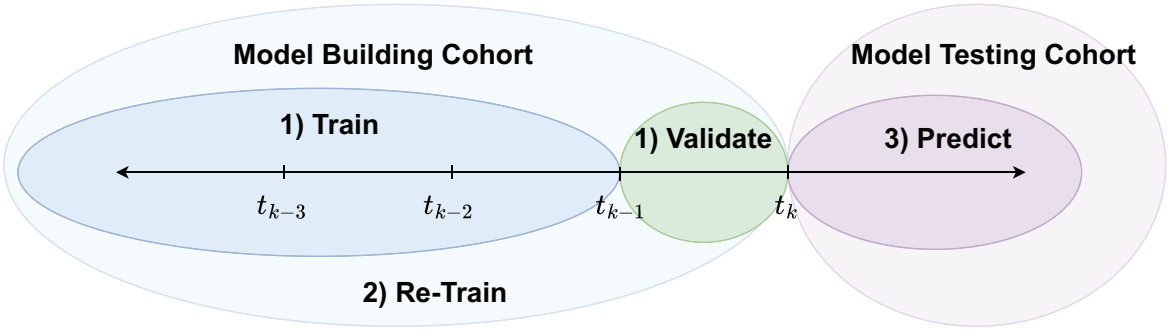


Figure 4.4: A figure depicting the training, validation, and testing split for the rolling-origin algorithm, mapped onto the censored longitudinal data structure. The model building component contains data from $\{t_0, t_1, \ldots t_{k-1}\}$, while model testing comes from data observed at t_k. Within the model building cohort, the network is first trained on data from $\{t_0, t_1, \ldots t_{k-2}\}$, while data observed in the window starting at time t_{k-1} are used to validate. When the optimal number of epochs has been reached, n_{opt}, initial training stops. Re-train the model using data observed at window start times $\{t_0, t_1, \ldots t_{k-1}\}$. This model is then tested for data at t_k, to obtain dynamic predictions of $E[\min\{T_i(t_k), \tau\}]$.

a longitudinal random forest, as proposed by Adler et al. [2011]. In estimating $\widehat{P}\{R_i(t) = 1|\bar{\mathbf{Z}}_i(t)\}$ via LSTM, we assume the network depicted in Figure 4.3 has a logistic link function g, and select $\beta_{\cdot,t}, \alpha_{\cdot,t}$ that minimize:

$$\arg \min_{\beta_{\cdot,t_k}, \alpha_{\cdot,t_k}} \sum_{t \leq t_k} \left[\mathbf{R}(t) \cdot \log\{g^{-1}(\eta_t)\}\right] + \left[\{\mathbf{1} - \mathbf{R}(t)\} \cdot \log\{\mathbf{1} - g^{-1}(\eta_t)\}\right].$$

This network is trained using observed values $R_i(t)$, $i = 1, \ldots n(t)$ for all time points up to and including $t_k \in \mathcal{T}$. We perform a 2-fold validation/training algorithm for selecting number of epochs, and obtain fit values from the validation set. Specifically, for each $t_k \in \mathcal{T}$, we split the data into two equal sized training and validation cohorts, based on subject IDs, and including all time points $t \leq t_k$. Training on a given learning rate, ∇_m proceeds with the training set, until validation loss fails to decrease for $\max_p$ epochs. The training model corresponding to the best validation loss is restored, and $\widehat{P}\{R_i(t_k) = 1|\bar{\mathbf{Z}}_i(t_k)\}$ is calculated for the validation cohort via the trained model. The procedure then repeats, with the cohort labels reversed. Because this is a cross-validation algorithm, and does not rely on the rolling-origin cross validation algorithm from Section 4.4.3, the model warm-up set is simply $\{t_0\}$; thus for values $t_1, \ldots, t_b$, one may obtain LSTM weights in this manner.

In this paper, we compare the performance of weights estimated via LSTM (as described above) and a longitudinal random forest. Random forests have many of the same desirable properties as neural networks; they can fit a model to potentially correlated covariates, as well

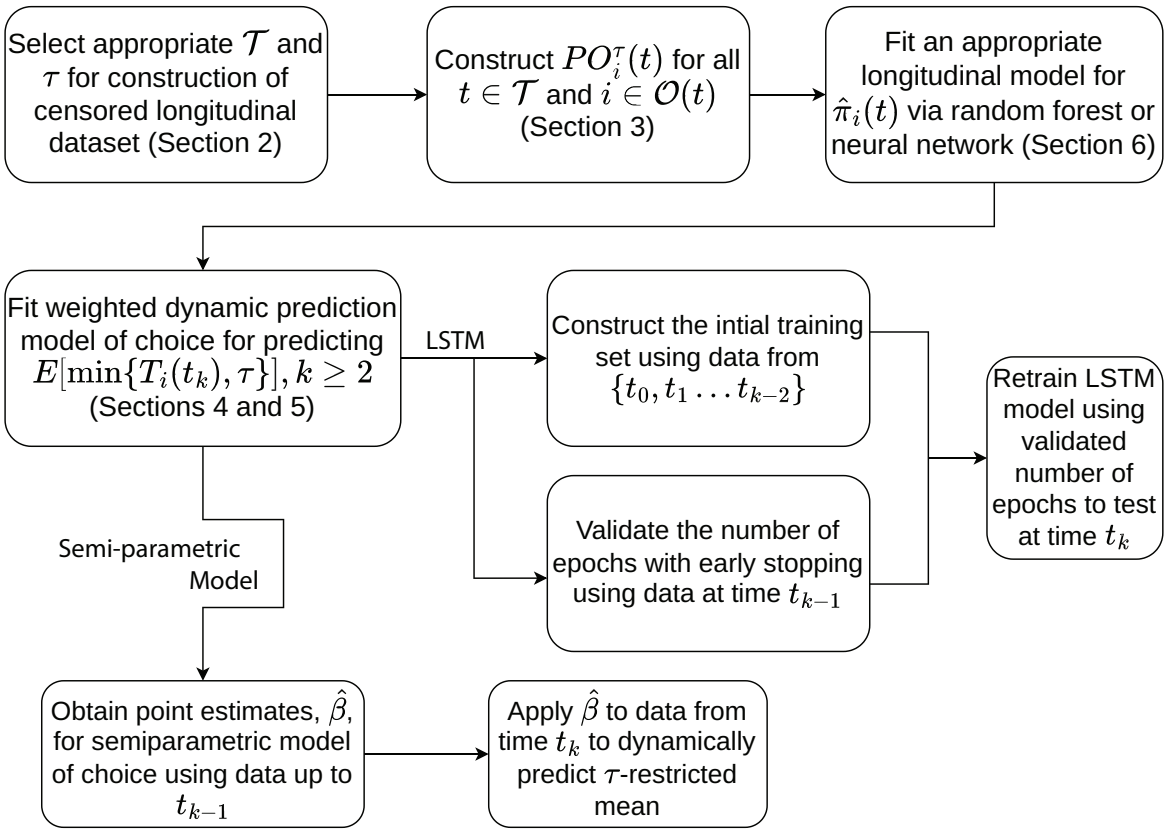


Figure 4.5: A figure displaying the analysis pipeline for on-line (real-time) prediction using random forest or LSTM IPW weights and the rolling-origin training and validation split for both the LSTM and GEE-based models.

as non-linear relationships between covariates and outcomes. Additionally, random forests may be programmed to always incorporate certain predictors into their model, providing the analyst with a greater degree of control than a more black-box neural network. Generally, they have fewer hyperparameters to tune (the number of trees, the minimum number of observations in a node, and the maximum number of splits within a tree), and rather than use a training and testing cohort, random forests use a strategy known as "bagging," which naturally creates distinct cohorts of data for training and testing (for further detail, see Loe et al. [2025b]). We apply the longitudinal random forest for dynamic prediction of at-risk status. A flowchart depicting the entire analysis pipeline is shown in Figure 4.5.

4.6 Weighted Semi-parametric Dynamic Prediction Model

At each $t_k \in \mathcal{T}$, we may fit $E[\min\{T(t_k), \tau\}|\bar{\mathbf{Z}}_{t_k}] = \beta^\top \mathbf{Z}_{t_k}$ via generalized estimating equations. As this paper is concerned with real-time dynamic predictions, at each t_k, we select data corresponding to $t < t_k - \tau$ for estimating $\widehat{\beta}$, and data at t_k to obtain predictions. Similar to the LSTM training, testing and validation algorithm described in Section 4.4.2, we need a model warm-up period while waiting for complete follow-up in the first τ-length window. Point estimates are then the solution to:

$$\mathbf{0} = \sum_{i=1}^{n} \bar{\mathbf{Z}}_i \mathbf{W}_i \mathbf{\Lambda}_i^{-1} (\widehat{\mathbf{PO}}_i^{\tau} - \bar{\mathbf{Z}}_i^\top \beta)$$

where $\mathbf{W}_i$ is a diagonal matrix (which weights least squares loss) with entries $\frac{R_i(t)}{\widehat{P}\{R_i(t)=1|\bar{\mathbf{Z}}_i(t)\}}$, for $t = t_0, t_1 \ldots t_{k-1} < t_k - \tau$ and $\mathbf{\Lambda}_i$ is the usual matrix product of variance functions and the working correlation structure. Estimates from this method will be termed "online GEE," in reference to the suitability for online, or real-time, predictions. Parameter estimates may then be applied to the cohort of data observed at t_k. Estimates from this model have the benefit of interpretability, and do not suffer from the same concerns surrounding tuning and convergence in deep learning approaches. However, in the presence of highly correlated covariates, such as the useful history covariates of Loe et al. [2025a], a semiparametric approach may not be fully estimable, particularly for elements of the variance-covariance matrix

4.7 Simulation Study

The goals of our simulation study are three-fold: (1) to examine the one-step online predictive ability of the proposed LSTM model for the primary recurrent event of interest when both primary and secondary events have non-trivial associations with covariates, (2) to describe pragmatic approaches to including history covariates in the algorithms, and (3) to evaluate the performance of a longitudinal random forest inverse weight compared to an LSTM inverse weight.

We first describe the generation of covariates, $\mathbf{Z}$ that are used in generated correlated gap times between alternating recurrent evens, $G_{i,j}^{(1)} = T_{i,j}^{(1)} - T_{i,j-1}^{(2)}$ and $G_{i,j}^{(2)} = T_{i,j}^{(2)} - T_{i,j}^{(1)}$. We draw $Z_1 \sim \text{Cauchy}(\theta = 4, \sigma = \frac{1}{100})$, $Z_2 \sim \Gamma(3, \frac{1}{3}), Z_3 \sim \text{Beta}(5, 1)$, $Z_4 \sim \text{Exponential}(1.7)$, $Z_5, Z_6 \sim \text{MVN}([13, -.5]^\top, \begin{pmatrix} 1 & .9 \\ .9 & 1 \end{pmatrix})$, Z_7 through Z_{100} are random normal noise covariates, with $\mu = \text{Unif}[-1, 1]$ and $\sigma^2 = 1$, $Z_{101} \ldots Z_{124}$ are multivariate normal, where Z_{101} to Z_{108} have mean 0.3, Z_{109} to Z_{116} have mean 0.1 and Z_{117} to Z_{124} have mean -0.3, each with variance

0.4 and correlation 0.28.

To obtain correlated gap times, we draw from a multivariate normal distribution with $\mu_i = (\mu_i^{(1)}, \mu_i^{(2)}, \mu_i^{(1)}, \mu_i^{(2)}, \ldots)$ and variance-covariance matrix Σ, which has a block diagonal form, with entries in each block:

$$a_{j,k} = \begin{cases} \rho^{(1,1)}\sigma^2 & j \neq k, j \text{ and } k \text{ are both odd} \\ \rho^{(2,2)}\sigma^2 & j \neq k, j \text{ and } k \text{ are both even} \\ \rho^{(1,2)}\sigma^2 & j \neq k, j \text{ and } k \text{ have different parity} \\ \sigma^2 & j = k \end{cases}$$

Here, $\rho^{(k,\ell)}$ is the correlation between $G_{i,j}^{(k)}, G_{i,j}^{(\ell)}$. We select $\rho^{(1,1)} = 0.3$, $\rho^{(2,2)} = 0.2$, $\rho^{(1,2)} = -0.2$, and $\sigma^2 = 10^{-5}$. We defined $\mu_i^{(1)} = \exp(-\sin Z_1 + 2Z_2Z_3 + (Z_2Z_4^2)^{\frac{1}{3}} - 3\sqrt{Z_4} + \cos(Z_5)Z_6^2 + \mid Z_6 \mid + 2\sin\{\cos(Z_4)\} - 2\tan(Z_3))$, and $\mu_i^{(2)} = \exp(Z_{101} + \ldots Z_{124})Z_3(\widetilde{Z}_6 + |\widetilde{Z}_6 + .3|)$, where $\widetilde{Z}_6$ is centered and scaled Z_6. Summing across gap-times yields alternating recurrent event time $T_{i,1}^{(1)} < T_{i,1}^{(2)} < \ldots$, and we follow the advice of Xia et al. [2020] and apply a burn-in period of at least $b_{pre} = \max \frac{5}{\lambda_i}$ duration prior to defining the analysis dataset. Combinations of covariates that yield $\mu_i^{(1)}$ values outside of 0.45 and 2 and $\mu_i^{(2)}$ outside of 0.2 to 0.6 are discarded. The censored longitudinal data structure takes the form $\{0, .5, 1, \ldots, 11.5\}$ months, with $\tau = 0.5$ months and $C_i - 2 \sim \text{Exponential}(\frac{1}{20})$ to mimic a year of follow-up with approximately 40% of observations censored.

As the goal of this paper is to dynamically model $E[\min\{T_i(t), \tau\}]$, we consider two settings of history covariates proposed by Loe et al. [2025a] and extended in Loe et al. [2025b]. The first setting is where subjects have an exact recollection of their last pair of alternating recurrent events within the previous year, termed $\boldsymbol{H}_h$ for "high-quality history," and the second set of history covariates, $\boldsymbol{H}_\ell$ for "low-quality history," where no history prior to study time is recorded, though once follow-up commences, these covariates may be recorded. In this latter case, Loe et al. [2025a] proposed a multiple imputation approach for handling missing history covariates. Here, we describe results using a pragmatic approach that incorporates a single imputation of the history covariates at t_0 from the average patient population. For later follow-up windows with sufficient observed patient history for each individual, this t_0 impute is dropped from patient history covariate calculations. The various algorithms include history covariates along with other covariates in estimating inverse probability weights as well as modeling $E[\min\{T_i(t), \tau\}]$; GEE models include history covariates as main effects.

Neural network inverse weights are estimated with an LSTM architecture with 1 dimensional subject embeddings, 8 dimensional time embeddings, 1 layer, $\eta \in \mathbb{R}^{n(t)\times 32}$ and $p_d = 0$. Hyperparameters for the inverse weight LSTM were selected via a grid search. Neural net-

work architecture for dynamic predictions of the τ-restricted mean was also selected using a rolling-origin grid-search; based on the types of covariates available ($\boldsymbol{H}_\ell$ or $\boldsymbol{H}_h$) as well as inverse weighting strategy (LSTM or RF), different architectures yielded lower testing losses. Exact hyperparameter specifications are shown in Table 4.1. Each network has 1 dimension for subject embeddings, and 8 dimensions for time embeddings. The grid search prioritizes deeper architectures (larger h) for LSTM weights, and more dropout for $\boldsymbol{H}_h$ settings.

We select n_{opt}, ∇_{opt} from candidates $1, \dots 250$ and $\{0.005, 0.001\}$, respectively, using the rolling-origin algorithm from Section 4.4.3. We compare this LSTM to two different online semiparametric GEE models. The first model is $E[\min\{T_i(t), \tau\} \mid \mathbf{Z}_i(t)] = \beta^\top \mathbf{Z}_i(t)$, where $\mathbf{Z}_i(t)$ includes covariates $\mathbf{Z}_1$ through $\mathbf{Z}_6$ in the correct functional form used for modeling $\mu_i^{(1)}$ as well as history covariates. We term this model "GEE Oracle," though exact oracle modeling of $T_i(t)$ would include terms that count number of observed events (for further details, see Section C.2). The second online GEE model, or "GEE Main Effects," considers $E[\min\{T_i(t), \tau\} \mid \mathbf{Z}_i(t)] = \beta^\top \mathbf{Z}_i(t)$ where $\mathbf{Z}_i(t)$ contains $\mathbf{Z}_1, \dots \mathbf{Z}_{124}$ as well as history covariates as main effects.

4.7.1 Evaluation Metrics

Model performance is evaluated in the testing set using average bias, the c-statistic (also referred to as the index of concordance or Harrell's C-index) [Harrell Jr et al., 1982] and root error sum of squares, $\sqrt{\frac{1}{n_{\mathcal{O}(t)}} \sum_{i \in \mathcal{O}(t)} [\widehat{E}[\min\{T_i(t), \tau\} | \bar{\mathbf{Z}}_i(t)\} - \widehat{PO}_i^\tau(t)]^2}$, denoted "Testing RESS." C-statistics measure predictive discrimination, and a value of approximately 0.50 indicates poor discriminative ability, while perfect prediction yields a c-statistic of 1. Testing RESS is the root of the average of error sum of squares; while it is not the function of theoretical expectation $\sqrt{E_{T_i(t)}[\{T_i(t) - \widehat{PO}_i^\tau(t)\}^2]}$, or RMSE, it is an observed average that approximates RMSE. As an estimator of RMSE, RESS provides information about both bias and variance of predictions in the held-out testing set.

The true value of $E[\min\{T_i(t), \tau\} \mid \mathbf{Z}_i(t)]$ needed to define bias is determined upon noting that the conditional distribution of $T_i(t)$ given $\mathbf{Z}_i(t)$ is $\text{Normal}\Big[\{N_i^{(1)}(t^-) + 1\}\mu_i^{(1)} + N_i^{(2)}(t^-)\mu_i^{(2)} - (t + b_{pre}), \widetilde{\sigma}^2\Big]$, where $N_i^{(k)}(t^-) = \sum_j I(T_{i,j}^{(k)} < t + b_{pre})$ is the number of events of type k experienced by subject i prior to time t including those accumulated during the burn-in period of duration, b_{pre}. $T_i(t)$ has variance $\widetilde{\sigma}^2 = \sum_{1 \le j \le \{N_i^{(1)}(t^-) + N_i^{(2)}(t^-) + 1\}} \sum_{1 \le k \le \{N_i^{(1)}(t^-) + N_i^{(2)}(t^-) + 1\}} a_{j,k}$, where $a_{j,k}$ is the j, k entry of Σ defined above. From this we use the relationship,

$$E[\min\{T_i(t), \tau\} \mid \mathbf{Z}_i(t)] = \int_{-\infty}^{\tau^-} y f_{T_i(t)}(y) dy + \tau\{1 - F_{T_i(t)}(\tau)\},$$

where $f_{T_i(t)}$ and $F_{T_i(t)}$ are the probability distribution functions and cumulative distribution functions of the time-to-first primary alternating recurrent event at time t, $T_i(t)$. For further details on the theoretical distribution of $T_i(t)$, see Supplemental Materials, C.2

4.7.2 Results

Figures 4.6 and 4.7 display dynamic prediction model results based on the testing dataset at each $t \in \mathcal{T}$. See Supplementary Materials Figures C.2 and C.3 for metrics with corresponding 95% confidence bands. Figure 4.6 reflects the scenario where only history variables collected during study follow-up are known ($\boldsymbol{H}_\ell$ setting) and Figure 4.7 reflects the setting where history covariates are known at all times, $t \in \mathcal{T}$ ($\boldsymbol{H}_h$ setting). Panels A, B, C and D show the average c-statistic, percent bias, RESS results, and percent absolute bias respectively, all in the testing set (data held out and not used in training). In each panel, GEE Oracle results are graphed in yellow, GEE Main Effects results are in black, and LSTM model results are in sky blue. Each dynamic prediction model is assessed with use of random forest or LSTM IPW weights, which are labeled in figure legends parenthetically after the model designation and graphed with a dashed line (RF weight), or a solid line (LSTM weight).

The GEE Oracle model performs well in either scenario, of course, with performance metrics continuing to improve as on-study history covariates enrich over time. All models generally improve as history covariate information grows over time. Evaluating inverse weighting methods for the GEE Main Effects model and GEE Oracle model, both RF and LSTM inverse weights perform similarly in terms of the c-statistic, RESS and bias metrics. For GEE Oracle, at later time points, RF weights appear slightly less biased in $\boldsymbol{H}_\ell$ settings, while LSTM weights are slightly less biased in $\boldsymbol{H}_h$ settings. When considering GEE Main effects, at later points LSTM weights have slightly less bias for $\boldsymbol{H}_\ell$ settings, and RF weights have slightly less bias for $\boldsymbol{H}_h$ settings. Bias at early windows is attributed to the online prediction nature of fitting both inverse weights and restricted means. Lack of data at early windows causes machine learning model based fitted values for $P\{R_i(t) = 1\}$ to be less reliable than fitted values at later windows (which are based on more data).

At later time points, the long short-term memory network outperforms the GEE Main Effects model in both $\boldsymbol{H}_\ell$ and $\boldsymbol{H}_h$ settings in terms of c-statistic and testing RESS, though LSTM-based models initially performed poorly at early follow-up windows. In the $\boldsymbol{H}_\ell$ case, the LSTM with LSTM weights and RF weights both have superior c-statistics and minimal bias at later windows; at some time points, the LSTM-based restricted mean model c-statistics even outperform the c-statistic corresponding to the semiparametric oracle model. The LSTM model with LSTM weights performs well in all follow-up windows in terms of

bias, with the least bias of all approaches in early follow-up windows; bias results for LSTM approaches over time were more variable than GEE approaches.

In the $\boldsymbol{H}_h$ setting, the richer covariate history greatly improves performance of the GEE Main Effects approach relative to the $\boldsymbol{H}_\ell$ setting, with higher c-statistics over time that are competitive with GEE oracle after midway through the study period. The LSTM based models initially perform poorly in terms of c-statistic, testing RESS, and bias. However, metrics rapidly improve over the first third of the study period, corresponding to a need for plenty of follow-up for the LSTM-based models. LSTM RESS for later time points is roughly half of that corresponding to the oracle model (Figure 4.7, Panel C), with outstanding c-statistics relative to both semiparametric models. However, percent bias appears to stabilize around 5% for the neural network architectures, while the oracle model stabilizes around 0, in a classic bias and variance trade-off.

In general, the selected architecture for the neural networks in simulation justifiably favored smaller RESS over smaller bias, as we selected architectures based on lowest testing loss (weighted squared errors). Excellent performance of LSTM-based prediction architectures in this setting echoes improvements seen in other neural network applications in machine learning literature. Specifically, use of embeddings allows for correlation within time and individuals to be modeled implicitly in the LSTM, while increased dimensions of $\eta(t)$, particularly with $\boldsymbol{H}_h$ covariates, allows for complex emergent relationship between covariates and outcomes. There are no equivalent such customizations in semiparametric modeling, though embeddings loosely resemble random effects modeling.

History	Weight	Subj. Embed Dim.	Time Embed Dim.	h	No. LSTM layers	p_d
$\boldsymbol{H}_\ell$	LSTM	1	8	8	1	0
	RF	1	8	4	1	0
$\boldsymbol{H}_h$	LSTM	1	8	64	2	0.2
	RF	1	8	4	1	0.2

Table 4.1: A table displaying the hyperparameters selected for the LSTM models in Section 4.7. Hyperparameters were selected based on lowest testing loss in a rolling-origin grid search. Weight refers to the weighting strategy used in the neural network loss function, while "Subj. Embed Dim." and "Time Embed Dim." refer to the dimension of the vector that the network uses to learn time and subject characterizations. Hyperparameter h refers to the dimension of the "hidden state," or number of columns of the linear predictor η, where $\eta \in \mathbb{R}^{n(t)\times h}$. The number of LSTM layers refers to the number of LSTM architectures stacked on top of each other, and p_d is the proportion of slopes and intercept randomly set to 0 during each epoch of training. For further details on these parameters, see Section 4.4.2.

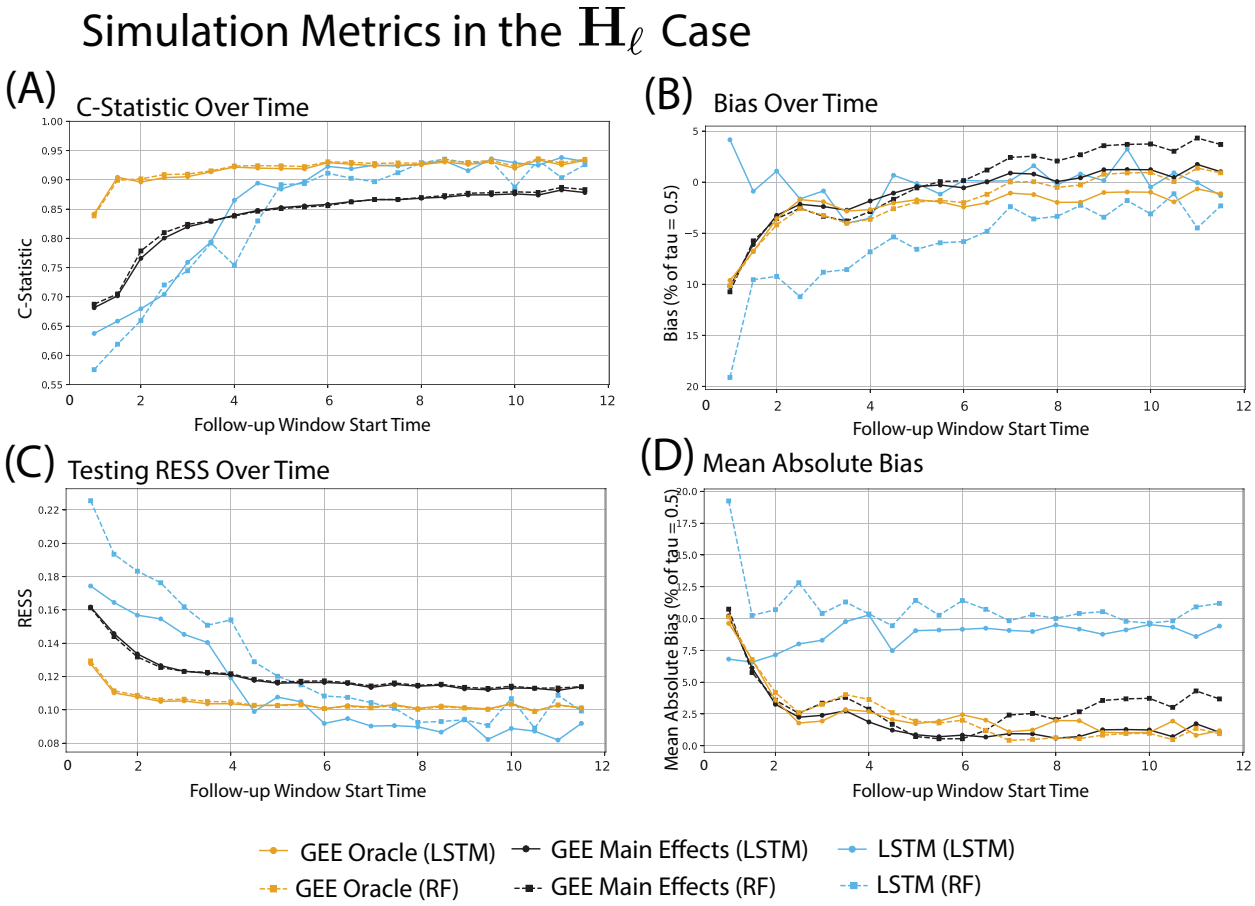


Figure 4.6: A panel graphic displaying the performance of the dynamic prediction models under consideration in Section 4.7 for the $\boldsymbol{H}_l$ setting. Counter clockwise from upper left is (A) c-statistic over time, (B) percent bias (in terms of τ) over time, (C) testing root mean squared error over time, and (D) mean absolute bias over time. The method used to predict $E[\min\{T_i(t), \tau\}] = \bar{\mathbf{Z}}^\top \beta$ is listed first (yellow for GEE Oracle, black for GEE Main Effects, and blue for LSTM), while the weight model is listed in parenthetical (with LSTM weights graphed in a solid line, while random forest weights are dashed). All results are from 200 replicates with $n = 2500$ each.

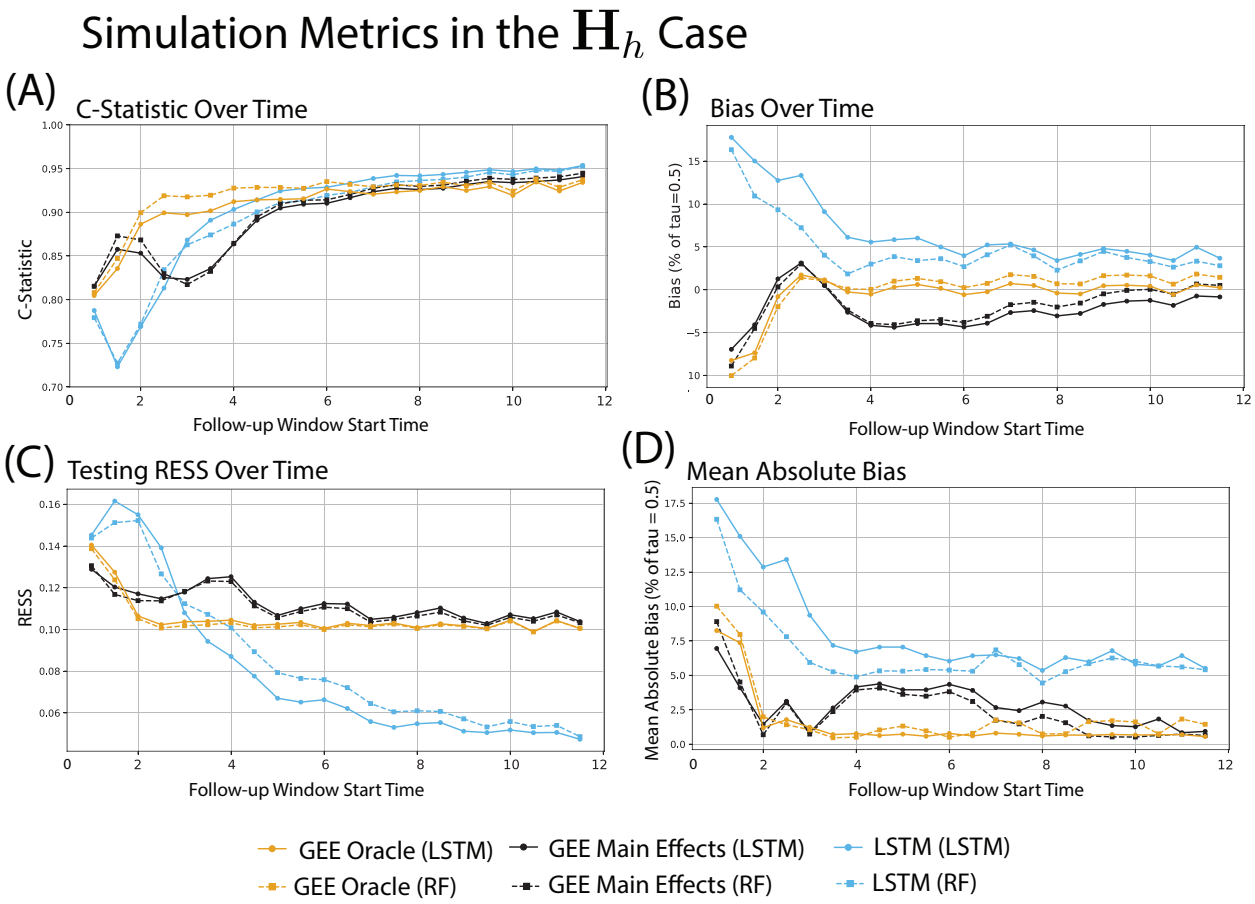


Figure 4.7: A panel graphic displaying the performance of the dynamic prediction models under consideration in Section 4.7 for the $\boldsymbol{H}_h$ setting. Counter clockwise from upper left is (A) c-statistic over time, (B) average bias over time, (C) testing root mean squared error over time, and (D) mean absolute bias over time. The method used to predict $E[\min\{T_i(t), \tau\}] = \bar{\mathbf{Z}}^\top \beta$ is listed first (yellow for GEE Oracle, black for GEE Main Effects, and blue for LSTM), while the weight model is listed in parenthetical (with LSTM weights graphed in a solid line, while random forest weights are dashed). All results are from 200 replicates with $n = 2500$ each.

4.8 Application to Intern Health Study Data

Medical residency, or the period of medical training after medical school, is known to be a stressful period of time that may lead to depression with a rate higher than in the general population [Guille et al., 2010], with negative effects on job satisfaction, physician retention and medical errors [Fang et al., 2022]. It is of interest to identify opportune moments of potential interventions, e.g., delivered via mobile app prompts, to improve the mental health and general well-being of the training physicians [NeCamp et al., 2020, Wang et al., 2023]. We uses data obtained from the 2023 cohort of the Intern Health Study, a multi-center study, where interns self-reported mood on a scale of 1 to 10 (10 being best) daily. It is of clinical interest to dynamically predict periods when mood for individual interns falls 3 points below baseline mood scores.

Covariates available for analysis include: age, gender, marital status, race, ethnicity, residency location and program, SEQ (a measure of adverse sexual experiences), Risky Families Questionnaire score (a measure of exposure to adverse conditions as a child), neuroticism at baseline, daily steps, sleep minutes, resting heart rate, mobile health device (Garmin, Apple Watch or FitBit), bedtime, wake time, if a participant has a history of depression, neuroticism or disability, any treatment for depression, substance use, and birth place. Incomplete baseline covariates are handled via a single imputation of data using random forest [van Buuren and Groothuis-Oudshoorn, 2011]. As there are no mood scores for individuals prior to study start time, history covariates are imputed using a random draw from an appropriate exponential distribution at t_0 (as in Loe et al. [2025a]), and updated at later window start times.

The censored longitudinal data structure assumes the form $\mathcal{T} = \{0, 14, 28, \ldots, 350\}$ days with $\tau = 14$ days. As in Section 4.7, IPW weights are estimated via longitudinal random forest or an LSTM model at each time point. The inverse probability LSTM uses 2 layers, $h = 70$, 2 dimensions for subject and time embeddings, and $p_d = 0.3$, with 2-fold cross-validation to select number of epochs to train.

We fit the following main effects model via GEE at each time point for dynamic prediction:

$$E[\min\{T_i(t), \tau\} \mid \mathbf{Z}_i(t)] = \beta^\top \mathbf{Z}_i(t),$$

where $\mathbf{Z}_i(t)$ contains all covariates listed above, and derived history covariates. History covariates include the gap times between last recurrent event pairing, time since last mood score drop, average time spent in the risk population ($R_i(t) = 1$), average time spent with lowered mood, and observed number of lowered mood events. Parameter estimates based on the GEE main effects model at $t = 350$ are displayed in Supplemental Materials C.3.

LSTM architectures for the neural network weights and the random forest weights are shown in Table 4.2. The network selected for fitting a restricted mean model using LSTM weights is relatively simple, particularly when compared to the model selected for use with RF weights. The architecture corresponding to LSTM with RF weights has fewer embeddings for subject and time, but the hidden size (number of columns of η) is over five times that of the hidden size of the LSTM with LSTM weights. In addition to this complexity, the rolling-origin grid search algorithm indicated the use of three separate LSTM layers with drop out of 0.2.

Figure 4.8 shows predictive performance via c-statistics over time and RESS in the testing cohort. The neural network algorithm had outstanding c-statistic performance, with a maximum of approximately 0.97 at day 266 or 238 (Panel A), depending on weighting scheme used. In comparison, the GEE Main Effects model c-statistic topped out at 0.823 on day 210. The manner of inverse weight estimation gave similar results in terms of c-statistics shown in Panel A. Panel B shows smaller RESS compared to the GEE Main Effects model. The LSTM model using LSTM inverse weights had the smallest RESS of all methods after approximately day 300 on study.

History	Weight	Subj. Embed Dim.	Time Embed Dim.	h	No. LSTM layers	p_d
$\boldsymbol{H}_\ell$	LSTM	1	8	6	1	0
	RF	1	1	32	3	0.2

Table 4.2: A table displaying the hyperparameters selected for the LSTM models in Section 4.8. Hyperparameters were selected based on lowest testing loss in a rolling-origin grid search at t_3. Weight refers to the weighting strategy used in the neural network loss function, while "Subj. Embed Dim." and "Time Embed Dim." refer to the dimension of the vector that the network uses to learn time and subject characterizations. Hyperparameter h refers to the dimension of the "hidden state," or number of columns of the linear predictor η, where $\eta \in \mathbb{R}^{n(t)\times h}$. The number of LSTM layers refers to the number of LSTM architectures stacked on top of each other, and p_d is the proportion of slopes and intercept randomly set to 0 during each epoch of training.

4.9 Discussion

In this manuscript we developed several dynamic prediction approaches for the alternating recurrent event setting that leverage the power of popular modern machine learning methods, and discussed the relative strengths and weaknesses of these methods. Of particular importance is the care with which we introduced the long short-term memory neural network machinery, with an eye towards bridging the notational and lexical gap between statistical

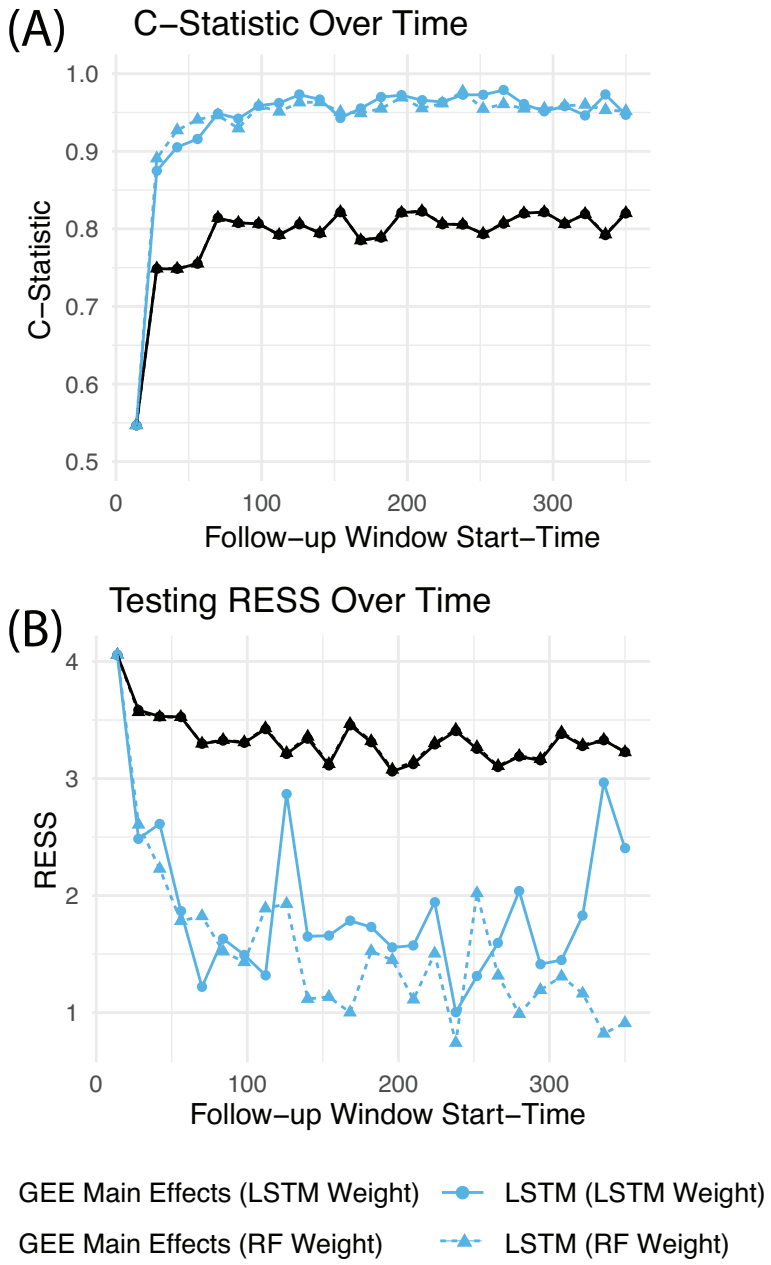


Figure 4.8: A panel graphic displaying results from predicting periods of low mood for medical interns. The censored longitudinal data strucutre takes the form $\tau = 14$, $\mathcal{T} = \{0, 14, 28, \ldots, 350\}$. Panel (A) displays c-statistic, while Panel (B) displays testing RESS. The main effects model fit via GEE is graphed in black, while the LSTM models are shown in sky blue. The method used to calculate IPW weights is listed in parenthetical after the model name.

and computer science practitioners. Strengths of neural networks (LSTMs included) and random forest approaches for building prediction tools in other settings have been well documented [e.g., Loe et al., 2025a, Breiman, 2001, Hochreiter and Schmidhuber, 1997, Mikolov and Zweig, 2012, Beck et al., 2024, Goodfellow et al., 2016]. This strong performance is somewhat offset by the lack of parameter estimates with well described statistical behavior that one may obtain from a parametric or semiparametric statistical modeling approach, such as the GEE approach specified in Section 4.6. Such parameter estimates are often helpful for understanding underlying associations between predictors and outcomes that generate further scientific hypotheses.

In Section 4.7, each method performed well in the alternating recurrent event setting, but the LSTM approaches truly shined as individual history data grew over time and was incorporated into these algorithms. If not for the parameter estimate information related to history covariates seen in GEE algorithm settings (where linear terms of history covariates were consistently the most statistically significant in Tables C.1 and C.2), the reason for such outstanding performance for LSTM methods would remain completely opaque. As it stands, the true functional form of these history covariates within the LSTM algorithm still remains unclear. While there have been strides made in extracting relative importance of covariates in deep learning paradigms (e.g., Shapley values [Lundberg and Lee, 2017], or the field of interpretable machine learning), no off-the-shelf, easily programmable method exists for our proposed LSTM.

On a more practical note, the proposed analysis methods are agnostic to the definitions of primary and secondary recurrent events. For example, in Section 4.7, we modeled the primary event type, corresponding to $\mu^{(1)}$; however, analysis of the secondary event type may be of equal import. In that case, an analyst may exchange the definitions of primary and secondary events, and obtain two dynamic online prediction models, one for each type.

In conducting simulations, we often observed that small differences in network architecture could affect results substantially (data not shown). If the goal is to minimize an objective function, neural networks tend to obtain a local minimum that produces high levels of discrimination. In simulation, this may coincide with some bias in the estimated predictions. In statistical methods literature, such a bias-precision tradeoff is carefully mapped out through statistical proofs of consistency, identifiability and limiting behavior before popular use commences. The merging of computing and statistical literature for prediction is growing quickly, with technical arguments on why machine learning methods work well lagging somewhat behind. Content experts in both fields are aware that hyperparameter selection is a key concern in such algorithms. In our early simulation attempts, we saw firsthand that LSTM methods are not robust to hyperparameter choices made at random; the identification

of hyperparameters from the testing set is key to a well-behaved LSTM algorithm. Further work in understanding exactly how slight changes of hyperparameters affect network performance in our setting requires an ablation study [Meyes et al., 2019], a direction we hope to continue working towards. We hope that our paper also helps demystify computational jargon so that more traditional statistical researchers can enter this promising field.

CHAPTER 5

Conclusion

In this thesis, we presented novel approaches for dynamic prediction of recurrent events and alternating recurrent events. Each chapter makes use of survival analysis tools, as well as more novel and opaque machine learning algorithms, placing this work at the interstices of statistics and applications of novel computer science algorithms. While healthcare and biomedical applications motivated our methods, (alternating) recurrent events appear is many disciplines, meaning that analysts working in multiple domains may apply our methods.

Chapter 2 proposed a novel random forest based prediction tool, `RFRE.PO`, for recurrent events, where the recovery state is assumed to be negligible. History covariates centered in our approach, as a way to accrue subject-level information useful for prediction. While these highly important covariates may not always be available for an analyst at the start of follow-up, Chapter 2 proposes a method for imputing baseline history measures, with values updated as follow-up commences. These covariates provide crucial information even for methods that do not make use of the highly flexible random forest algorithm. The `RFRE.PO` algorithm proposed in Chapter 2 lacks the clear interpretations of associations between covariates and outcomes that an analyst obtains when fitting a semiparametric model; in an attempt to provide comparable associations, we provide results from permutation based importance scores, a Wald-type estimate. Our permutation based important scores may not be appropriate for interpretations of treatment effect, particularly in a random forest, where effect may be adjusted away based on post-treatment measures; yet where prediction is the goal, permutation importance metrics squeeze some interpretability from the `RFRE.PO` method.

Chapter 3 introduced a wrinkle to the data structure in Chapter 2. Rather than assuming negligible recovery time (e.g. COPD patients discharged on the same day as a breathing emergency), Chapter 3 allowed for extended periods of recovery. We showed that if an analyst fails to account for periods of removal from risk population, that is, the secondary

event state, a variant of length sampling bias creeps in to parameter estimates. Dynamic prediction centers here not on prediction of the restricted mean (the estimand of interest), but rather a debiasing of the parameter estimates, a natural use of the type of random forest models proposed in Chapter 2. In estimating the IPW weights, we are not concerned with interpretability of estimates of $\widehat{P}\{R_i(t) = 1\}$, rather, simple consistency of the IPW model is key. As such, covariates that may be perfectly predictive of $R_i(t)$ are permissible for use in the random forest IPW model. Specifically, we consider the ways that different patient history covariates, introduced in Chapter 2, aid in adjustment for informative missingness due to the second state. A highlight of Chapter 3 is the performance of reflective history. While both Chapters 2 and 3 rely on longitudinal random forests, there is little literature on hyperparameter selection for these algorithms, a gap in the literature for longitudinal data analysis.

Finally, in Chapter 4, we combined ideas from both Chapter 2 and 3. Like in Chapter 2, we were no longer concerned with interpretability of parameter estimates, but rather, in creating as powerful and accurate of a model as possible. However, our dynamic prediction model utilizes the weights regression concepts from Chapter 3, and unlike 2, Chapter 4 constructs an online model rather than a post-hoc prediction tool. Our model had outstanding predictive performance compared to a more traditional semiparametric model when applied to data from the intern health study. Future work may develop a predictive model in conjunction with the study team for deploying in future cohorts, ablation studies, development of permutation metrics for LSTMs, and a deeper dive into the role of the inverse probability weights.

Taken together, the work in this dissertation provides a new framework for analysis of recurrent and alternating recurrent event data. Scientific understanding of the disease of interest remains critical; in Chapter 2, the recurrent event (COPD exacerbations) was not subject to a secondary state, and rendering weighted pseudo-observations unnecessary. The decision to weight, then, remains a question of disease mechanism, and an understanding of the biological refractory period assumed. In Chapter 3 we saw a reduction in bias for all inverse probability weighted models, with relatively minimal additional computational burden. Throughout this dissertation, exact event-times were assumed. Component-wise censoring, common in epidemiological studies, provides a further statistical challenge, though work by Eaton et al. [2022] proposes a pseudo-observation for composite endpoints where at least one component is subject to interval censoring, a possible adaptations to handle various kinds of censoring over study time. Finally, censoring in each chapter was assumed to be independent of event times. While weights for alternating recurrent events and at-risk status may ameliorate violations of non-informative censoring, this idea remains untested.

Future work will develop methods appropriate for recurrent and alternating recurrent events subject to covariate dependent censoring.

APPENDIX A

Appendix for Chapter 2

A.1 Kaplan-Meier Estimation

Using the notation from Section 2.2, let $t \in \mathcal{T}$, and define:

$$N_t(\tau) = \sum_{i=1}^{n(t)} I[X_i(t) \leq \tau, \delta_i(t) = 1]$$
$$dN_t(u) = \sum_{i=1}^{n(t)} I[X_i(t) = u, \delta_i(t) = 1]$$
$$Y_t(u) = \sum_{i=1}^{n(t)} I[X_i(t) \geq u]$$

where $n(t)$ is the number of subjects available for follow-up at time point t. Briefly, $N_t(\tau)$ counts the number of events before horizon τ, $dN_t(u)$ counts the number of events at precisely u, and $Y_t(u)$ keeps track of the number of subjects at-risk for a recurrent event at time point u. Then the Kaplan-Meier estimator for probability of survival greater than τ at $t \in \mathcal{T}$ is defined as

$$\widehat{P}[T_i(t) \geq \tau] = \prod_{i:X_i(t) \leq \tau} \Big\{1 - \frac{dN_t[X_i(t)]}{Y_t[X_i(t)]}\Big\}.$$

This probability can be estimated for each $t \in \mathcal{T}$ using the quantities calculated in Section 2.2.

A.2 Model Calibration via Mean Squared Error

A calibration estimate of the m-XMT predictions averaged over all individuals, $i = 1, \ldots, N$ and follow-up windows within individual, $t_0, \ldots, t_{b_i}$ becomes

$$MSE = \frac{1}{\sum_{i=1}^{N}(b_i + 1)} \sum_{i=1}^{N} \sum_{j=0}^{b_i} [\widehat{S}_i^{\tau}(t_j) - \widehat{P}\{T_i(t_j) \geq \tau | \mathbf{Z}_i(t_j)\}]^2.$$

To obtain a similar calibration estimate for `RFRE.PO`, we replace the fitted values from the m-XMT model, $\widehat{P}\{T_i(t_j) \geq \tau | \mathbf{Z}_(t_j)\}$, with the predictions from the `RFRE.PO` method, $\widetilde{P}\{T_i(t_j) \geq \tau | \mathbf{Z}_i(t_j)\}$

A time-dependent MSE estimate pertaining to m-XMT predictions for the follow-up window starting at $t \in \mathcal{T}$ is:

$$MSE(t) = \frac{1}{n(t)} \sum_{i=1}^{n(t)} [\widehat{S}_i^{\tau}(t) - \widehat{P}\{T_i(t) \geq \tau | \mathbf{Z}_i(t)\}]^2.$$

Similarly, we can calculate $MSE(t)$ for `RFRE.PO` by replacing the predictions from an m-XMT model with the predictions from the `RFRE.PO` method, as described in Section 2.5.

A.3 Additional Simulation Results: MSE

Figure 2.4 and Table A.2 summarize MSE and $MSE(t), t \in \mathcal{T}$, simulation results for the `RFRE.PO` algorithm, the m-XMT (Main Effects) model, and the m-XMT (True) model. In general, the m-XMT (Main Effects) model has the highest MSE {Panel (A)}. In settings with correlation, the inclusion of history covariates decreases MSE for the m-XMT (Main Effects) model and the `RFRE.PO` algorithm. In all simulation settings, the `RFRE.PO` method outperforms the m-XMT (Main Effects) method, and upon the inclusion of history covariates, competes with the m-XMT (True) model in settings with highest correlation and lowest censoring. Examination of $MSE(t)$ reveals a similar stabilization as the stabilization observed in the C-statistic (Figure 2.4).

A.4 Additional Simulation Results: Power and Type I Error

Tables A.3 and A.4 display the simulated proportion of times predictors were deemed statistically important in dynamic prediction methods using the `RFRE.PO` algorithm (based on the random forest permutation test) and the m-XMT (Main Effects) model (based on the Wald test), respectively. Columns Z_5 and Z_7 of these tables show low rates of deeming noise covariates as important predictors with more conservative rates seen using the `RFRE.PO` algorithm than the m-XMT (Main Effects) model. For the m-XMT (Main Effects) model, these rates reflect type I error rates for inclusion of these predictors in the model that are reasonably close to the nominal 0.05 type I level subject to simulation variability. For the `RFRE.PO` algorithm, these rates do not reflect the rate at which the predictors were used in the algorithm, but rather the rate at which these variables were significantly improving

prediction outputs when included as an input in the algorithm.

Similarly, columns Z_1, Z_2, Z_3, Z_4, and Z_6 of these tables show the rate at which covariates associated with recurrent events in the m-XMT (True) model were deemed important in the `RFRE.PO` algorithm and the m-XMT (Main Effects) model in settings where (1) no history covariates were used, (2) $H_p^{(1)}(t)$ was included in the model or (3) $H_f^{(1)}(t)$ was used in the model. For the m-XMT (Main Effects) model, these rates correspond to power for detecting the association of these predictors in these 3 covariate history settings. For the `RFRE.PO` algorithm, these rates reflect the rates at which these variables were significantly improving prediction outputs when included as an input in the algorithm in these three history covariate settings.

When no history covariates are included as algorithm inputs, the `RFRE.PO` algorithm is more likely to deem these predictors as significantly helpful in making dynamic predictions when compared to the m-XMT (Main Effects) model. In settings where there is positive correlation between recurrent events and history covariates are included as inputs in the prediction algorithms, the relative importance of these history covariates tends to diminish rates of importance seen for Z_1, Z_2, Z_3, Z_4, Z_6 in Tables A.3 and A.4.

We were initially puzzled when looking at the importance of the covariate, t, in these tables, as t was not directly used in generating data via the m-XMT (True) model. Yet, the covariate, t, tends to show high importance in models with covariate history variables and correlation between recurrent event times. The explanation seems to lie with Panel B of Figure 2.3, where the C-statistic improves over the early follow-up windows when covariate history variables are used. This result seems to indicate an interaction between t and history variables in settings with correlated recurrent events that were not explicitly modeled in the data generation stage.

A.5 Simulation Study: Additional Tables and Figures

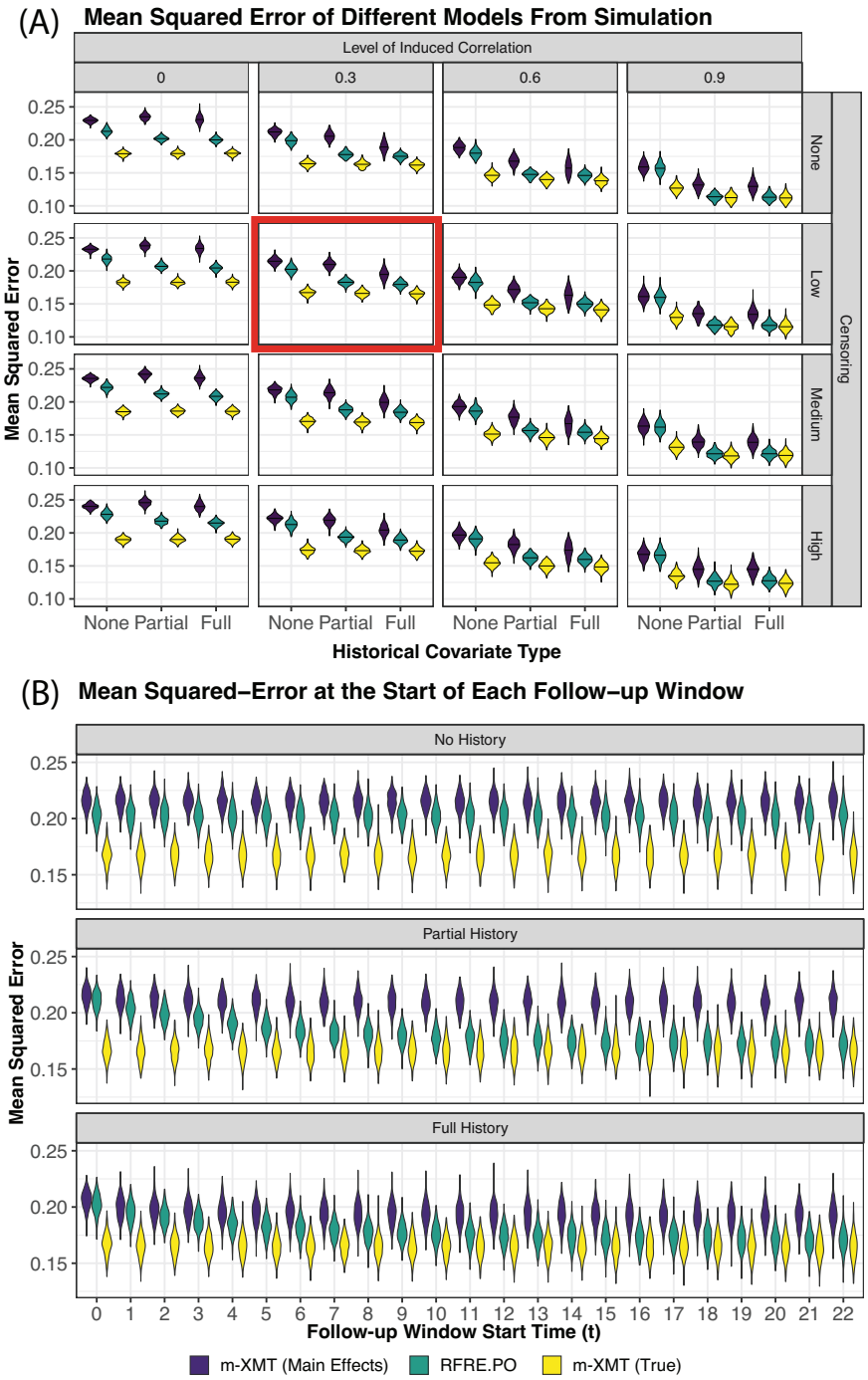


Figure A.1: Panel (A) displays the mean squared error (MSE) for simulation settings displayed in Figure 2.3. Censoring increases from top to bottom, and the amount of induced correlation increases from left to right. Within a box, the quality of historical information increases from left to right. Panel (B) shows the MSE at the start of each follow-up window for the setting boxed in red in Panel (A). From top to bottom, more historical information is added in Panel (B), and the MSE corresponding to `RFRE.PO` in $H_p^{(1)}(t)$ settings. 500 replicates were performed, each with $n = 500$ subjects.

Table A.1: Simulated C-statistics listed as Mean (ESD) for varying prediction algorithms when Z_1, Z_2 are correlated with each other, censoring percentages (none is 0%, light is 23%, moderate is 45%, and heavy is 63%), and gap-time correlations ($\rho = 0$, 0.3, 0.6, 0.9). 500 replicates were performed with $n = 500$ for each replicate, and the same data is reported in Figure 2.3.

	$\rho = 0$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.9$
No Censoring				
m-XMT (Main Effects)	0.534 (0.009)	0.533 (0.010)	0.533 (0.010)	0.531 (0.011)
`RFRE.PO`	0.605 (0.012)	0.596 (0.014)	0.584 (0.015)	0.563 (0.016)
m-XMT (True)	0.677 (0.006)	0.673 (0.007)	0.665 (0.008)	0.651 (0.010)
m-XMT (Main Effects) with $H_p^{(1)}(t)$	0.495 (0.021)	0.581 (0.018)	0.668 (0.012)	0.730 (0.008)
`RFRE.PO` with $H_p^{(1)}(t)$	0.633 (0.007)	0.653 (0.007)	0.685 (0.006)	0.712 (0.006)
m-XMT (True) with $H_p^{(1)}(t)$	0.677 (0.007)	0.674 (0.007)	0.685 (0.007)	0.711 (0.007)
m-XMT (Main Effects) with $H_f^{(1)}(t)$	0.522 (0.035)	0.648 (0.017)	0.687 (0.017)	0.732 (0.010)
`RFRE.PO` with $H_f^{(1)}(t)$	0.636 (0.008)	0.659 (0.008)	0.688 (0.008)	0.711 (0.007)
m-XMT (True) with $H_f^{(1)}(t)$	0.677 (0.006)	0.678 (0.007)	0.692 (0.010)	0.711 (0.010)
Light Censoring				
m-XMT (Main Effects)	0.533 (0.009)	0.533 (0.009)	0.534 (0.010)	0.532 (0.011)
`RFRE.PO`	0.601 (0.015)	0.593 (0.014)	0.580 (0.017)	0.562 (0.017)
m-XMT (True)	0.677 (0.007)	0.672 (0.008)	0.664 (0.009)	0.650 (0.011)
m-XMT (Main Effects) with $H_p^{(1)}(t)$	0.496 (0.023)	0.574 (0.018)	0.668 (0.012)	0.724 (0.009)
`RFRE.PO` with $H_p^{(1)}(t)$	0.628 (0.008)	0.649 (0.007)	0.681 (0.007)	0.708 (0.006)
m-XMT (True) with $H_p^{(1)}(t)$	0.676 (0.006)	0.674 (0.007)	0.682 (0.008)	0.705 (0.008)
m-XMT (Main Effects) with $H_f^{(1)}(t)$	0.526 (0.032)	0.643 (0.017)	0.681 (0.019)	0.725 (0.011)
`RFRE.PO` with $H_f^{(1)}(t)$	0.634 (0.008)	0.656 (0.008)	0.685 (0.008)	0.708 (0.008)
m-XMT (True) with $H_f^{(1)}(t)$	0.677 (0.007)	0.676 (0.008)	0.689 (0.011)	0.706 (0.011)
Moderate Censoring				
m-XMT (Main Effects)	0.534 (0.010)	0.535 (0.010)	0.535 (0.011)	0.532 (0.012)
`RFRE.PO`	0.597 (0.015)	0.590 (0.015)	0.577 (0.016)	0.558 (0.017)
m-XMT (True)	0.677 (0.007)	0.672 (0.008)	0.664 (0.009)	0.649 (0.012)
m-XMT (Main Effects) with $H_p^{(1)}(t)$	0.495 (0.022)	0.564 (0.018)	0.661 (0.013)	0.716 (0.011)
`RFRE.PO` with $H_p^{(1)}(t)$	0.625 (0.009)	0.646 (0.009)	0.677 (0.008)	0.704 (0.007)
m-XMT (True) with $H_p^{(1)}(t)$	0.677 (0.007)	0.674 (0.008)	0.681 (0.008)	0.700 (0.008)
m-XMT (Main Effects) with $H_f^{(1)}(t)$	0.529 (0.030)	0.638 (0.017)	0.676 (0.022)	0.719 (0.013)
`RFRE.PO` with $H_f^{(1)}(t)$	0.632 (0.009)	0.654 (0.008)	0.682 (0.009)	0.705 (0.008)
m-XMT (True) with $H_f^{(1)}(t)$	0.676 (0.007)	0.676 (0.008)	0.688 (0.011)	0.700 (0.010)
Heavy Censoring				
m-XMT (Main Effects)	0.535 (0.010)	0.535 (0.011)	0.535 (0.012)	0.532 (0.012)
`RFRE.PO`	0.592 (0.015)	0.585 (0.016)	0.569 (0.017)	0.551 (0.017)
m-XMT (True)	0.677 (0.008)	0.673 (0.008)	0.664 (0.010)	0.649 (0.012)
m-XMT (Main Effects) with $H_p^{(1)}(t)$	0.498 (0.022)	0.557 (0.018)	0.641 (0.015)	0.707 (0.014)
`RFRE.PO` with $H_p^{(1)}(t)$	0.621 (0.009)	0.642 (0.008)	0.672 (0.008)	0.701 (0.007)
m-XMT (True) with $H_p^{(1)}(t)$	0.677 (0.008)	0.673 (0.008)	0.678 (0.009)	0.695 (0.009)
m-XMT (Main Effects) with $H_f^{(1)}(t)$	0.535 (0.031)	0.635 (0.019)	0.669 (0.025)	0.709 (0.017)
`RFRE.PO` with $H_f^{(1)}(t)$	0.628 (0.010)	0.652 (0.009)	0.678 (0.010)	0.701 (0.009)
m-XMT (True) with $H_f^{(1)}(t)$	0.676 (0.008)	0.676 (0.008)	0.685 (0.011)	0.695 (0.012)

Table A.2: Simulated MSE listed as Mean (ESD) for varying prediction algorithms considered in Section 2.7, with censoring percentages (None, Light, Moderate, and Heavy), and gap-time correlations ($\rho = 0$, 0.3, 0.6, 0.9). 500 replicates were performed with $n = 500$ for each replicate.

	$\rho = 0$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.9$
		No Censoring		
m-XMT (Main Effects)	0.229 (0.003)	0.212 (0.005)	0.188 (0.006)	0.159 (0.008)
`RFRE.PO`	0.213 (0.005)	0.199 (0.005)	0.180 (0.006)	0.157 (0.008)
m-XMT (True)	0.179 (0.003)	0.164 (0.004)	0.146 (0.005)	0.127 (0.006)
m-XMT (Main Effects) with $H_p^{(1)}(t)$	0.235 (0.005)	0.206 (0.006)	0.168 (0.007)	0.131 (0.008)
`RFRE.PO` with $H_p^{(1)}(t)$	0.202 (0.003)	0.178 (0.005)	0.147 (0.005)	0.114 (0.005)
m-XMT (True) with $H_p^{(1)}(t)$	0.179 (0.004)	0.163 (0.005)	0.140 (0.005)	0.112 (0.006)
m-XMT (Main Effects) with $H_f^{(1)}(t)$	0.231 (0.007)	0.189 (0.009)	0.157 (0.011)	0.130 (0.009)
`RFRE.PO` with $H_f^{(1)}(t)$	0.200 (0.004)	0.175 (0.004)	0.146 (0.005)	0.113 (0.005)
m-XMT (True) with $H_f^{(1)}(t)$	0.180 (0.004)	0.162 (0.004)	0.138 (0.005)	0.112 (0.006)
		Light Censoring		
m-XMT (Main Effects)	0.233 (0.004)	0.215 (0.005)	0.190 (0.006)	0.161 (0.008)
`RFRE.PO`	0.218 (0.005)	0.203 (0.006)	0.183 (0.007)	0.160 (0.008)
m-XMT (True)	0.182 (0.004)	0.167 (0.005)	0.148 (0.006)	0.129 (0.007)
m-XMT (Main Effects) with $H_p^{(1)}(t)$	0.238 (0.005)	0.210 (0.006)	0.172 (0.007)	0.135 (0.008)
`RFRE.PO` with $H_p^{(1)}(t)$	0.207 (0.004)	0.183 (0.005)	0.152 (0.006)	0.118 (0.006)
m-XMT (True) with $H_p^{(1)}(t)$	0.182 (0.004)	0.166 (0.005)	0.143 (0.006)	0.115 (0.006)
m-XMT (Main Effects) with $H_f^{(1)}(t)$	0.234 (0.006)	0.195 (0.008)	0.163 (0.011)	0.134 (0.009)
`RFRE.PO` with $H_f^{(1)}(t)$	0.204 (0.004)	0.180 (0.005)	0.150 (0.006)	0.118 (0.006)
m-XMT (True) with $H_f^{(1)}(t)$	0.183 (0.004)	0.166 (0.005)	0.141 (0.006)	0.116 (0.007)
		Moderate Censoring		
m-XMT (Main Effects)	0.236 (0.004)	0.218 (0.005)	0.193 (0.007)	0.163 (0.008)
`RFRE.PO`	0.222 (0.005)	0.207 (0.006)	0.186 (0.007)	0.162 (0.008)
m-XMT (True)	0.185 (0.004)	0.170 (0.005)	0.151 (0.006)	0.131 (0.007)
m-XMT (Main Effects) with $H_p^{(1)}(t)$	0.242 (0.005)	0.214 (0.006)	0.172 (0.007)	0.139 (0.008)
`RFRE.PO` with $H_p^{(1)}(t)$	0.212 (0.004)	0.188 (0.005)	0.156 (0.006)	0.122 (0.006)
m-XMT (True) with $H_p^{(1)}(t)$	0.186 (0.004)	0.169 (0.005)	0.146 (0.006)	0.119 (0.006)
m-XMT (Main Effects) with $H_f^{(1)}(t)$	0.237 (0.006)	0.200 (0.008)	0.169 (0.012)	0.139 (0.009)
`RFRE.PO` with $H_f^{(1)}(t)$	0.209 (0.004)	0.184 (0.005)	0.154 (0.006)	0.122 (0.006)
m-XMT (True) with $H_f^{(1)}(t)$	0.185 (0.004)	0.168 (0.005)	0.144 (0.006)	0.119 (0.007)
		Heavy Censoring		
m-XMT (Main Effects)	0.240 (0.004)	0.221 (0.005)	0.197 (0.008)	0.167 (0.009)
`RFRE.PO`	0.228 (0.005)	0.212 (0.006)	0.191 (0.008)	0.166 (0.009)
m-XMT (True)	0.190 (0.004)	0.173 (0.005)	0.155 (0.006)	0.135 (0.007)
m-XMT (Main Effects) with $H_p^{(1)}(t)$	0.246 (0.005)	0.219 (0.006)	0.182 (0.008)	0.145 (0.009)
`RFRE.PO` with $H_p^{(1)}(t)$	0.218 (0.004)	0.194 (0.005)	0.162 (0.006)	0.127 (0.007)
m-XMT (True) with $H_p^{(1)}(t)$	0.190 (0.004)	0.174 (0.005)	0.151 (0.006)	0.123 (0.007)
m-XMT (Main Effects) with $H_f^{(1)}(t)$	0.240 (0.006)	0.205 (0.008)	0.173 (0.011)	0.145 (0.01)
`RFRE.PO` with $H_f^{(1)}(t)$	0.215 (0.004)	0.189 (0.006)	0.160 (0.007)	0.128 (0.007)
m-XMT (True) with $H_f^{(1)}(t)$	0.190 (0.004)	0.172 (0.006)	0.148 (0.007)	0.124 (0.008)

Table A.3: Simulated proportion of times that the corresponding covariate appeared as statistically significant according to the Wald test for the m-XMT (Main Effects) method. History covariates considered were of the type τ^*-restricted mean survival time in previous windows, where τ^* is $\frac{1}{12}$ years (described in further detail in Section 2.2.1). In the $H_p^{(1)}(t)$ case, results were from $m = 10$ multiply imputed datasets combined via Rubin's rules. 500 replicates were performed with $n = 500$ for each replicate. Covariates with signal (Z_1, Z_2, Z_3, Z_4, Z_6) are displayed first, and noise covariates are shown second (Z_5, Z_7). Within a censoring and correlation designation, quality of historical information included in the model increases, from top to bottom.

		Signal					Noise		History Covariates		
Censoring	ρ	Z_1	Z_2	Z_3	Z_4	Z_6	Z_5	Z_7	t	$H_p^{(1)}(t)$	$H_f^{(1)}(t)$
None	0.0	0.642	0.370	0.078	0.056	0.380	0.054	0.066	0.050	–	–
None	0.0	0.641	0.293	0.050	0.077	0.378	0.054	0.085	0.046	0.402	–
None	0.0	0.646	0.372	0.078	0.060	0.380	0.060	0.066	0.046	–	0.346
None	0.3	0.624	0.230	0.082	0.044	0.388	0.060	0.056	0.064	–	–
None	0.3	0.635	0.232	0.076	0.057	0.374	0.052	0.052	0.071	0.431	–
None	0.3	0.528	0.222	0.074	0.050	0.248	0.054	0.040	0.358	–	0.998
None	0.6	0.662	0.200	0.062	0.056	0.460	0.070	0.072	0.044	–	–
None	0.6	0.495	0.168	0.060	0.071	0.304	0.043	0.092	0.522	1.000	–
None	0.6	0.470	0.172	0.062	0.054	0.190	0.092	0.072	0.672	–	1.000
None	0.9	0.592	0.134	0.098	0.082	0.474	0.066	0.066	0.066	–	–
None	0.9	0.381	0.165	0.072	0.036	0.137	0.065	0.065	0.986	1.000	–
None	0.9	0.336	0.122	0.070	0.072	0.148	0.066	0.050	0.772	–	1.000
Low	0.0	0.573	0.244	0.074	0.046	0.299	0.072	0.044	0.062	–	–
Low	0.0	0.598	0.274	0.074	0.095	0.385	0.054	0.051	0.071	0.439	–
Low	0.0	0.581	0.246	0.082	0.044	0.303	0.066	0.046	0.066	–	0.289
Low	0.3	0.592	0.206	0.072	0.060	0.398	0.054	0.040	0.052	–	–
Low	0.3	0.546	0.272	0.043	0.060	0.301	0.073	0.043	0.066	0.222	–
Low	0.3	0.504	0.192	0.068	0.062	0.238	0.060	0.036	0.230	–	0.998
Low	0.6	0.636	0.212	0.086	0.060	0.478	0.052	0.066	0.056	–	–
Low	0.6	0.487	0.170	0.057	0.073	0.243	0.053	0.097	0.413	1.000	–
Low	0.6	0.474	0.192	0.072	0.074	0.226	0.068	0.070	0.528	–	1.000
Low	0.9	0.582	0.168	0.096	0.094	0.500	0.054	0.064	0.052	–	–
Low	0.9	0.351	0.120	0.060	0.043	0.187	0.060	0.060	0.967	1.000	–
Low	0.9	0.338	0.150	0.070	0.076	0.176	0.064	0.078	0.582	–	1.000
Medium	0.0	0.618	0.250	0.068	0.060	0.358	0.094	0.088	0.080	–	–
Medium	0.0	0.580	0.279	0.061	0.065	0.370	0.059	0.057	0.055	0.475	–
Medium	0.0	0.622	0.248	0.066	0.064	0.370	0.094	0.084	0.084	–	0.236
Medium	0.3	0.624	0.250	0.072	0.074	0.372	0.066	0.060	0.066	–	–
Medium	0.3	0.599	0.246	0.068	0.054	0.393	0.086	0.052	0.072	0.132	–
Medium	0.3	0.508	0.230	0.080	0.080	0.234	0.070	0.064	0.172	–	0.994

Continued on next page

Table A.3: (continued)

Censoring	ρ	Signal					Noise		History Covariates		
		Z_1	Z_2	Z_3	Z_4	Z_6	Z_5	Z_7	t	$H_p^{(1)}(t)$	$H_f^{(1)}(t)$
Medium	0.6	0.620	0.160	0.078	0.074	0.464	0.060	0.058	0.060	–	–
Medium	0.6	0.485	0.190	0.080	0.052	0.251	0.064	0.052	0.289	1.000	–
Medium	0.6	0.442	0.166	0.074	0.070	0.236	0.072	0.054	0.410	–	1.000
Medium	0.9	0.582	0.182	0.076	0.062	0.488	0.042	0.062	0.052	–	–
Medium	0.9	0.412	0.151	0.050	0.072	0.207	0.062	0.054	0.922	1.000	–
Medium	0.9	0.376	0.176	0.068	0.064	0.190	0.042	0.064	0.466	–	1.000
High	0.0	0.580	0.254	0.060	0.066	0.322	0.048	0.054	0.058	–	–
High	0.0	0.584	0.263	0.066	0.078	0.400	0.056	0.068	0.066	0.436	–
High	0.0	0.584	0.246	0.062	0.064	0.322	0.048	0.058	0.056	–	0.244
High	0.3	0.562	0.260	0.080	0.078	0.358	0.064	0.076	0.084	–	–
High	0.3	0.583	0.228	0.070	0.060	0.315	0.040	0.114	0.068	0.044	–
High	0.3	0.462	0.228	0.078	0.078	0.238	0.058	0.078	0.146	–	0.988
High	0.6	0.570	0.198	0.074	0.076	0.424	0.088	0.048	0.068	–	–
High	0.6	0.536	0.216	0.074	0.064	0.304	0.056	0.058	0.198	0.984	–
High	0.6	0.428	0.174	0.082	0.074	0.208	0.064	0.044	0.302	–	1.000
High	0.9	0.570	0.160	0.088	0.076	0.460	0.064	0.066	0.068	–	–
High	0.9	0.433	0.122	0.064	0.068	0.230	0.066	0.034	0.770	1.000	–
High	0.9	0.390	0.154	0.076	0.082	0.226	0.058	0.070	0.276	–	1.000

Table A.4: Simulated proportion of times that the corresponding covariate appeared as statistically significant according to the random forest permutation test. 500 replicates were performed, with $n = 500$ subjects in each replicate. History covariates considered were of the type τ^*-restricted mean survival time in previous windows, where τ^* is $\frac{1}{12}$ years (described in further detail in Section 2.2.1). In the $H_p^{(1)}(t)$ case, results were from $m = 10$ multiply imputed datasets, combined via Rubin's rules. Covariates with signal (Z_1, Z_2, Z_3, Z_4, and Z_6) are displayed first, and noise covariates are shown second (Z_5, Z_7). Within a censoring and correlation designation, quality of historical information included in the model increases, from top to bottom.

		Signal					Noise		History Covariates		
Censoring	ρ	Z_1	Z_2	Z_3	Z_4	Z_6	Z_5	Z_7	t	$H_p^{(1)}(t)$	$H_f^{(1)}(t)$
None	0.00	1.000	1.000	0.954	0.066	1.000	0.050	0.050	0.062	–	–
None	0.0	1.000	0.552	0.008	0.000	1.000	0.000	0.000	0.538	1.000	–
None	0.00	1.000	0.988	0.866	0.050	1.000	0.038	0.036	0.992	–	1.000
None	0.3	1.000	0.994	0.838	0.052	1.000	0.044	0.034	0.046	–	–
None	0.3	1.000	0.332	0.006	0.002	1.000	0.000	0.002	0.994	1.000	–
None	0.3	1.000	0.942	0.578	0.032	1.000	0.044	0.038	0.996	–	1.00
None	0.6	1.000	0.954	0.610	0.048	1.000	0.038	0.042	0.046	–	–
None	0.6	1.000	0.223	0.010	0.026	1.000	0.006	0.000	1.000	1.000	–
None	0.6	1.000	0.808	0.356	0.032	1.000	0.036	0.036	1.000	–	1.00
None	0.9	1.000	0.792	0.334	0.032	1.000	0.032	0.034	0.046	–	–
None	0.9	1.000	0.123	0.012	0.127	0.998	0.002	0.006	1.000	1.000	–
None	0.9	1.000	0.488	0.136	0.034	1.000	0.034	0.026	1.000	–	1.00
Low	0.0	1.000	1.000	0.890	0.044	1.000	0.020	0.030	0.022	–	–
Low	0.0	1.000	0.413	0.004	0.000	1.000	0.000	0.000	0.451	1.000	–
Low	0.0	1.000	0.970	0.739	0.038	1.000	0.014	0.026	0.824	–	1.00
Low	0.3	1.000	0.982	0.780	0.032	1.000	0.038	0.044	0.038	–	–
Low	0.3	1.000	0.280	0.008	0.002	1.000	0.000	0.000	0.976	1.000	–
Low	0.3	1.000	0.884	0.476	0.028	1.000	0.028	0.030	0.840	–	1.00
Low	0.6	1.000	0.938	0.506	0.018	1.000	0.016	0.040	0.042	–	–
Low	0.6	1.000	0.172	0.010	0.024	1.000	0.002	0.002	1.000	1.000	–
Low	0.6	1.000	0.712	0.260	0.022	1.000	0.016	0.040	0.948	–	1.00
Low	0.9	1.000	0.716	0.264	0.052	1.000	0.016	0.034	0.048	–	–
Low	0.9	0.996	0.118	0.006	0.104	1.000	0.008	0.024	1.000	1.000	–
Low	0.9	1.000	0.424	0.122	0.022	1.000	0.024	0.018	0.996	–	1.000
Medium	0.0	1.000	0.998	0.832	0.032	1.000	0.032	0.030	0.042	–	–
Medium	0.0	1.000	0.303	0.004	0.000	1.000	0.000	0.000	0.403	1.000	–
Medium	0.0	1.000	0.926	0.618	0.034	1.000	0.042	0.026	0.696	–	1.00
Medium	0.3	1.000	0.982	0.694	0.038	1.000	0.026	0.030	0.070	–	–
Medium	0.3	1.000	0.222	0.010	0.006	1.000	0.000	0.000	0.967	1.000	–
Medium	0.3	1.000	0.818	0.406	0.032	1.000	0.022	0.012	0.638	–	1.00

Continued on next page

Table A.4: (continued)

Censoring	ρ	Signal					Noise		History Covariates		
		Z_1	Z_2	Z_3	Z_4	Z_6	Z_5	Z_7	t	$H_p^{(1)}(t)$	$H_f^{(1)}(t)$
Medium	0.6	1.000	0.898	0.448	0.036	1.000	0.020	0.022	0.064	–	–
Medium	0.6	1.000	0.161	0.010	0.024	1.000	0.000	0.000	1.000	1.000	–
Medium	0.6	1.000	0.622	0.196	0.020	1.000	0.012	0.020	0.824	–	1.00
Medium	0.9	1.000	0.688	0.258	0.028	1.000	0.016	0.016	0.064	–	–
Medium	0.9	1.000	0.113	0.014	0.138	1.000	0.006	0.023	1.000	1.000	–
Medium	0.9	1.000	0.440	0.112	0.034	1.000	0.020	0.018	0.926	–	1.00
High	0.0	1.000	0.966	0.688	0.036	1.000	0.024	0.020	0.060	–	–
High	0.0	1.000	0.211	0.000	0.000	1.000	0.000	0.000	0.310	1.000	–
High	0.0	1.000	0.836	0.490	0.038	1.000	0.022	0.022	0.582	–	1.00
High	0.3	1.000	0.946	0.506	0.028	1.000	0.014	0.016	0.098	–	–
High	0.3	1.000	0.192	0.004	0.002	1.000	0.000	0.000	0.953	1.000	–
High	0.3	1.000	0.706	0.268	0.018	1.000	0.012	0.016	0.590	–	1.000
High	0.6	1.000	0.812	0.368	0.014	1.000	0.022	0.026	0.094	–	–
High	0.6	1.000	0.126	0.010	0.024	1.000	0.002	0.006	1.000	1.000	–
High	0.6	1.000	0.524	0.138	0.020	1.000	0.018	0.026	0.716	–	1.000
High	0.9	1.000	0.602	0.226	0.016	1.000	0.016	0.014	0.086	–	–
High	0.9	0.998	0.106	0.033	0.110	0.994	0.004	0.018	1.000	1.000	–
High	0.9	1.000	0.314	0.094	0.018	1.000	0.022	0.012	0.818	–	1.000

A.6 Tables and Figures for the Analysis of the Azithromycin for the Prevention of COPD Cohort

Table A.5: m-XMT parameter estimates, based on covariates selected by Xia et al. [2020], applied to censored longitudinal data with $\tau = 180$, $a = 30$.

Covariate	$e^{\widehat{\beta}\dagger}$	95% CI	p
Azithromycin (versus placebo)	1.43	(1.14, 1.79)	**<0.01**
Smoker at baseline (versus non-smoker)	1.07	(0.80, 1.43)	0.66
FEV_1 in liters at baseline	1.32	(1.05, 1.66)	**0.02**
Male gender (versus female)	1.36	(1.07, 1.74)	**0.01**
Age in decades	1.14	(0.99, 1.32)	0.07

† $e^{\widehat{\beta}}$ is the estimated odds ratio for being event-free during a 180-day follow-up period corresponding to a 1-unit increase in the (continuous) predictor, or the presence of a binary covariate, adjusted for other predictors.
FEV_1 is forced expiratory volume in one second, a clinical measure of lung function.

Table A.6: Parameter estimates from the m-XMT (Forward Selection) model based on the first of $m = 10$ imputed datasets. For the analysis, we created three history variables, $H_f(t), H_p^{(2)}(t)$ and $H_p^{(3)}(t)$. $H_f(t)$ was an indicator if the subject had experienced a severe exacerbation on study time prior to time t. $H_p^{(2)}(t)$ was assumed to take a categorical form based on whether the participant had an exacerbation in the past zero to 31 days ($H_p^{(2)}(t) = 4$), 31-92 days ($H_p^{(2)}(t) = 3$), 93-182 days ($H_p^{(2)}(t) = 2$), 183 days to 365 days ($H_p^{(2)}(t) = 1$), or finally, if that time was more than 365 days or never ($H_p^{(2)}(t) = 0$). $H_p^{(3)}(t)$ is the estimated 30-day event rate using all follow-up prior to time t. A very recent exacerbation, bronchitis, gender, treatment, the ratio of FEV to FVC %predicted, SGRQ symptoms, oxygen use, recent hospitalization, rate of exacerbations, and use of ICS medication remain significant across the $m = 10$ imputed datasets.

Covariate	$e^{\widehat{\beta}}$ †	95% CI	p
Categorical $H_{p,i}^{(2)}(t) = 0$ (versus 1, 2, 3 or 4)	2.32	(2.01, 2.68)	**<0.01**
Hospitalization in the year before $t = 0$ (versus not)	0.70	(0.53, 0.92)	**0.01**
Corticosteroid use in the year before $t = 0$ (versus no use)	0.61	(0.40, 0.94)	**0.03**
$H_{f,i}(t) = 1$ (versus $H_{f,i}(t) = 0$)	1.81	(1.23, 2.68)	**0.01**
Time t (in months)	0.96	(0.93, 1.00)	**0.02**
Male gender (versus female gender)	1.69	(1.28, 2.23)	**<0.01**
Race = Black (versus non-Black)	0.57	(0.37, 0.89)	**0.01**
Sudden feelings of panic at time t (versus no such feelings)	0.84	(0.74, 0.95)	**0.01**
Sudden feelings of "butterfly" fright at time t (versus no such feelings)	0.89	(0.78, 1.01)	0.08
Seen a mental health provider at time t (versus has not seen a provider)	0.75	(0.61, 0.93)	**0.01**
Has pain at time t (versus no pain)	0.92	(0.86, 0.98)	**0.03**
St. George's Respiratory Questionnaire symptoms score at time t	0.99	(0.99, 1.00)	**0.01**
Antiarrhythmic use at time t (versus no use)	0.07	(0.00, 6.56)	0.25
FEV_1/FVC % predicted at time t	1.01	(1.00, 1.02)	**<0.01**
Azithromycin treatment group (versus placebo)	0.65	(0.49, 0.85)	**<0.01**
Inhaled corticosteroid or long-acting beta-agonist use at time t	1.28	(1.03, 1.60)	**0.02**
Other baseline medications (versus no miscellaneous medications)	0.15	(0.03, 0.81)	**0.02**
Supplemental oxygen use at time t (versus no supplemental oxygen)	0.55	(0.41, 0.73)	**<0.01**
Mixed beta-blocker use at time t (versus selective or no beta-blocker use)	3.79	(1.08, 13.24)	**0.04**
Bronchiectasis diagnosis at time t (versus no such diagnosis)	0.64	(0.48, 0.85)	**<0.01**

†: $e^{\widehat{\beta}}$ is the estimated odds ratio for being event-free during a 180-day follow-up period corresponding to a 1-unit increase in the (continuous) predictor, or the presence of a binary covariate, adjusted for other predictors.

FEV_1 is forced expiratory volume in one second and FVC is forced vital capacity, clinical measures of lung function.

SGRQ is the Saint George's Respiratory Questionnaire, a patient reported outcome measuring quality of life.

Table A.7: P-values from the `RFRE.PO` permutation test applied to the out-of-bag-samples combined across $m = 10$ multiply imputed datasets. Included predictors had a statistically significant permutation test z-score for at least one of the multiply imputed training datasets. Additional predictors that did not achieve statistical significance via the permutation test, but were used in the `RFRE.PO` algorithm in some form, are not shown. For analysis, we created three history variables, $H_f(t), H_p^{(2)}(t)$ and $H_p^{(3)}(t)$. $H_f(t)$ was an indicator if the subject had experienced a severe exacerbation on study time prior to time t. $H_p^{(2)}(t)$ was assumed to take a categorical form based on whether the participant had an exacerbation in the past zero to 31 days ($H_p^{(2)}(t) = 4$), 31-92 days ($H_p^{(2)}(t) = 3$), 93-182 days ($H_p^{(2)}(t) = 2$), 183 days to 365 days ($H_p^{(2)}(t) = 1$), or finally, if that time was more than 365 days or never ($H_p^{(2)}(t) = 0$). $H_p^{(3)}(t)$ is the estimated 30-day event rate using all follow-up prior to time t.

Covariate	p
Categorical $H_{p,i}^{(2)}(t) = 0$ (versus Categorical $H_{p,i}^{(2)}(t) = 1, 2, 3, 4$)	0.06
Categorical $H_{p,i}^{(2)}(t) = 1$ (versus Categorical $H_{p,i}^{(2)}(t) = 0, 2, 3, 4$)	**<0.04**
Categorical $H_{p,i}^{(2)}(t) = 4$ (versus Categorical $H_{p,i}^{(2)}(t) = 0, 1, 2, 3$)	**<0.01**
Categorical $H_{p,i}^{(2)}(t)$	**<0.01**
$H_{p,i}^{(3)}(t)$	**<0.01**
$H_{f,i}(t)$	**<0.01**
Respiratory Hospitalization in the Year before Study Start	**0.01**
Corticosteroid Use in the Year before Study Start	**<0.01**
Gender	0.99
Smoking at Baseline	0.98
General Health Score at Time t	**<0.01**
Standardized General Health Z-Score at Time t	**<0.01**
Pain Score at Time t	**0.03**
Standardized Pain Score at Time t	**0.05**
SGRQ Symptom Score at Time t	**<0.01**
Seen a Mental Health Professional for Anxiety/Depression at Time t	1.00
Anti-Coagulant Use at Time t	1.00
Inhaled Corticosteroids, Long-Acting Muscarinic, and Long-Acting Bronchodilator Agent at Time t	0.99
Long-Acting Muscarinic Use at Time t	1.00
Treatment Group	**0.02**
Leukocytosis at Time t	**0.02**
Oxygen Use at Time t	1.00
Bronchiectasis Diagnosis at Time t	**<0.01**
Phlegm at Time t	1.00

FEV_1 is forced expiratory volume in one second and FVC is forced vital capacity in liters, clinical measures of lung function.

SGRQ is the Saint George's Respiratory Questionnaire, a patient reported outcome measuring quality of life.

Table A.8: Comparison of the C-statistics and mean-squared errors (MSE) of models fit to the validation set of the Azithromycin in the Prevention of COPD cohort. The Wald forward selection model and the model corresponding to clinical input from Xia et al. [2020] are fit using the modified XMT approach. The forward selection model includes predictors through automated forward selection requiring $p < 0.05$ for entry into the model. Missing values in the forward selection model and `RFRE.PO` for time since most recent exacerbation are multiply imputed using the coding scheme described in Section 2.8. C-statistic and mean-squared error estimates and standard errors for all methods and models are averaged across results from multiply imputed data with Rubin's rule applied to bootstrapped ($b = 100$) standard errors.

Model Name	C-Statistic		MSE	
	Mean	SD	Mean	SD
`RFRE.PO`	0.613	0.009	0.242	0.003
m-XMT (Clinical Input)	0.565	0.009	0.252	0.002
m-XMT (Forward Selection)	0.570	0.009	0.264	0.004

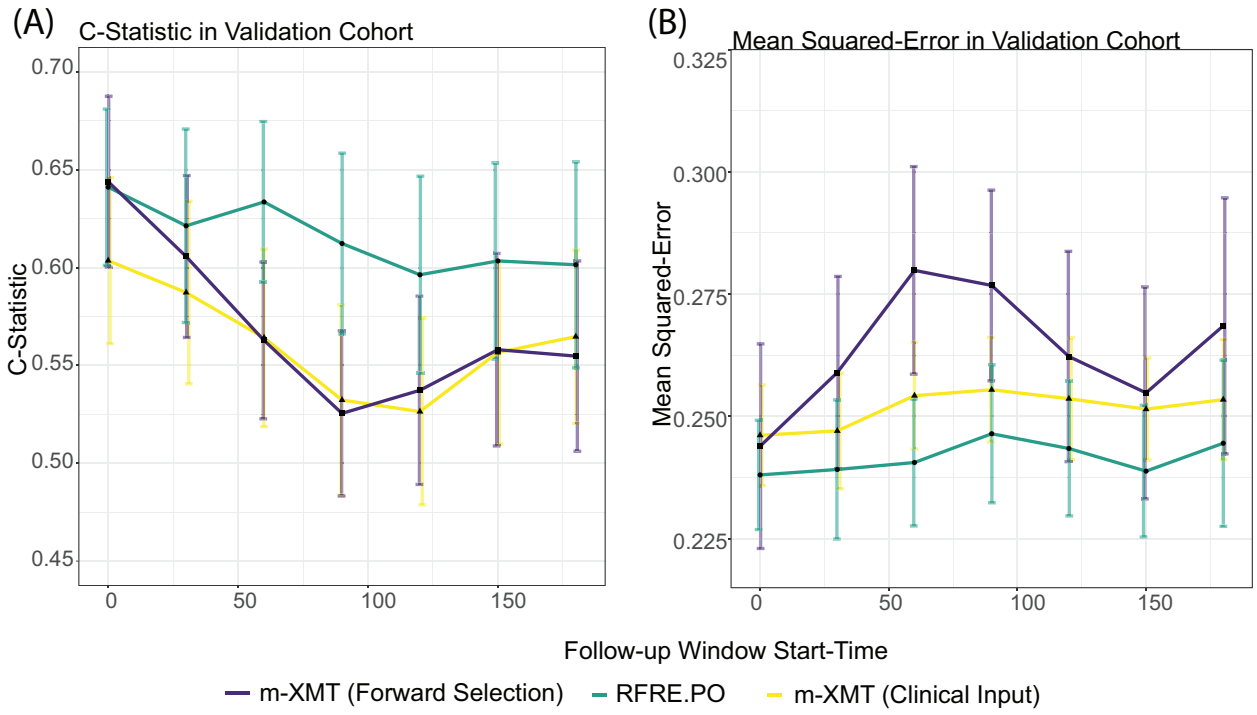


Figure A.2: A panel graphic displaying time-dependent C-statistics (Panel A) and MSEs (Panel B) in the validation dataset of the Azithromycin for the Prevention of COPD study, calculated at each follow-up window $t \in \{0, 30, 60, 90, 120, 150, 180\}$ for $\tau = 180$; 95% confidence intervals also displayed. Compared to the other algorithms, `RFRE.PO` has fairly stable and attractive performance metrics across the various follow-up windows, likely because interactions with the various predictors over time are naturally incorporated if useful for prediction.

APPENDIX B

Appendix for Chapter 3

B.1 Clinical Relevance of $T_i(t)$

Understanding how an exposure may increase or decrease time-to-primary event is a common question of clinical interest. Consider the scenario where $Z_i(t)$, dynamic predictors of alternating recurrent event behaviors contains an exposure of interest, and examine Figure B.1. Subjects 4, 5, and 6 are at-risk rarely, and thus, appear in the censored longitudinal data structure infrequently. This by itself is not a problem, as our censored longitudinal data structure is constructed to handle such missingness, though if an exposure of interest shortens time spent at-risk, these individuals will be missing from the data, and their shorter times-to-primary alternating recurrent event will not contribute information to the proposed regression model. This will produce sampling bias, and may obscure potentially significant associations between exposures and outcomes.

An analyst may also attempt a "naive" approach to modeling the censored longitudinal data structure, and model $T_i^{(1)}(t)$ rather than $T_i(t)$. One may notice that $T_i^{(1)}(t)$ is always defined, whereas $T_i(t)$ is subject to missingness. That is, for subjects with $R_i(t) = 1$, $T_i^{(1)}(t) = T_i(t)$ but for subjects with $R_i(t) = 0$, $T_i^{(1)}(t) = \{T^{(1)}_{i,\eta_i^{(1)}(t)} - T^{(2)}_{i,\eta_i^{(2)}(t)}\} + T_i^{(2)}(t)$. Thus, changes in $T_i^{(1)}(t)$ for $R_i(t) = 0$ are a function of both the primary and secondary types of events, which cannot be untangled in a GEE approach. To understand why this is a problem, consider the example of cancer patients with venous thromboembolisms, where the primary event is time-to-hospitalization and the secondary alternating recurrent event is time-to-discharge from hospital. Imagine a certain kind of chemotherapy, Chemo A, is associated with the same time-to-hospitalization as Chemo B, but an increase in duration of hospital stay. For subjects with $R_i(t) = 1$ and Chemo A, $T_i^{(1)}(t)$ is the same as $T_j^{(1)}(t)$ for those with Chemo B. However, for subjects with $R_i(t) = 0$ and Chemo A, $T_i^{(1)}(t)$ is actually greater than $T_j^{(1)}(t)$; this may lead to an analyst concluding that Chemo A, the therapy that increases hospitalization time, is a superior treatment, as it increases average $T_i^{(1)}(t)$.

However, the observed increase was due to an increase in the secondary alternating recurrent event, and in this situation is something that most clinicians would prefer to decrease.

B.2 Proof of Theorem 1

$$\mathbb{E}\left\{\frac{R_i(t)}{\pi_i(t)}PO_i^{\tau}(t)\middle|Z_i(t)\right\} = \mathbb{E}\left\{\frac{R_i(t)}{P[R_i(t)=1|W_i(t)]}PO_i^{\tau}(t)|Z_i(t)\right\}$$

But by conditioning on a finer partition, we obtain:

$$= \mathbb{E}\left\{\mathbb{E}\left\{\frac{R_i(t)}{P[R_i(t)=1|W_i(t)]}PO_i^{\tau}(t)|Z_i(t),W_i(t),R_i(t)\right\}|Z_i(t)\right\}$$

However, if we have conditioned on $W_i(t)$ and $R_i(t)$, then they may be treated as fixed in the inner expectation and factored, and we also apply Assumption 3:

$$= \mathbb{E}\left\{\frac{R_i(t)}{P[R_i(t)=1|W_i(t)]}\mathbb{E}\left\{PO_i^{\tau}(t)|Z_i(t),R_i(t)\right\}|Z_i(t)\right\}$$

Now, apply the definition of our pseudo-observation and expectation of a Bernoulli random variable to obtain

$$\begin{aligned}
&= \mathbb{E}\left\{\frac{R_i(t)}{P[R_i(t)=1|W_i(t)]}R_i(t)\mathbb{E}[\min\{T_i^{(1)}(t),\tau\}|R_i(t)=1,Z_i(t)]|Z_i(t)\right\} \\
&= \mathbb{E}\left\{\frac{R_i(t)}{P[R_i(t)=1|W_i(t)]}\mathbb{E}[\min\{T_i^{(1)}(t),\tau\}|R_i(t)=1,Z_i(t)]|Z_i(t)\right\}
\end{aligned}$$

However, if $R_i(t)=1$, then by definition $\min\{T_i^{(1)}(t),\tau\} = \min\{T_i(t),\tau\}$, and by Assumption 3, we can condition on $W_i(t)$:

$$\begin{aligned}
&= \mathbb{E}\left\{\frac{R_i(t)}{P[R_i(t)=1|W_i(t)]}\mathbb{E}[\min\{T_i(t),\tau\}|R_i(t)=1,Z_i(t),W_i(t)]|Z_i(t)\right\} \\
&= \mathbb{E}\left\{\mathbb{E}\left[\frac{R_i(t)}{P[R_i(t)=1|W_i(t)]}\min\{T_i(t),\tau\}|R_i(t)=1,Z_i(t),W_i(t)\right]|Z_i(t)\right\}
\end{aligned}$$

But marginalizing over $W_i(t)$ yields:

$$= \mathbb{E}\left\{\frac{R_i(t)}{P[R_i(t)=1|W_i(t)]}\min\{T_i(t),\tau\}|Z_i(t)\right\}$$

The reasoning now boils down to a more classic argument involving inverse probability weights.

$$= \mathbb{E}\left\{ \frac{R_i(t)}{P[R_i(t)=1|W_i(t)]} \min\{T_i(t),\tau\} | Z_i(t) \right\}$$
$$= \mathbb{E}\left\{ \mathbb{E}\left\{ \frac{R_i(t)}{P[R_i(t)=1|W_i(t)]} \min\{T_i(t),\tau\} | W_i(t), Z_i(t) \right\} | Z_i(t) \right\}$$

However, by conditioning on $W_i(t)$, we may again treat $P[R_i(t)=1|W_i(t)]$ as a constant:

$$= \mathbb{E}\left\{ \frac{1}{P[R_i(t)=1|W_i(t)]} \mathbb{E}\left\{ R_i(t) \min\{T_i(t),\tau\} | W_i(t), Z_i(t) \right\} | Z_i(t) \right\}$$

By Assumption 2:

$$= \mathbb{E}\left\{ \frac{1}{P[R_i(t)=1|W_i(t)]} \mathbb{E}\left\{ R_i(t) | W_i(t), Z_i(t) \right\} \mathbb{E}\left\{ \min\{T_i(t),\tau\} | W_i(t), Z_i(t) \right\} | Z_i(t) \right\}$$
$$= \mathbb{E}\left\{ \frac{1}{P[R_i(t)=1|W_i(t)]} P\left\{ R_i(t)=1 | W_i(t), Z_i(t) \right\} \mathbb{E}\left\{ \min\{T_i(t),\tau\} | W_i(t), Z_i(t) \right\} | Z_i(t) \right\}$$

As $W_i(t)$ is assumed to contain covariates $Z_i(t)$ (Assumption 2), the above line reduces to $\mathbb{E}\{\min\{T_i(t),\tau\}|Z_i(t)\}$, as desired.

B.3 History Covariates

When recurrent event times with negligible secondary alternating recurrent event length are highly correlated within an individual, Loe et al. [2025a] found that "history covariates," whether or not they are available for analysis at the beginning of follow-up, are some of the most important types of covariates included in $Z_i(t)$ for dynamic prediction of recurrent events. In this section we first develop history covariates for: (1) *full* history setting (those where $Z_i(t)$ is known for all $t \in \mathcal{T}$) and (2) *partial* history covariates (those where $Z_i(t)$ is not completely known at the start of follow-up), and develop a new type of history covariate for the prediction of $\widehat{R}_i(t)$. For all defined covariates, we first consider the full history covariate setting, where past subject behavior within t_h units of time prior to t_0 may be used in defining covariates, and then describe different cases of partial history information that a practitioner may encounter.

For subject i, let q_i count the number of primary events that occur in the time period from t_h to t_0. As subjects are assumed to be at-risk at $t_0 = 0$, the events that occur prior to t_0 may be denoted $\max(T^{(2)}_{i,-(q_i+1)}, t_h) < T^{(1)}_{i,-q_i} < T^{(2)}_{i,-q_i} < \ldots T^{(1)}_{i,-1} < T^{(2)}_{i,-1}$, where calendar

times for events happening prior to t_0 and t_h are less than zero. First, consider a dynamic predictor that estimates the mean observed time at-risk for a primary alternating recurrent event for subject i. This covariate may be written as

$$H_{f,1}^{(1)}(t) = \frac{1}{\sum_{j=-q} I(T_{i,j}^{(1)} < t)} \sum_{j=-q} \{T_{i,j}^{(1)} - \max(T_{i,(j-1)}^{(2)}, t_h)\} \times I(T_{i,j}^{(1)} < t). \tag{B.1}$$

This quantity is well defined for all $t \in \mathcal{T}$, and for later $t \in \mathcal{T}$, the quality of the information carried by $H_{f,1}^{(k)}$ increases due to an increased number of gap times used in averaging.

We also consider the most recent primary event time defined as

$$H_{f,2}^{(1)}(t) = t - \max(\{T_{i,j}^{(1)} | T_{i,j}^{(1)} < t, -q_i < j\}, t_h) \tag{B.2}$$

and the length of last period at-risk for the primary event:

$$H_{f,3}^{(1)}(t) = T_{i,\eta_i^{(1)}(t)-1}^{(1)} - T_{i,\eta_i^{(2)}(t)-1}^{(2)} \tag{B.3}$$

where t_h is used if either of these quantities does not exist. Finally, $q_i + \eta_i(t) - 1$ may be used as a historical covariate, counting the number of primary events occurring before time point t. For all of the history covariates described, we focus on using the primary alternating recurrent event time, but note that the secondary alternating recurrent event times may also be used to derive other dynamic predictors.

While complete historical information on subjects yields the best results in simulation, we recognize that such information may not always be available. The first partial history setting generates the same possible covariates as the fully observed history setting, but restricts q_i to be at most four. This means that for an alternating recurrent event history on a subject, within the time span from t_h to t_0, at most four events will be used in constructing covariates that follow equations B.1- B.3. We refer to covariates of this type as $\boldsymbol{H}_{p_1}$; similarly covariates of type $\boldsymbol{H}_{p_2}$ have q_i equal to at most two. A third partial history setting may arise when a proportion of subjects have complete subject history t_h, and the remaining subjects have no recorded event history. In simulation, we refer to covariates arising from this historical setting as $\boldsymbol{H}_{p_3}$. Moreover, an analyst may not have any historical event information, which results in only history covariates constructed from events on study ($\boldsymbol{H}_{p_4}$), or simply not include any history covariates, $\boldsymbol{H}_n$.

In considering updating history covariates, the random forest described in Section 3.5 is only concerned with predicting $R_i(t)$ at each time point; thus covariates that may perfectly predict $\min(T_i(t), \tau)$ are permissible for use in our random forest algorithm. Furthermore, an individual's tendency to enter into the risk-set at other windows may be an important

covariate in fitting a model to $R_i(t)$, particularly if they rarely appear in the risk-set for the primary event-type or almost always appear in the risk-set for the primary event-type; that is $R_i(t_l) = 0$ for all $t \neq t_l$ or $R_i(t_l) = 1$ for all $t \neq t_l$ respectively. Therefore, we introduce a new way to consider historical information, called *reflective* history, or $\boldsymbol{H}_{p*}$.

Consider the prediction of $R_i(t_1)$, or the prediction of risk set membership at the start of the second follow-up window in the $\boldsymbol{H}_{p_4}$ setting. There are likely few observations of either type of recurrent event outcome. However, there is a rich set of covariates useful for prediction of behavior in the window starting at t_1; those are the events occurring after t_1, particularly in the absence of censoring. Note that we are not using future primary alternating recurrent event times to predict current primary alternating recurrent event times. We are simply using future, already observed primary and secondary event types to better fit a model for $R_i(t)$. Such settings commonly arise when analyzing data from clinical trials; all outcomes have been recorded for the entire follow-up at data lock, so future primary and secondary events have been fully observed. Indeed, this is an extension of an argument made by Sun and Tchetgen Tchetgen [2018], where missingness patterns across all time points was considered in maximum likelihood estimation with longitudinal data subject to nonmonotone dropout. In the $\boldsymbol{H}_{p*}$ setting, an analyst keeps track of the number of observed events both prior to and subsequent to all $t \in \mathcal{T}$, and then calculates covariates of the type of Equations B.1-B.3 using events prior to t if there are more events before the window start time, or events later than t if the opposite is true. For window start-times where there are relatively few observations available for small t and no censoring, we expect improved model performance in the $\boldsymbol{H}_{p*}$ setting compared to the $\boldsymbol{H}_{p_4}$ setting.

B.4 Independent Event Times

We assume that gap times between alternating recurrent events are independent and identically distributed with an exponential distribution, though primary and secondary event types may have different rate parameters, λ_i and γ_i respectively. Xia et al. [2020] showed that for independent exponentially distributed gap-times, the time-to-first-event in a follow-up window for the censored longitudinal data structure remains exponentially distributed with rate λ. We first simulate $Z_1 \sim \text{Beta}(5,1)$, $Z_2, Z_3 \sim MVN(\vec{0}, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}$. The corresponding λ_i that will generate exponential gap times that follow our model are taken as the solution of:

$$\mathbb{E}\left[\min\{T_i(t), \tau\} \mid Z_i(t)\right] = \int_{0^+}^{\tau^-} y\lambda_i e^{-\lambda_i y} dy + \tau e^{-\lambda_i \tau} = \beta_0 + \beta_1 Z_{i1} + \beta_2 Z_{i2} + \beta_3 I(Z_{i3} \geq 0) \times Z_{i1},$$

where $\beta_0 = 0.5$, $\beta_1 = 0.25$, $\beta_2 = -0.4$ and $\beta_3 = 0.3$. Combinations of covariates that generate hazards outside of the range of $\frac{1}{6}$ to $\frac{5}{8}$ for the primary alternating recurrent event-type will be discarded. The secondary alternating recurrent event-type has a hazard described by $\gamma_i = \exp\{X_{i2} + \frac{1}{2}X_{i3} + 10X_{i6}^2 - 5(Z_{i1} + Z_{i3})\}$, where $X_2 \sim N(1, .25)$, $X_3 \sim \chi_3^2$, and $X_6 \sim Bern(.5)$. There are three included noise covariates (X_1, X_4, X_5), but the inclusion of Z_1 and Z_3 in the generation of γ_i induces light correlation between λ_i and γ_i (ρ of approximately -0.10). Results, displayed in Table B.2, are comforting. As with most longitudinal and recurrent event models, if events are independent of each other, all models work well, are relatively unbiased, and demonstrate appropriate asymptotic behavior. The distribution of point estimates in the presence of censoring are displayed in Figure B.3.

B.5 Tables and Figures

Table B.1: A table displaying results from simulation in the case where primary recurrent events have correlation of $\rho = 0.8$, and secondary recurrent events are independent of each other. The hazard of primary alternating recurrent event completely determines the hazard of the alternating secondary recurrent event, no subjects are censored, and results come from 500 replicates of $n = 750$ each.

Model	Covariate	$\widehat{\beta} - \beta$	$\lvert\widehat{\beta} - \beta\rvert/\beta$	Coverage	$\widehat{SE}$	ESD	$\widehat{SE}$/ESD
PAIR-GEE, with $\boldsymbol{H}_f$	$\widehat{\beta}_1$	-0.001	0.001	0.950	0.013	0.013	0.985
	$\widehat{\beta}_2$	-0.003	0.004	0.919	0.015	0.016	0.931
	$\widehat{\beta}_3$	0.006	0.007	0.949	0.011	0.011	1.037
	$\widehat{\beta}_4$	-0.001	0.001	0.913	0.013	0.014	0.906
PAIR-GEE with $\boldsymbol{H}_{p_1}$	$\widehat{\beta}_1$	0.001	0.002	0.936	0.014	0.014	0.951
	$\widehat{\beta}_2$	-0.004	0.006	0.913	0.015	0.016	0.912
	$\widehat{\beta}_3$	0.004	0.005	0.947	0.011	0.011	1.028
	$\widehat{\beta}_4$	-0.006	0.008	0.881	0.012	0.014	0.904
PAIR-GEE, with $\boldsymbol{H}_{p_2}$	$\widehat{\beta}_1$	0.000	0.001	0.937	0.013	0.014	0.949
	$\widehat{\beta}_2$	-0.008	0.012	0.887	0.014	0.015	0.921
	$\widehat{\beta}_3$	0.004	0.005	0.953	0.011	0.011	1.019
	$\widehat{\beta}_4$	-0.008	0.011	0.846	0.012	0.014	0.897
PAIR-GEE, with $\boldsymbol{H}_{p*}$	$\widehat{\beta}_1$	-0.002	0.002	0.969	0.013	0.012	1.085
	$\widehat{\beta}_2$	-0.003	0.004	0.931	0.015	0.016	0.965
	$\widehat{\beta}_3$	-0.002	0.002	0.947	0.010	0.010	1.051
	$\widehat{\beta}_4$	-0.021	0.028	0.510	0.011	0.012	0.945
PAIR-GEE, with $\boldsymbol{H}_{p_3}$	$\widehat{\beta}_1$	-0.013	0.018	0.793	0.012	0.012	1.009
	$\widehat{\beta}_2$	-0.025	0.039	0.507	0.013	0.013	0.977
	$\widehat{\beta}_3$	-0.001	0.001	0.953	0.010	0.010	1.030
	$\widehat{\beta}_4$	-0.013	0.017	0.763	0.012	0.013	0.906
PAIR-GEE, with $\boldsymbol{H}_{p_4}$	$\widehat{\beta}_1$	-0.016	0.023	0.697	0.012	0.011	1.018
	$\widehat{\beta}_2$	-0.033	0.052	0.277	0.012	0.013	0.984
	$\widehat{\beta}_3$	-0.001	0.001	0.963	0.010	0.010	1.030
	$\widehat{\beta}_4$	-0.020	0.027	0.539	0.011	0.012	0.925
PAIR-GEE, with $\boldsymbol{H}_n$	$\widehat{\beta}_1$	-0.024	0.034	0.424	0.011	0.011	0.991
	$\widehat{\beta}_2$	-0.036	0.056	0.204	0.012	0.013	0.961
	$\widehat{\beta}_3$	-0.021	0.025	0.317	0.009	0.009	1.000
	$\widehat{\beta}_4$	-0.042	0.055	0.017	0.010	0.011	0.914
Unweighted GEE	$\widehat{\beta}_1$	-0.026	0.036	0.357	0.011	0.011	1.019
	$\widehat{\beta}_2$	-0.043	0.068	0.059	0.012	0.012	1.009
	$\widehat{\beta}_3$	-0.022	0.026	0.280	0.008	0.008	1.025
	$\widehat{\beta}_4$	-0.047	0.061	0.001	0.009	0.009	0.968

Table B.2: A table displaying results from simulation in the case where primary recurrent events and secondary recurrent events are independent of each other. The hazard of primary alternating recurrent event completely determines the hazard of the alternating secondary recurrent event, approximately 30% of subjects are censored, and results come from 300 replicates of $n = 750$ each.

Model	Covariate	$\widehat{\beta} - \beta$	$\lvert\widehat{\beta} - \beta\rvert/\beta$	Coverage	$\widehat{SE}$	ESD	$\widehat{SE}$/ESD
PAIR-GEE, with $\boldsymbol{H}_f$	β_0	-0.001	0.002	0.957	0.034	0.035	0.980
	β_1	0.003	0.012	0.950	0.052	0.052	0.987
	β_2	0.000	0.001	0.937	0.036	0.038	0.947
	β_3	-0.002	0.008	0.950	0.037	0.038	0.974
PAIR-GEE, with $\boldsymbol{H}_{p_1}$	β_0	-0.001	0.002	0.953	0.034	0.034	0.983
	β_1	0.003	0.012	0.957	0.051	0.052	0.985
	β_2	0.000	0.001	0.933	0.036	0.038	0.947
	β_3	-0.002	0.008	0.947	0.037	0.038	0.974
PAIR-GEE, with $\boldsymbol{H}_{p_1}$	β_0	-0.001	0.002	0.957	0.034	0.034	0.983
	β_1	0.003	0.012	0.950	0.052	0.052	0.987
	β_2	0.000	0.000	0.933	0.036	0.038	0.947
	β_3	-0.002	0.008	0.947	0.037	0.038	0.974
PAIR-GEE, with $\boldsymbol{H}_{p*}$	β_0	0.004	0.008	0.947	0.034	0.035	0.984
	β_1	0.000	0.002	0.947	0.052	0.053	0.986
	β_2	0.006	0.014	0.923	0.036	0.038	0.953
	β_3	-0.006	0.021	0.953	0.037	0.038	0.978
PAIR-GEE, with $\boldsymbol{H}_{p_3}$	β_0	-0.001	0.002	0.953	0.034	0.034	0.983
	β_1	0.003	0.013	0.947	0.052	0.052	0.987
	β_2	0.000	0.000	0.933	0.036	0.038	0.950
	β_3	-0.002	0.008	0.950	0.037	0.038	0.974
PAIR-GEE, with $\boldsymbol{H}_{p_4}$	β_0	-0.001	0.002	0.953	0.034	0.035	0.980
	β_1	0.003	0.012	0.947	0.052	0.052	0.985
	β_2	0.000	0.000	0.930	0.036	0.038	0.950
	β_3	-0.002	0.008	0.953	0.037	0.038	0.974
PAIR-GEE, with $\boldsymbol{H}_n$	β_0	-0.001	0.002	0.957	0.034	0.035	0.980
	β_1	0.003	0.012	0.947	0.051	0.052	0.983
	β_2	0.000	0.001	0.927	0.036	0.038	0.949
	β_3	-0.002	0.008	0.947	0.037	0.038	0.976
Unweighted GEE	β_0	-0.001	0.002	0.947	0.034	0.035	0.977
	β_1	0.003	0.011	0.947	0.052	0.053	0.974
	β_2	0.000	0.000	0.930	0.036	0.037	0.960
	β_3	-0.002	0.008	0.943	0.037	0.037	0.992

Table B.3: A table displaying results from simulation in the case where primary recurrent events have correlation of $\rho = 0.8$, and secondary recurrent events are independent of each other. This is the same case as 3.1, yet simulation results are calculated using the weighted bootstrap summarized in Williams and Savitsky [2021]. The hazard of primary alternating recurrent events completely determines the hazard of the alternating secondary recurrent event, approximately 30% of subjects are censored before the final follow-up window, and results come from 500 replicates of $n = 750$ each.

Model	Covariate	$\widehat{\beta} - \beta$	$\lvert\widehat{\beta} - \beta/\beta$	Coverage	$\widehat{SE}$	ESD	$\widehat{SE}$/ESD
PAIR-GEE, with $\boldsymbol{H}_f$	$\widehat{\beta}_1$	-0.001	0.002	0.992	0.018	0.014	1.321
(Bootstrap)	$\widehat{\beta}_2$	-0.003	0.004	0.986	0.021	0.016	1.270
	$\widehat{\beta}_3$	0.005	0.006	0.992	0.016	0.011	1.431
	$\widehat{\beta}_4$	-0.002	0.002	0.982	0.018	0.014	1.277
PAIR-GEE, with $\boldsymbol{H}_{p_1}$	$\widehat{\beta}_1$	0	0	0.978	0.019	0.015	1.255
(Bootstrap)	$\widehat{\beta}_2$	-0.004	0.006	0.980	0.020	0.016	1.232
	$\widehat{\beta}_3$	0.003	0.004	0.990	0.016	0.011	1.393
	$\widehat{\beta}_4$	-0.007	0.009	0.964	0.018	0.014	1.255
PAIR-GEE, with $\boldsymbol{H}_{p_2}$	$\widehat{\beta}_1$	-0.001	0.001	0.990	0.018	0.014	1.271
(Bootstrap)	$\widehat{\beta}_2$	-0.008	0.013	0.964	0.019	0.015	1.293
	$\widehat{\beta}_3$	0.003	0.003	0.992	0.015	0.011	1.387
	$\widehat{\beta}_4$	-0.008	0.011	0.952	0.017	0.014	1.270
PAIR-GEE, with $\boldsymbol{H}_{p*}$	$\widehat{\beta}_1$	-0.003	0.004	0.996	0.018	0.013	1.441
(Bootstrap)	$\widehat{\beta}_2$	-0.005	0.008	0.982	0.021	0.015	1.397
	$\widehat{\beta}_3$	-0.003	0.004	0.988	0.014	0.010	1.464
	$\widehat{\beta}_4$	-0.023	0.030	0.744	0.016	0.012	1.374
PAIR-GEE, with $\boldsymbol{H}_{p_3}$	$\widehat{\beta}_1$	-0.014	0.019	0.908	0.016	0.012	1.355
(Bootstrap)	$\widehat{\beta}_2$	-0.025	0.039	0.792	0.018	0.013	1.368
	$\widehat{\beta}_3$	-0.001	0.002	0.996	0.015	0.010	1.408
	$\widehat{\beta}_4$	-0.013	0.017	0.920	0.016	0.013	1.299
PAIR-GEE, with $\boldsymbol{H}_{p_4}$	$\widehat{\beta}_1$	-0.017	0.024	0.864	0.016	0.012	1.353
(Bootstrap)	$\widehat{\beta}_2$	-0.033	0.052	0.552	0.018	0.013	1.357
	$\widehat{\beta}_3$	-0.001	0.002	0.988	0.015	0.010	1.394
	$\widehat{\beta}_4$	-0.020	0.026	0.806	0.016	0.012	1.295
PAIR-GEE, with $\boldsymbol{H}_n$	$\widehat{\beta}_1$	-0.025	0.034	0.680	0.015	0.012	1.330
(Bootstrap)	$\widehat{\beta}_2$	-0.034	0.055	0.480	0.017	0.013	1.303
	$\widehat{\beta}_3$	-0.021	0.026	0.594	0.012	0.009	1.407
	$\widehat{\beta}_4$	-0.042	0.054	0.104	0.014	0.011	1.324

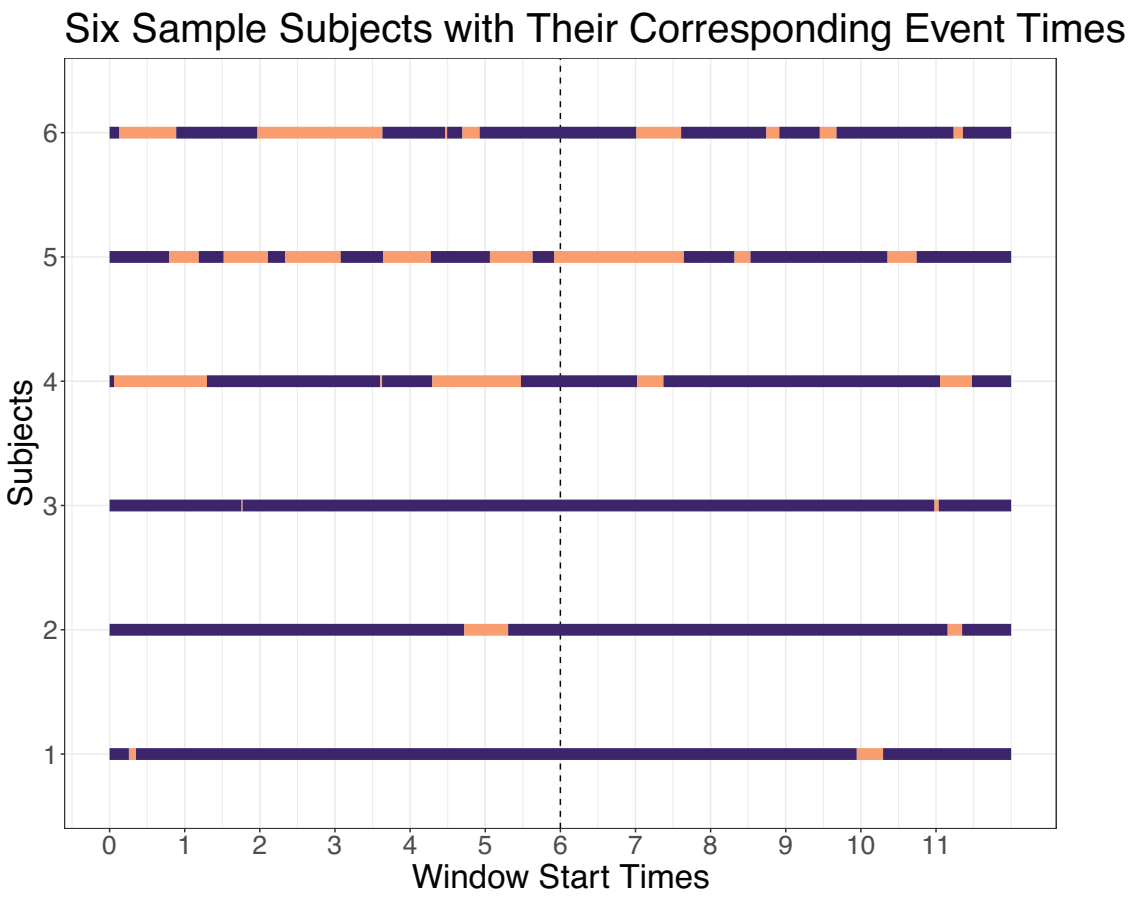


Figure B.1: Six representative simulated subjects' alternating recurrent event data, following the model described in Section 3.6. Individuals $i = 1, 2, 3$ have $Z_i = 0$, and all other individuals have covariate profile $Z_i = 1$. Window start times, $t \in \mathcal{T}$, are displayed on the horizontal axis. Primary and secondary event types are the start and the end of the light salmon segments, respectively. Purple segments represent that at-risk periods of the primary event type. The time at which subjects experience a bend in hazards is distinguished with a vertical dashed line. Subjects 1 through 3 experience more time at risk relative to subjects $4, 5$, and 6. The latter subjects are more likely to be missing from the complete case analysis at window start times, $t \in \mathcal{T}$.

Table B.4: A table displaying results from the same setting as results displayed in Table 3.1; however, the standard errors are calculated via a bootstrap for weighted, clustered data [Williams and Savitsky, 2021]. The mean of the variances of bootstrap replicates is reported as the estimated variance.

	Covariate	$\widehat{\beta}-\beta$	$\lvert\widehat{\beta}-\beta\rvert/\beta$	Coverage	$\widehat{SE}$	ESD	$\widehat{SE}$/ESD
PAIR-GEE, with$\boldsymbol{H}_f$	$\widehat{\beta}_1$	-0.001	0.002	0.992	0.018	0.014	1.321
(Bootstrap)	$\widehat{\beta}_2$	-0.003	0.004	0.986	0.021	0.016	1.270
	$\widehat{\beta}_3$	0.005	0.006	0.992	0.016	0.011	1.431
	$\widehat{\beta}_4$	-0.002	0.002	0.982	0.018	0.014	1.277
PAIR-GEE, with $\boldsymbol{H}_{p_1}$	$\widehat{\beta}_1$	0	0	0.978	0.019	0.015	1.255
(Bootstrap)	$\widehat{\beta}_2$	-0.004	0.006	0.980	0.020	0.016	1.232
	$\widehat{\beta}_3$	0.003	0.004	0.990	0.016	0.011	1.393
	$\widehat{\beta}_4$	-0.007	0.009	0.964	0.018	0.014	1.255
PAIR-GEE, with $\boldsymbol{H}_{p_2}$	$\widehat{\beta}_1$	-0.001	0.001	0.990	0.018	0.014	1.271
(Bootstrap)	$\widehat{\beta}_2$	-0.008	0.013	0.964	0.019	0.015	1.293
	$\widehat{\beta}_3$	0.003	0.003	0.992	0.015	0.011	1.387
	$\widehat{\beta}_4$	-0.008	0.011	0.952	0.017	0.014	1.270
PAIR-GEE, with $\boldsymbol{H}_{p*}$	$\widehat{\beta}_1$	-0.003	0.004	0.996	0.018	0.013	1.441
(Bootstrap)	$\widehat{\beta}_2$	-0.005	0.008	0.982	0.021	0.015	1.397
	$\widehat{\beta}_3$	-0.003	0.004	0.988	0.014	0.010	1.464
	$\widehat{\beta}_4$	-0.023	0.030	0.744	0.016	0.012	1.374
PAIR-GEE, with $\boldsymbol{H}_{p_3}$	$\widehat{\beta}_1$	-0.014	0.019	0.908	0.016	0.012	1.355
(Bootstrap)	$\widehat{\beta}_2$	-0.025	0.039	0.792	0.018	0.013	1.368
	$\widehat{\beta}_3$	-0.001	0.002	0.996	0.015	0.010	1.408
	$\widehat{\beta}_4$	-0.013	0.017	0.920	0.016	0.013	1.299
PAIR-GEE, with $\boldsymbol{H}_{p_4}$	$\widehat{\beta}_1$	-0.017	0.024	0.864	0.016	0.012	1.353
(Bootstrap)	$\widehat{\beta}_2$	-0.033	0.052	0.552	0.018	0.013	1.357
	$\widehat{\beta}_3$	-0.001	0.002	0.988	0.015	0.010	1.394
	$\widehat{\beta}_4$	-0.020	0.026	0.806	0.016	0.012	1.295
PAIR-GEE, with $\boldsymbol{H}_n$	$\widehat{\beta}_1$	-0.025	0.034	0.680	0.015	0.012	1.330
(Bootstrap)	$\widehat{\beta}_2$	-0.034	0.055	0.480	0.017	0.013	1.303
	$\widehat{\beta}_3$	-0.021	0.026	0.594	0.012	0.009	1.407
	$\widehat{\beta}_4$	-0.042	0.054	0.104	0.014	0.011	1.324

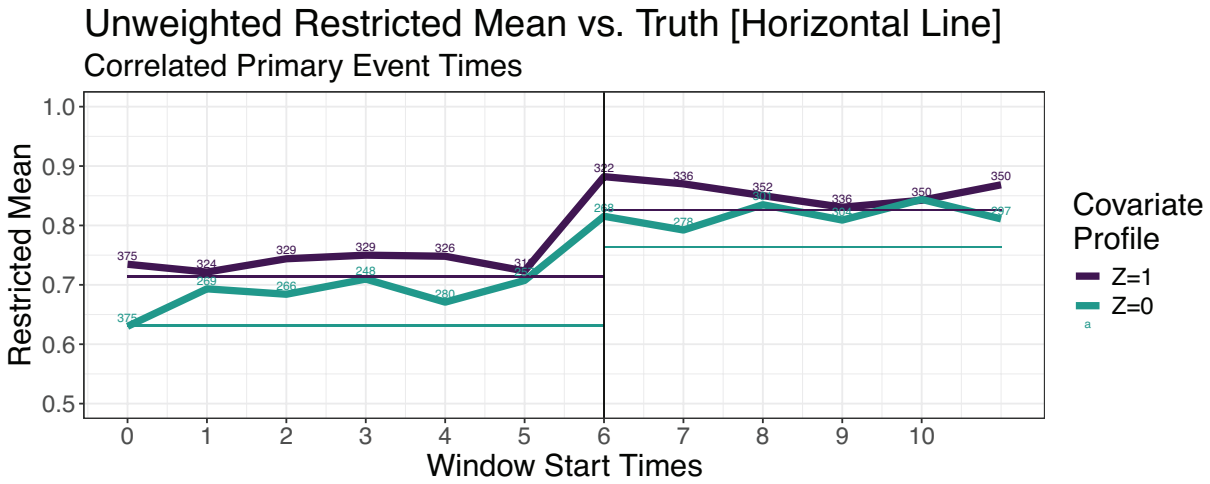


Figure B.2: A graphic displaying the true τ-restricted mean compared to the empirical, or observed, τ-RMST in the highly correlated case for a single replicate. True parameter estimates are displayed using horizontal lines, while the empirical value is displayed on the y axis. Lines connecting the estimated τ-RMST are annotated with the number of subjects in the risk set for each window start time and each covariate profile.

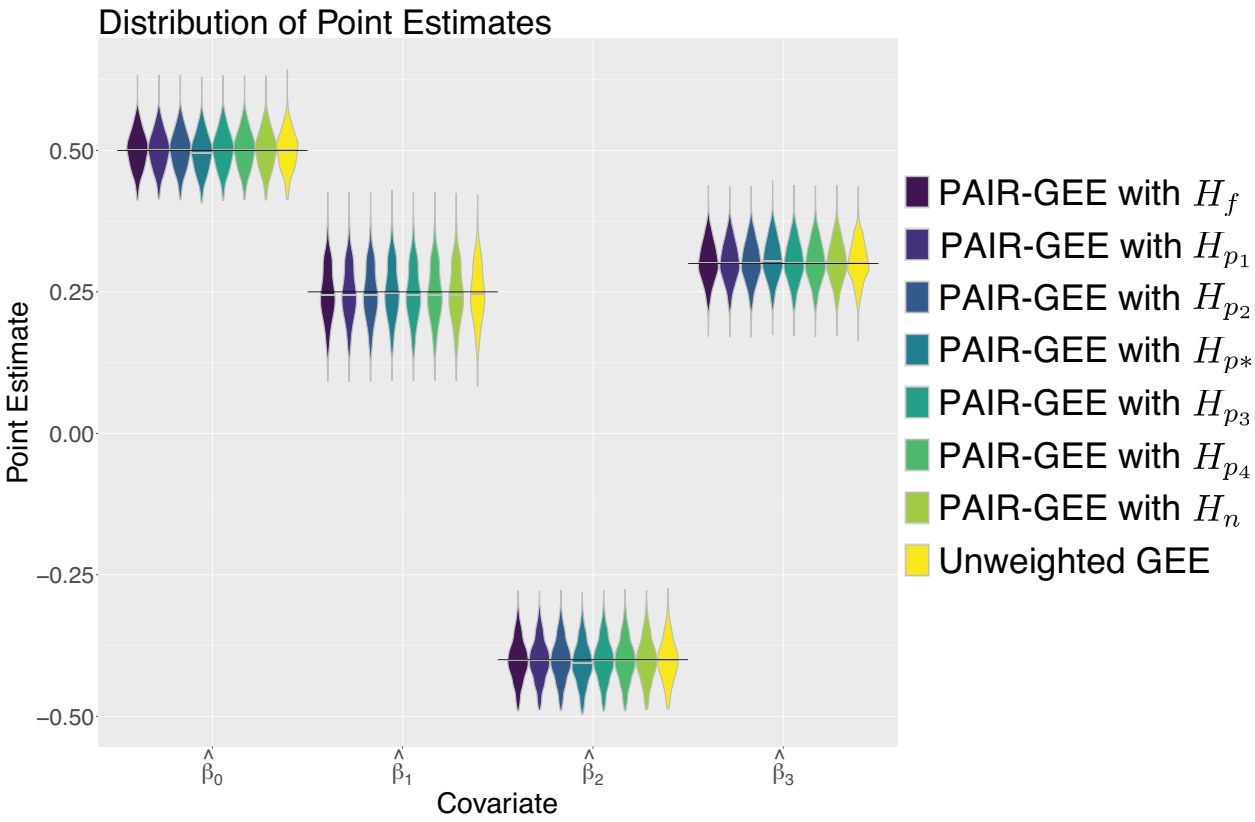


Figure B.3: Distribution of point estimates in the presence of censoring, when primary alternating recurrent events within a subject are independent This corresponds to the case of missing completely at random for exponentially distributed gap times. Results come from 300 replicates of $n = 750$ each. Features included in the random forest algorithm for estimating PAIR-GEE weights involve the seven history covariates described in Section 3.5 with varying degrees of availability that decreases from left to right for each displayed coefficient

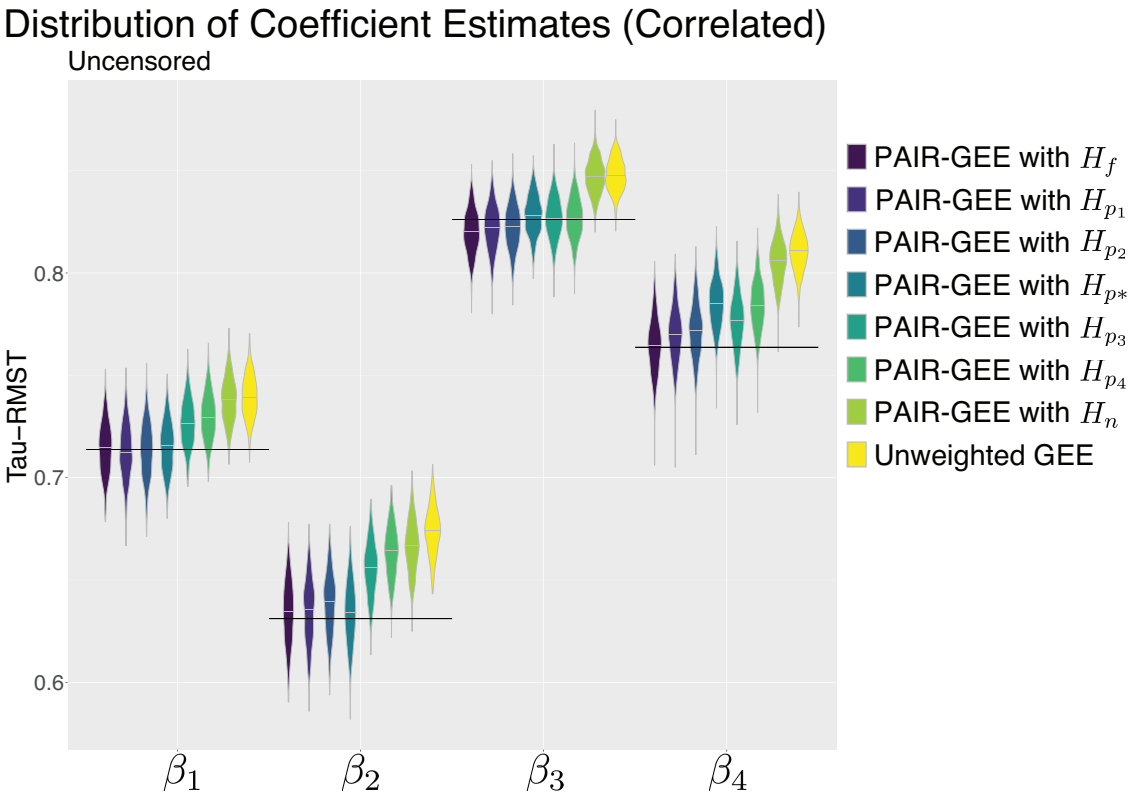


Figure B.4: Violin plots of uncensored point estimates seen in simulation (500 replicates, $n = 750$). Coefficients β_1 and β_2 reflect those with coefficients $Z_i = 1$ and $Z_i = 0$ before $t = 6$ and coefficients β_3 and β_4 after and including $t = 6$, respectively. Features included in the random forest algorithm for estimating PAIR-GEE weights involve the seven history covariates described in Section 3.5 with varying degrees of availability that decreases from left to right for each displayed coefficient. Here, $\boldsymbol{H}_f$ indicates the full year of patient history was available prior to $t = 0$. $\boldsymbol{H}_{p_1}$ and $\boldsymbol{H}_{p_2}$ indicate that the patient history included information on the last two or one alternating recurrent events, respectively, if these occurred within the prior year. $\boldsymbol{H}_{p*}$ indicates that no historical information prior to $t = 0$ was available and that $\boldsymbol{H}_{p^*}$ history covariates (as described in Section 3.5) were used. $\boldsymbol{H}_{p_3}$ indicates that complete history was available on 50% of individuals, but was otherwise unavailable. $\boldsymbol{H}_{p_4}$ indicates no availability of history covariates prior to $t = 0$. $\boldsymbol{H}_n$ indicates $\boldsymbol{H}_n$ were included in the random forest algorithm.

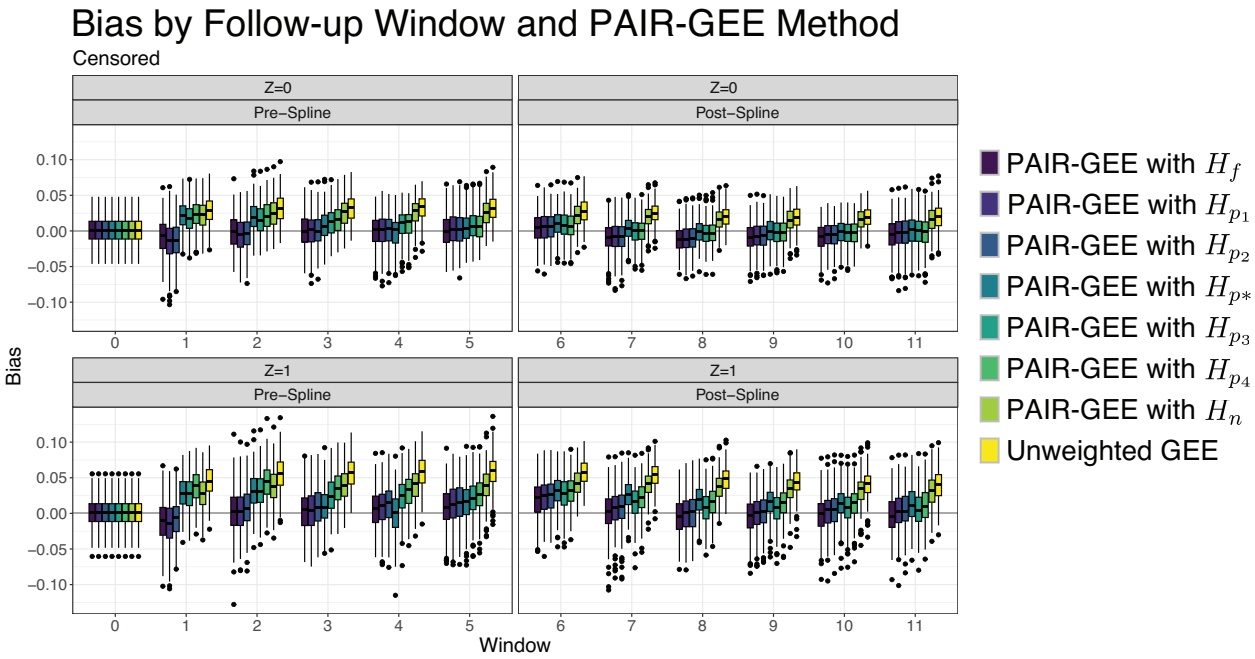


Figure B.5: Bias in point estimation for each $t \in \mathcal{T}$ when alternating recurrent events are correlated with each other. The top half of the graphic displays results from the uncensored simulation, with the bottom half displaying results from the censored results, described in the main paper. Notice the persistent bias for the unweighted methods, as well as the PAIR-GEE with $\boldsymbol{H}_n$, while other methods of incorporating history covariates are able, to varying extents, adjust for patient specific history in a more elegant manner, and approach non-zero bias in later follow-up windows.

APPENDIX C

Appendix for Chapter 4

C.1 History Covariates Used in Simulation

We derive seven history covariates of the type described in Loe et al. [2025a]. Covariates include: (1) mean time in the at-risk population, (2) time since last return to risk-population, (3) length of last removal from risk population, (4) average time removed from the risk population, (5) length of last fully observed time at-risk, (6) most recent observed event time, and (7) the number of alternating recurrent events observed. In calculating these covariates for $\boldsymbol{H}_\ell$ settings, we use a single imputation for time t_0, and continue to use the imputation until at least one pair of alternating recurrent events has been observed on study time. These covariates are used in the restricted mean LSTM models.

We also include indicators of lagged at-risk status; that is, at t_k we may define earlier at risk status for $t_{k-\ell}, \ell = 1, \ldots, (k-1)$ via the indicator, $\mathcal{L}_i(t_k, \ell) = R_i(t_{k-\ell})$. Collectively, this past recurrent event status history as measured by $\mathcal{L}_i(t_k, \ell)$ summarizes the missingness pattern of our data [see, for example, Little and Rubin, 2002, Sun and Tchetgen Tchetgen, 2018, Chen et al., 2026].

In incorporating history covariates in the longitudinal random forest of the type from Loe et al. [2025b], an analyst selects variables to always include in the construction of each random tree. For $\boldsymbol{H}_\ell$ settings, where only history observed on study time is recorded, and history covariates at t_0 are imputed, an analyst may take a pragmatic approach, and only force splits on history covariates after enough follow-up has commenced. In our simulations, required splits on history covariates for random forest weight estimation occurred only after t_5.

C.2 Oracle Model

Here we provide details for those interested in the derivation of the oracle model for $\min\{T_i(t), \tau\}$. Figure C.1 displays a sample individual with burn-in length b_{pre} and follow-

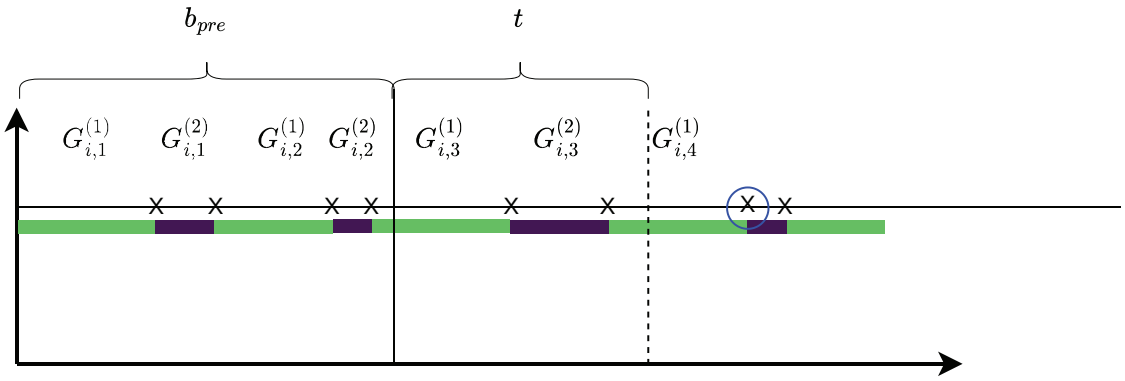


Figure C.1: A graphic displaying the generating model for simulation, where b_{pre} denotes burn-in, t is window start-time beyond burn-in, and $G_{i,j}^{(k)}$ denote the correlated multivariate normal gap-times. The primary alternating recurrent event of interest that forms $T_i(t)$ is circled in blue, while primary event gap times are displayed in green, and secondary event times are displayed in purple.

up window start time t. Events are denoted by "X," with primary events displayed in green, and secondary events in purple. The $T_{i,j}^{(1)}$ that corresponds to the same event in creating $T_i(t)$ is circled in blue. We may re-write $T_{i,j}^{(1)}$ as $G_{i,j}^{(1)} + \sum_{k<j} G_{i,k}^{(1)} + G_{i,k}^{2}$, where the marginal distribution of each $G_{i,j}^{(1)}$ is normal, with mean $\mu_i^{(1)}$ and variance σ^2, and $G_{i,j}^{(2)}$ is normal, with mean $\mu_i^{(2)}$ and variance σ^2. Re-arranging and subtracting t, we obtain $T_i(t) \sim \text{Normal}\Big[\{N_i^{(1)}(t^-)+1\}\mu_i^{(1)} + N_i^{(2)}(t^-)\mu_i^{(2)} - (t+b_{pre}), \tilde{\sigma}^2\Big]$, where $N_i^{(k)}(t^-) = \sum_j I(T_{i,j}^{(k)} < t + b_{pre})$ is the number of events of type k experienced by subject i prior to time t including those accumulated during the burn-in period of duration, b_{pre}.

C.3 Data Application

Table C.1: The point estimates with corresponding standard errors for the main effects model at time $t = 350$ for the IHS data with LSTM weights. The reference levels for marital status is ""ingle, device reference is "unknown", tobacco use at baseline is "none", and alcohol use is "none." The reference level for sex is "male." History covariates of the type $\boldsymbol{H}_\ell$ are denoted with a †.

Covariate	Estimate	SE	Z	P-value
Intercept	11.6873	0.4730	610.4776	**<0.0001**
Age	0.0099	0.0092	1.1539	0.2827
Sex	-0.0583	0.0607	0.9226	0.3368

Covariate	Estimate	SE	Z	P-value
Marital Status: Married	-0.0043	0.0987	0.0019	0.9655
Marital Status: Separated/Divorced	-0.8771	0.6560	1.7880	0.1812
Marital Status: Not in Committed Relationship	0.0403	0.0954	0.1788	0.6724
Marital Status: In a Committed Relationship	-0.0122	0.0940	0.0167	0.8971
Neuroticism Score at Baseline	-0.0054	0.0034	2.5320	0.1116
Sexual Experiences Questionnaire Score	0.0051	0.0088	0.3286	0.5665
Early Environment Exposures at Baseline	0.0075	0.0027	7.5134	**0.0061**
Depression Treatment at Baseline	0.0395	0.0751	0.2763	0.5991
Last Measured Daily Steps (Standardized)	-0.0510	0.0253	4.0439	**0.0443**
Last Measured Resting Heart Rate	-0.0037	0.0013	8.1701	**0.0043**
Specialty: Child Neurology	0.3999	0.3374	1.4047	0.2359
Specialty: Dermatology	-0.3200	0.5579	0.3290	0.5663
Specialty: Emergency Medicine	0.1041	0.1313	0.6285	0.4279
Specialty: Family Medicine	0.1224	0.1453	0.7098	0.3995
Specialty: General Surgery	-0.0667	0.1524	0.1917	0.6615
Specialty: Internal Medicine	0.3034	0.1170	6.7249	**0.0095**
Specialty: Neurological Surgery	0.2012	0.3173	0.4020	0.5260
Specialty: Neurology	0.4683	0.1883	6.1819	0.0129
Specialty: Obstetrics And Gynecology	-0.0459	0.1696	0.0734	0.7865
Specialty: Ophthalmology	-0.2862	0.3056	0.8766	0.3491
Specialty: Orthopedic Surgery	0.1094	0.2204	0.2464	0.6196
Specialty: Otolaryngology	0.1096	0.2705	0.1641	0.6854
Specialty: Pathology	-0.2980	0.2573	1.3416	0.2468
Specialty: Pediatrics	0.3517	0.1330	6.9919	**0.0082**
Specialty: Physical Medicine	-0.2590	0.3647	0.5045	0.4775
Specialty: Plastic Surgery	-0.9375	0.5966	2.4692	0.1161
Specialty: Psychiatry	0.1332	0.1449	0.8445	0.3581
Specialty: Radiology	-0.1434	0.3688	0.1512	0.6974
Specialty: Surgery - General	1.2243	0.4464	7.5219	**0.0061**
Specialty: Thoracic Surgery - Integrated	-0.5654	0.4566	1.5334	0.2156
Specialty: Transitional Year	0.2116	0.1625	1.6950	0.1929
Specialty: Triple Board	1.9607	1.3371	2.1501	0.1426
Specialty: Urology	-0.2539	0.2504	1.0275	0.3108
Specialty: Vascular Surgery	0.3300	0.3852	0.7341	0.3916
Depression at Baseline	-0.2191	0.1537	2.0320	0.1540

Covariate	Estimate	SE	Z	P-value
Tobacco Use at Baseline: At least some	-0.0382	0.0906	0.1780	0.6731
Alcohol: Low	-0.0127	0.0898	0.0200	0.8875
Alcohol: High	0.0536	0.0863	0.3860	0.5344
Depression Age	0.0003	0.0068	0.0024	0.9610
Disability (Yes)	-0.1752	0.0216	65.6693	**<0.0001**
Device: Apple Watch	0.3150	0.3445	0.8361	0.3605
Device: Fitbit	0.3509	0.3512	0.9980	0.3178
Device: Garmin	0.1696	0.3648	0.2161	0.6420
Last Measured Sleep Duration in Minutes	-0.0004	0.0001	7.5397	**0.0060**
Time of Since Most Recent Event†	-0.0094	0.0009	106.9214	**<0.0001**
Average Time Since Last Event†	0.0199	0.0010	408.0342	**<0.0001**
Average Cycle Duration†	-0.0138	0.0014	93.7753	**<0.0001**
Number Of Events Observed†	-0.2715	0.0086	987.7060	**<0.0001**
Average Time Recovered†	0.0128	0.0007	384.2739	**<0.0001**
Average Time Depressed†	0.0168	0.0050	11.2241	**0.0008**

Table C.2: The point estimates with corresponding standard errors for the main effects model at time $t = 350$ for the IHS data with RF weights. The reference levels for marital status is "single," device reference is "unknown", tobacco use at baseline is "none", and alcohol use is "none." The reference level for sex is "male." History covariates of the type $\boldsymbol{H}_\ell$ are denoted with a †.

Covariate	Estimate	SE	Z	P-value
Intercept	11.6871	0.4866	576.7915	**<0.0001**
Age	0.0111	0.0094	1.4070	0.2356
Sex	-0.0567	0.0627	0.8198	0.3653
Marital Status: Married	0.0008	0.1004	0.0001	0.9940
Marital Status: Separated/Divorced	-0.9118	0.7043	1.6764	0.1954
Marital Status: Not in Committed Relationship	0.0346	0.0969	0.1275	0.7210
Marital Status: In a Committed Relationship	-0.0068	0.0958	0.0051	0.9431
Neuroticism Score at Baseline	-0.0069	0.0036	3.7278	0.0535
Sexual Experiences Questionnaire Score	0.0046	0.0090	0.2585	0.6111
Early Environment Exposures at Baseline	0.0085	0.0028	9.2391	**0.0024**
Depression Treatment at Baseline	0.0402	0.0775	0.2690	0.6040
Last Measured Daily Steps (Standardized)	-0.0524	0.0258	4.1166	**0.0425**

Covariate	Estimate	SE	Z	P-value
Last Measured Resting Heart Rate	-0.0040	0.0013	9.0606	**0.0026**
Specialty: Child Neurology	0.5175	0.3518	2.1641	0.1413
Specialty: Dermatology	-0.2775	0.5460	0.2583	0.6113
Specialty: Emergency Medicine	0.1201	0.1358	0.7822	0.3765
Specialty: Family Medicine	0.1489	0.1493	0.9950	0.3185
Specialty: General Surgery	-0.0826	0.1594	0.2686	0.6043
Specialty: Internal Medicine	0.3327	0.1221	7.4234	**0.0064**
Specialty: Neurological Surgery	0.2930	0.3515	0.6949	0.4045
Specialty: Neurology	0.4508	0.1960	5.2897	**0.0215**
Specialty: Obstetrics And Gynecology	-0.0042	0.1733	0.0006	0.9808
Specialty: Ophthalmology	-0.1588	0.3151	0.2541	0.6142
Specialty: Orthopedic Surgery	0.1915	0.2229	0.7375	0.3905
Specialty: Otolaryngology	-0.0587	0.3019	0.0378	0.8459
Specialty: Pathology	-0.3031	0.2655	1.3037	0.2535
Specialty: Pediatrics	0.3804	0.1374	7.6700	**0.0056**
Specialty: Physical Medicine	-0.9196	0.4488	4.1983	**0.0405**
Specialty: Plastic Surgery	-1.2734	0.6796	3.5109	0.0610
Specialty: Psychiatry	0.1615	0.1491	1.1738	0.2786
Specialty: Radiology	-0.1946	0.3808	0.2612	0.6093
Specialty: Surgery - General	1.2241	0.4545	7.2542	**0.0071**
Specialty: Thoracic Surgery - Integrated	-0.6304	0.4358	2.0923	0.1480
Specialty: Transitional Year	0.1749	0.1694	1.0661	0.3018
Specialty: Triple Board	1.9272	1.3967	1.9041	0.1676
Specialty: Urology	-0.2547	0.2575	0.9786	0.3225
Specialty: Vascular Surgery	0.3604	0.3839	0.8811	0.3479
Depression at Baseline	-0.2088	0.1577	1.7524	0.1856
Tobacco Use at Baseline: At least some	-0.0525	0.0941	0.3110	0.5770
Alcohol: Low	-0.0124	0.0936	0.0175	0.8949
Alcohol: High	0.0659	0.0898	0.5378	0.4633
Depression Age	-0.0008	0.0069	0.0124	0.9115
Disability (Yes)	-0.1817	0.0220	67.9450	**<0.0001**
Device: Apple Watch	0.2810	0.3570	0.6193	0.4313
Device: Fitbit	0.3334	0.3641	0.8381	0.3599
Device: Garmin	0.1256	0.3767	0.1111	0.7389
Last Measured Sleep Duration in Minutes	-0.0004	0.0001	9.2393	**0.0024**

Covariate	Estimate	SE	Z	P-value
Time of Since Most Recent Event†	-0.0093	0.0010	94.6185	**<0.0001**
Average Time Since Last Event†	0.0200	0.0010	375.7507	**<0.0001**
Average Cycle Duration†	-0.0135	0.0015	82.3069	**<0.0001**
Number Of Events Observed†	-0.2712	0.0086	989.0799	**<0.0001**
Average Time Recovered†	0.0131	0.0007	396.4536	**<0.0001**
Average Time Depressed†	0.0127	0.0053	5.6517	**0.0174**

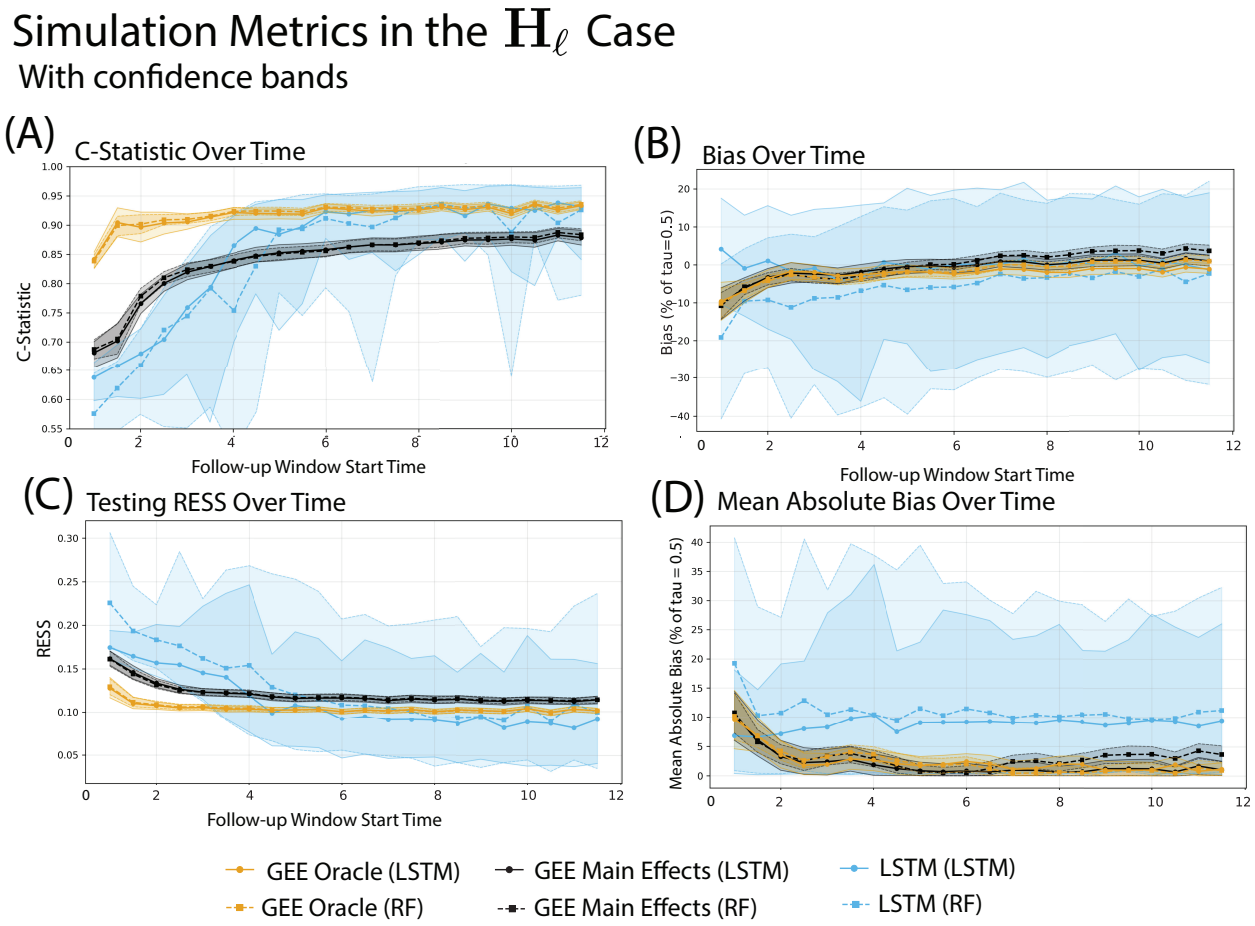


Figure C.2: A panel graphic displaying the performance of the dynamic prediction models under consideration in Section 4.7 for the $\boldsymbol{H}_l$ setting. Confidence bands display the 97.5% percentile and 2.5% percentile respectively. Counter clockwise from upper left is (A) c-statistic over time, (B) percent bias (in terms of τ) over time, (C) testing root mean squared error over time, and (D) mean absolute bias over time. The method used to predict $E[\min\{T_i(t), \tau\}] = \bar{\mathbf{Z}}^\top \beta$ is listed first (yellow for GEE Oracle, black for GEE Main Effects, and blue for LSTM), while the weight model is listed in parenthetical (with LSTM weights graphed in a solid line, while random forest weights are dashed). All results are from 200 replicates with $n = 2500$ each.

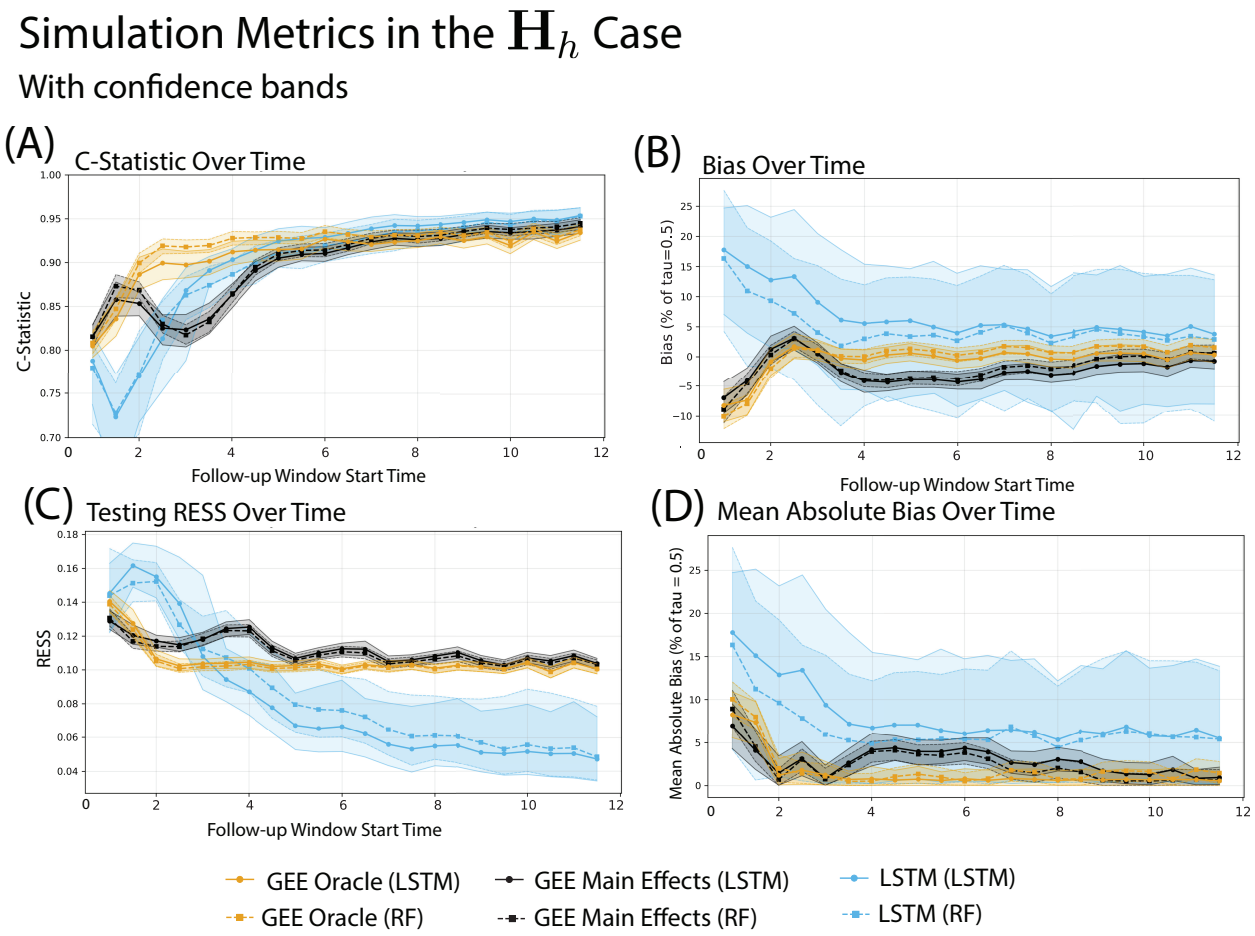


Figure C.3: A panel graphic displaying the performance of the dynamic prediction models under consideration in Section 4.7 for the $\boldsymbol{H}_h$ setting. Confidence bands display the 97.5% percentile and 2.5% percentile respectively. Counter clockwise from upper left is (A) c-statistic over time, (B) percent bias (in terms of τ) over time, (C) testing root mean squared error over time, and (D) mean absolute bias over time. The method used to predict $E[\min\{T_i(t), \tau\}] = \bar{\mathbf{Z}}^\top \beta$ is listed first (yellow for GEE Oracle, black for GEE Main Effects, and blue for LSTM), while the weight model is listed in parenthetical (with LSTM weights graphed in a solid line, while random forest weights are dashed). All results are from 200 replicates with $n = 2500$ each.

BIBLIOGRAPHY

Werner Adler, Sergej Potapov, and Berthold Lausen. Classification of repeated measurements data using tree-based ensemble methods. Computational Statistics, 26:355–369, Mar 2011. URL https://link.springer.com/content/pdf/10.1007/s00180-011-0249-1.pdf.

Richard K. Albert, John Connett, William Bailey, Richard Casaburi, Allen Cooper, Gerard Criner, Jeffrey Curtis, Mark Dransfield, MeiLan Han, Stephen Lazarus, Barry Make, Nathaniel Marchetti, Fernando Martinez, Nancy Madinger, Charlene McEvoy, Dennis Niewoehner, Janos Porsasz, Connie Price, John Reily, Paul Scanlon, Frank Sciurba, Steven Scharf, George Washko, Prescott Woodruff, and Nicholas Anthonisen. Azithromycin for prevention of exacerbations of copd. New England Journal of Medicine, 365(8):689–698, 2011.

Per Kragh Andersen and Richard D. Gill. Cox's regression model for counting processes: A large sample study. The Annals of Statistics, 10(4):1100–1120, 1982. ISSN 00905364. URL http://www.jstor.org/stable/2240714.

Per Kragh Andersen, John P Klein, and Susanne Rosthøj. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. Biometrika, 90 (1):15–27, 2003.

Per Kragh Andersen, Mette Gerster Hansen, and John P Klein. Regression analysis of restricted mean survival time based on pseudo-observations. Lifetime Data Analysis, 10 (4):335–350, 2004.

Adin-Cristian Andrei. Nonparametric and semiparametric censored survival analysis of correlated times to event and their sequelae. ProQuest Dissertations, 2005.

Adin-Cristian Andrei and Susan Murray. Regression models for the mean of the quality-of-life-adjusted restricted survival time using pseudo-observations. Biometrics, 63:398–404, Jun 2007. URL https://hdl.handle.net/2027.42/91895.

YC Ang, L Meyerson, YQ Tang, and N Qian. Statistical methods for the analysis of relapse data in m.s. clinical trials. Journal of Neurological Sciences, 285(1-2):206–211, Aug 2009. doi: 10.1016/j.jns.2009.07.017. URL https://pubmed.ncbi.nlm.nih.gov/19709676/.

J Scott Armstrong and Michael C Grohman. A comparative study of methods for long-range market forecasting. Management Science, 19(2):211–221, 1972.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. Advances in Neural Information Processing Systems, 37:107547–107603, 2024.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 5(2):157–166, 1994.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning, 2018. URL https://arxiv.org/abs/1606.04838.

Leo Breiman. Random forests. Machine learning, 45(1):5–32, 2001.

Leo Breiman, Jerome Friedman, Richard A Olshen, and Charles J Stone. Classification and regression trees. Chapman and Hall/CRC, 2017.

Carles Bretó, Daihai He, Edward L Ionides, and Aaron A King. Time series analysis via mechanistic models. The Annals of Applied Statistics, pages 319–348, 2009.

Jianwen Cai and Douglas E Schaubel. Marginal means/rates models for multiple type recurrent event data. Lifetime Data Analysis, 10(2):121–138, 2004.

Louis Capitaine, Robin Genuer, and Rodolphe Thiébaut. Random forests for high-dimensional longitudinal data. Statistical methods in medical research, 30(1):166–184, 2021.

Noelle E Carlozzi, Jonathan P Troost, Srijan Sen, Sung Won Choi, Zhenke Wu, Jennifer A Miner, Wendy L Lombard, Christopher Graves, and Angelle M Sander. Improving outcomes for care partners of individuals with traumatic brain injury: Results for a mhealth randomized control trial of the careqol app. Archives of Physical Medicine and Rehabilitation, 106(4):548–561, 2025.

Sandra Castro-Pearson, Aparajita Sur, Chi Hyun Lee, Chiung-Yu Huang, and Xianghua Luo. Bivrec: an r package for the nonparametric and semiparametric analysis of bivariate alternating recurrent events. BMC Medical Research Methodology, 22(1):92, 2022.

Kwun Chuen Gary Chan and Mei-Cheng Wang. Semiparametric modeling and estimation of the terminal behavior of recurrent marker processes before failure events. Journal of the American Statistical Association, 112(517):351–362, 2017. ISSN 01621459. URL http://www.jstor.org/stable/45027924.

Xingran Chen, Cheng-Han Yang, Zhenke Wu, and Bhramar Mukherjee. Prediction-based inference in electronic health record (ehr)-linked biobanks, 2026. URL https://arxiv.org/abs/2603.14356.

Sangbum Choi, Xuelin Huang, Hyunsu Ju, and Jing Ning. Semiparametric accelerated intensity models for correlated recurrent and terminal events. Statistica Sinica, 27(2): 625–643, 2017. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/26383293.

Richard J Cook, Jerald F Lawless, et al. The statistical analysis of recurrent events. Springer, 2007.

GC Donaldson, TA Seemungal, A Bhowmik, and JA Wedzicha. Relationship between exacerbation frequency and lung function in chronic obstructive pulmonary disease. Thorax, 57(10):847–852, October 2002. doi: 10.1136/thorax.57.10.847.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research, 12(7), 2011.

Anne Eaton, Yifei Sun, James Neaton, and Xianghua Luo. Nonparametric estimation in an illness-death model with component-wise censoring. Biometrics, 78(3):1168–1180, 2022.

Yu Fang, Amy SB Bohnert, Karina Pereira-Lima, Jennifer Cleary, Elena Frank, Zhuo Zhao, Walter Dempsey, and Srijan Sen. Trends in depressive symptoms and associated factors during residency, 2007 to 2019: a repeated annual cohort study. Annals of Internal Medicine, 175(1):56–64, 2022.

Charles Fefferman et al. Reconstructing a neural net from its output. Revista Matemática Iberoamericana, 10(3):507–556, 1994.

JP Fine, J Yan, and MR Kosorok. Temporal process regression. Biometrika, 91(3):683–703, 2004.

Guillaume Gerber, Yohann Le Faou, Olivier Lopez, and Michael Trupin. The impact of churn on client value in health insurance, evaluation using a random forest under various censoring mechanisms. Journal of the American Statistical Association, 116(536):2053–2064, 2021.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning, volume 1. MIT press Cambridge, 2016.

Alex Graves. Generating sequences with recurrent neural networks, 2014. URL https://arxiv.org/abs/1308.0850.

Constance Guille, Heather Speller, Rachel Laff, C Neill Epperson, and Srijan Sen. Utilization and barriers to mental health services among depressed medical interns: a prospective multisite study. Journal of Graduate Medical Education, 2(2):210–214, 2010.

R.F. Gunst and J.T. Webster. Regression analysis and problems of multicollinearity. Communications in statistics, 4(3):277–292, 1975. doi: 10.1080/03610927308827246.

Frank E. Harrell Jr, Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the Yield of Medical Tests. Journal of the American Medical Association, 247 (18):2543–2546, 05 1982. ISSN 0098-7484. doi: 10.1001/jama.1982.03320430047030. URL https://doi.org/10.1001/jama.1982.03320430047030.

Johannes Hengelbrock, Johanna Gillhaus, Sebastian Kloss, and Friedhelm Leverkus. Safety data from randomized controlled trials: applying models for recurrent events. Pharmaceutical Statistics, 15(4):315–323, 2016. doi: https://doi.org/10.1002/pst.1757. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pst.1757.

Sepp Hochreiter and Jürgen Schmidhuber. Lstm can solve hard long time lag problems. Advances in neural information processing systems, 9, 1996.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

Jianchang Hu and Silke Szymczak. A review on longitudinal data analysis with random forest. Briefings in Bioinformatics, 2(24):1–11, 2023. doi: https://doi.org/10.1093/bib/bbad002.

X Joan Hu, Maria Lorenzi, John J Spinelli, S Celes Ying, and Mary L McBride. Analysis of recurrent events with non-negligible event duration, with application to assessing hospital utilization. Lifetime Data Analysis, 17(2):215–233, 2011.

Edward L Ionides, Carles Bretó, and Aaron A King. Inference for nonlinear dynamical systems. Proceedings of the National Academy of Sciences, 103(49):18438–18443, 2006.

Hemant Ishwaran, Eugene H Blackstone, Claire E Pothier, and Michael S Lauer. Relative risk forests for exercise heart rate recovery as a predictor of mortality. Journal of the American Statistical Association, 99(467):591–600, 2004. doi: 10.1198/016214504000000638. URL https://doi.org/10.1198/016214504000000638.

Dae Hyun Kim, Hajime Uno, and Lee-Jej Wei. Restricted mean survival time as a measure to interpret clinical trial results. JAMA cardiology, 2(11):1179–1180, 2017. doi: 10.1001/jamacardio.2017.2922.

Dae Hyun Kim, Sandra M Shi, Danette Carroll, Mehdi Najafzadeh, and Lee-Jen Wei. Restricted mean survival time versus conventional measures for treatment decision-making. Journal of the American Geriatrics Society (JAGS), 69(8):2282–2289, 2021. doi: 10.1111/jgs.17195.

Diederik P Kingma and Jimmy Lei Lei. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

M. Kong, S. Xu, S.M. Levy, and S. Datta. G.E.E. type inference for clustered zero-inflated negative binomial regression with application to dental caries. Computational Statistics and Data Analysis, 85:54–66, May 2016. doi: 10.1016/j.csda.2014.11.014. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4303594/.

Jooyoung Lee and Richard J Cook. Dependence modeling for multi-type recurrent events via copulas. Statistics in medicine, 38(21):4066–4082, 2019a.

Jooyoung Lee and Richard J Cook. On estimands arising from misspecified semiparametric rate-based analysis of recurrent episodic conditions. Statistics in Medicine, 38(25):4977–4998, 2019b.

Xiaoqi Li, Yong Chen, and Ruosha Li. A frailty model for recurrent events during alternating restraint and non-restraint time periods. Statistics in Medicine, 36(4):643–654, 2017.

Yimei Li, E Paul Wileyto, and Daniel F Heitjan. Modeling smoking cessation data with alternating states and a cure fraction using frailty models. Statistics in medicine, 29(6): 627–638, 2010.

Baosheng Liang, Xingwei Tong, Donglin Zeng, and Yuanjia Wang. Semiparametric regression analysis of repeated current status data. Statistica Sinica, 27(3):1079–1100, 2017. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/26383245.

Danyu Y Lin, Lee-Jej Wei, I Yang, and Zhiliang Ying. Semiparametric regression for the mean and rate functions of recurrent events. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 62(4):711–730, 2000.

Roderick J. A. Little and Donald B. Rubin. Statistical Analysis with Missing Data. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., second edition. edition, 2002.

Abigail Loe, Susan Murray, and Zhenke Wu. Random forest for dynamic risk prediction of recurrent events: a pseudo-observation approach. Biostatistics, 26(1):1–20, 2025a.

Abigail Loe, Susan Murray, and Zhenke Wu. A random forest inverse probability weighted pseudo-observation framework for alternating recurrent events, 2025b. URL https://arxiv.org/abs/2510.23764.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30, 2017.

Lu Mao and D Y Lin. Semiparametric regression for the weighted composite endpoint of recurrent and terminal events. U.S. National Library of Medicine, Apr 2016. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4804115/.

Yassin Mazroui, Simone Mathoulin-Pélissier, Gaetan MacGrogan, Véronique Brouste, and Virginie Rondeau. Multivariate frailty models for two types of recurrent events with a dependent terminal event: application to breast cancer data. Biometrical Journal, 55(6): 866–884, 2013a.

Yassin Mazroui, Simone Mathoulin-Pélissier, Gaetan MacGrogan, Véronique Brouste, and Virginie Rondeau. Multivariate frailty models for two types of recurrent events with a dependent terminal event: application to breast cancer data. Biometrical Journal, 55(6): 866–884, 2013b.

Zachary R McCaw, Ariela R Orkaby, Lee-Jej Wei, DaeHyun Kim, and Michael W Rich. Applying evidence-based medicine to shared decision making: Value of restricted mean survival time. The American Journal of Medicine, 132, 2019.

Richard Meyes, Melanie Lu, Constantin Waubert de Puiseau, and Tobias Meisen. Ablation studies in artificial neural networks. CoRR, abs/1901.08644, 2019. URL http://arxiv.org/abs/1901.08644.

Tomas Mikolov and Geoffrey Zweig. Context dependent recurrent neural network language model. In 2012 IEEE Spoken Language Technology Workshop (SLT), pages 234–239, 2012. doi: 10.1109/SLT.2012.6424228.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. URL https://arxiv.org/abs/1301.3781.

Maja Miloslavsky, Sündüz Keleş, Mark J. van der Laan, and Steve Butler. Recurrent events analysis in the presence of time-dependent covariates and dependent censoring. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 66(1):239–257, 2004. ISSN 13697412, 14679868. URL http://www.jstor.org/stable/3647637.

Timothy NeCamp, Srijan Sen, Elena Frank, Maureen A Walton, Edward L Ionides, Yu Fang, Ambuj Tewari, and Zhenke Wu. Assessing real-time moderation for developing adaptive mobile health interventions for medical interns: micro-randomized trial. Journal of medical Internet research, 22(3):e15033, 2020.

Ganapati P Patil and Calyampudi R Rao. Weighted distributions and size-biased sampling with applications to wildlife populations and human families. Biometrics, pages 179–189, 1978.

Yulin Peng, Tao Chen, Fei Xiao, and Shaojie Zhang. Remaining useful lifetime prediction methods of proton exchange membrane fuel cell based on convolutional neural network-long short-term memory and convolutional neural network-bidirectional long short-term memory. Fuel Cells, 23(1):75–87, 2023.

Kaci L Pickett, Krithika Suresh, Kristen R Campbell, Scott Davis, and Elizabeth Juarez-Colunga. Random survival forests for dynamic predictions of a time-to-event outcome using a longitudinal biomarker. BMC medical research methodology, 21(1):216–230, 2021. doi: 10.1186/s12874-021-01375-x.

Lutz Prechelt. Automatic early stopping using cross validation: quantifying the criteria. Neural networks, 11(4):761–767, 1998.

Mustafa Radha, Pedro Fonseca, Arnaud Moreau, Marco Ross, Andreas Cerny, Peter Anderer, Xi Long, and Ronald M Aarts. Sleep stage classification from heart-rate variability using long short-term memory neural networks. Scientific reports, 9(1):14149, 2019.

C Radhakrishna Rao. On discrete distributions arising out of methods of ascertainment. Sankhyā: The Indian Journal of Statistics, Series A, pages 311–324, 1965.

Herbert Robbins and Sutton Monro. A stochastic approximation method. The Annals of Mathematical Statistics, pages 400–407, 1951.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association, 90(429):106–121, 1995.

Patrick Royston and Mahesh KB Parmar. 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(1):152, 2013.

Donald B. Rubin. Multiple imputation for nonresponse in surveys. Wiley series in probability and mathematical statistics. Applied probability and statistics, 0271-6232. Wiley, 1987.

Donald B Rubin. Multiple imputation after 18+ years. Journal of the American Statistical Association, 91(434):473–489, 1996.

Abdollah Safari, John Petkau, Mark J FitzGerald, and Mohsen Sadatsafavi. A parametric model to jointly characterize rate, duration, and severity of exacerbations in episodic diseases. BMC Medical Informatics and Decision Making, 23(1):1–11, 2023.

Saeid Safiri, Kristin Carson-Chahhoud, Maryam Noori, Seyed Aria Nejadghaderi, Mark J M Sullman, Javad Ahmadian Heris, Khalil Ansarin, Mohammad Ali Mansournia, Gary S Collins, Ali-Asghar Kolahi, and Jay S Kaufman. Burden of chronic obstructive pulmonary disease and its attributable risk factors in 204 countries and territories, 1990-2019: results from the global burden of disease study 2019. BMJ, 378:e069679, 2022. doi: 10.1136/bmj-2021-069679. URL https://www.bmj.com/content/378/bmj-2021-069679.

Erwan Scornet. Tuning parameters in random forests. ESAIM: Proceedings and Surveys, 60:144–162, 2018. doi: 10.1051/proc/201760144.

TA Seemungal, GC Donaldson, EA Paul, DJ Jeffries, and JA Wedzicha. Effect of Exacerbation on quality of life in patients with chronic obstructive pulmonary disease. American Journal of Critical Care Medicine, 157(5):1418–22, May 1998.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. The journal of Machine Learning Research, 15(1):1929–1958, 2014.

BaoLuo Sun and Eric J Tchetgen Tchetgen. On inverse probability weighting for nonmonotone missing at random data. Journal of the American Statistical Association, 113(521):369–379, 2018.

Héctor J Sussmann. Uniqueness of the weights for minimal feedforward nets with a given input-output map. Neural networks, 5(4):589–593, 1992.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks, 2014. URL https://arxiv.org/abs/1409.3215.

Leonard J Tashman. Out-of-sample tests of forecasting accuracy: an analysis and review. International Journal of Forecasting, 16(4):437–450, 2000.

Nabihah Tayob and Susan Murray. Nonparametric tests of treatment effect based on combined endpoints for mortality and recurrent events. Biostatistics, 16(1):73–83, January 2015. doi: 10.1093/biostatistics/kxu013.

Hajime Uno, Brian Claggett, Lu Tian, Eisuke Inoue, Paul Gallo, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. Journal of Clinical Oncology, 32, 2014.

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. Journal of Statistical Software, 45(3):1–67, 2011. doi: 10.18637/jss.v045.i03.

Hannah Van Wyk, Andrew F Brouwer, Gwenyth O Lee, Sully Márquez, Paulina Andrade, Edward L Ionides, Josefina Coloma, and Joseph NS Eisenberg. Early detection of dengue outbreaks: transmission model analysis of a dengue outbreak in a remote setting in ecuador. Epidemiology, 36(5):636–645, 2025.

Yehuda Vardi. Nonparametric estimation in the presence of length bias. The Annals of Statistics, 10(2):616–620, 1982.

Verner Vlačić and Helmut Bölcskei. Neural network identifiability for a family of sigmoidal nonlinearities. Constructive Approximation, 55(1):173–224, 2022.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839. URL https://doi.org/10.1080/01621459.2017.1319839.

Changfeng Wang, Santosh Venkatesh, and J Judd. Optimal stopping and effective machine complexity in learning. Advances in neural information processing systems, 6, 1993.

Jitao Wang, Yu Fang, Elena Frank, Maureen A Walton, Margit Burmeister, Ambuj Tewari, Walter Dempsey, Timothy NeCamp, Srijan Sen, and Zhenke Wu. Effectiveness of gamified team competition as mhealth intervention for medical interns: a cluster micro-randomized trial. npj Digital Medicine, 6(1):4, 2023.

Lili Wang, Kevin He, and Douglas E Schaubel. Penalized survival models for the analysis of alternating recurrent event data. Biometrics, 76(2):448–459, 2020.

Matthew R. Williams and Terrance D. Savitsky. Uncertainty estimation for pseudo-bayesian inference under complex sampling. International statistical review, 89(1):72–107, 2021. doi: 10.1111/insr.12376.

Meng Xia and Susan Murray. Commentary on Tayob and Murray (2014) with a useful update pertaining to study design. Biostatistics, 20(3):542—-545, July 2019. doi: https://doi.org/10.1093/biostatistics/kxy051.

Meng Xia, Susan Murray, and Nabihah Tayob. Regression analysis of recurrent-event-free time from multiple follow-up windows. Statistics in Medicine, 39:1–15, January 2020. doi: https://doi.org/10.1002/sim.8385.

Gongjun Xu, Sy Han Chiou, Chiung-Yu Huang, Mei-Cheng Wang, and Jun Yan. Joint scale-change models for recurrent events and failure time. Journal of the American Statistical Association, 112(518):794–805, 2017. ISSN 01621459. URL http://www.jstor.org/stable/45028409.

Jun Yan and Jason P Fine. Analysis of episodic data with application to recurrent pulmonary exacerbations in cystic fibrosis patients. Journal of the American Statistical Association, 103(482):498–510, 2008.

I. Yokota and Y. Matsuyama. Dynamic prediction of repeated events data based on landmarking model: application to colorectal liver metastases data. BMC Medical Research Methodology, 19(31):1–11, Feb 2019. URL https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-019-0677-0.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. arXiv, 2014. doi: 10.48550/arxiv.1409.2329.

Ruben Zazo, Alicia Lozano-Diez, Javier Gonzalez-Dominguez, Doroteo T. Toledano, and Joaquin Gonzalez-Rodriguez. Language identification in short utterances using long short-term memory (lstm) recurrent neural networks. PloS one, 11(1):e0146917, 2016.

Lili Zhao and Dai Feng. Deep neural networks for survival analysis using pseudo values. IEEE Journal of Biomedical and Health Informatics, 24(11):3308–3314, 2020.

Lili Zhao, Susan Murray, Laura H Mariani, and Wenjun Ju. Incorporating longitudinal biomarkers for dynamic risk prediction in the era of big data: A pseudo-observation approach. Statistics in Medicine, 39(26):3685–3699, 2020.

Hong Zhu. Non-parametric analysis of gap times for multiple event data: An overview. International Statistical Reviw/Revue Internationale de Statistque), 82(1):106–122, 2014.

Daniella A Zipkin, Craig Z Umscheid, Nancy L Keating, Elizabeth Allen, KoKo Aung, Rebecca Beyth, Scott Kaatz, Devin M Mann, Jeremy B Sussman, Deborah Korenstein, Connie Schardt, Avishek Nagi, Richard Sloane, and David A Feldstein. Evidence-based risk communication: A systematic review. Annals of Internal Medicine, 161:270–280, 2014.