Published in final edited form as:
*Statistics in Medicine* (2019), **38**(25): 4977–4998
DOI: 10.1002/sim.8345

# On estimands arising from misspecified semiparametric rate-based analysis of recurrent episodic conditions

JOOYOUNG LEE

*Department of Statistics and Actuarial Science*,
*University of Waterloo, Waterloo, ON, N2L 3G1, Canada*
*E-mail: j463lee@uwaterloo.ca*

RICHARD J. COOK

*Department of Statistics and Actuarial Science*,
*University of Waterloo, Waterloo, ON, N2L 3G1, Canada*

**Summary**

Marginal rate-based analyses are widely used for the analysis of recurrent events in clinical trials. In many areas of application, the events are not instantaneous but rather signal the onset of a symptomatic episode representing a recurrent infection, respiratory exacerbation, or bout of acute depression. In rate-based analyses, it is unclear how to best handle the time during which individuals are experiencing symptoms and hence are not at risk. We derive the limiting value of theNelson-Aalen estimator and estimators of the regression coefficients under a semiparametric rate-based model in terms of an underlying two-state process. We investigate the impact of the distribution of the episode durations, heterogeneity, and dependence on the asymptotic and finite sample properties of standard estimators. We also consider the impact of these features on power in trials designed to test intervention effects on rate functions. An application to a trial of individuals with herpes simplex virus is given for illustration.

*Keywords*: estimands, heterogeneity, intensity function, misspecification, rate function, recurrent episodes

# 1 INTRODUCTION

Many chronic diseases involve the recurrent onset and resolution of episodes during which individuals are in an adverse health state. Examples include recurrent exacerbations in chronic bronchitis (Grossman et al., 1998), recurrent bouts of acute depression in affective disorder (Kessing et al., 1999), and recurrent outbreaks of symptoms among individuals with herpes simplex virus infection (Romanowski et al., 2003). Common statistical methods for recurrent event analysis are geared toward the analysis of instantaneous events and include methods based on the semiparametric Andersen-Gill model (Andersen et al., 1993), multiplicative models involving rate or mean functions (Lawless and Nadeau, 1995; Lin et al., 2000), and frailty models (Lawless, 1987; Klein, 1992; Wienke, 2010). Such methods have seen widespread application in clinical trials involving recurrent episodic conditions where the "events" are taken to be the onset of the symptomatic periods (Hu et al., 2011). During the symptomatic periods, however, individuals are not truly at risk of the "event" since they are already symptomatic. It is unclear how to handle the risk-free periods in the recurrent event analyses. It is also unclear what impact any decision might have on inferences that follow. Options for handling these symptomatic episodes include (i) retaining individuals in the risk set during episodes, (ii) removing individuals from the risk set while they are experiencing an episode, or (iii) modeling the onset and duration times based on an alternating two-state model. Alternating renewal processes (Cox, 1967) are useful when the two types of sojourn times (waiting times between episodes and episode durations) can be assumed statistically independent. Several random effect (frailty) models have been developed to relax these independence conditions (Xue and Brookmeyer, 1996; Ng and Cook, 1997; Lee and Cook, 2018). Intensity-based two-state models offer another powerful approach for studying process dynamics, but they require conditioning on the process history and robust inference is not possible in this framework. Moreover, intensity-based analyses do not lead naturally to estimates of average causal treatment effects (Hernán and Robins, 2016), which are typically of interest in clinical trials.

Marginal methods based on partially conditional rate functions are increasingly used for the analysis of recurrent outcomes in recent years. Although such methods can be robust to misspecification of the variance function or dependence structure for point processes, they do not protect against misspecification of the risk set. The objective of this article is to study the asymptotic and finite sample properties of estimators from marginal rate-based recurrent event analyses (Lin et al., 2000) for two of the common approaches taken for handling the risk-free period. Upon specifying a quite general alternating two-state model for the onset and resolution of episodes, we study the limiting behavior of estimators from semiparametric rate-based analyses of the onset times of symptomatic periods.

The remainder of this paper is organized as follows. In Section 2, we define notation and intensity functions for an alternating two-state process that we use in our investigation of the consequences of risk-set misspecification. In Section 3, we review the formulation, estimating equations, and large sample results for estimators from a semiparametric multiplicative rate-based analysis. The effect of model misspecification on the limiting behavior of estimators is investigated in Section 4 for both the one-sample problem and the regression setting. Section 4.1 considers the setting where the data are generated according to a two-state process without any between-individual heterogeneity in the process intensities, whereas Section 4.2 considers a more general data generating process incorporating heterogeneity in risk for the onset and duration of exacerbations; a dependence between associated random effects is also accommodated. We study the implications of model misspecification due to failure to account for episode duration on study power for clinical trials in Section 5. An application to a randomized trial of individuals with herpes simplex virus infection is given in Section 6 and concluding remarks are made in Section 7.

## 2 AN ALTERNATING TWO-STATE PROCESS

In this section, we introduce a two-state data generating process which we use to study the limiting behavior of estimators from semiparametric rate-based analyses when they are applied to recurrent episodic conditions.

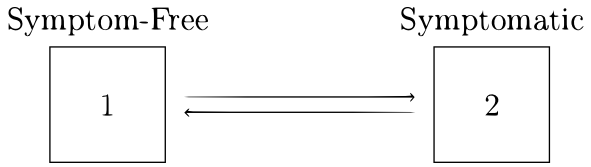



Figure 1: A two-state diagram for a chronic disease featuring recurrent symptomatic episodes

Suppose an individual with a chronic disease experiences recurrent symptomatic episodes arising according to a two-state model depicted in Figure 1. Let $Z_i(s) = 1$ if individual $i$ is symptom-free and $Z_i(s) = 2$ if they are symptomatic at $s > 0$, and suppose all individuals start in state 1 at time $t = 0$. We let $S_{ik}$ and $T_{ik}$ denote the start (onset) and termination (resolution) time of the $k$th episode for individual $i$ which is of duration $W_{ik} = T_{ik} - S_{ik}$, $k = 1, \ldots$. A schematic of a hypothetical sample path is given in Figure 2. Let $N_{i1}(t) = \sum_{k=1}^{\infty} I(S_{ik} \leq t)$ and let $N_{i2}(t) = \sum_{k=1}^{\infty} I(T_{ik} \leq t)$ record the cumulative number of onset and resolution times over $(0, t]$, respectively, and $\{N_i(s), 0 < s\}$ be a bivariate counting process with $N_i(s) = (N_{i1}(s), N_{i2}(s))'$. If $X_i$ is a set of fixed covariates, the history of the process is denoted by $H_i(t) = \{N_i(s), 0 < s < t, X_i\}$.

Consider a trial with the goal of observing individuals over a fixed period $(0, A]$ where $A$ is a common administrative censoring time. A random drop-out time $D_i$ for individual $i$ is assumed to be independent of the event process $\{N_i(s), 0 < s\}$ given covariates $X_i$, and $C_i = \min(A, D_i)$ is the right censoring time. We let $Y_i(s) = I(s \leq C_i)$ and $\bar{Y}_{ij}(s) = Y_i(s)Y_{ij}(s)$ where $Y_{ij}(s) = I(Z_i(s^-) = j)$, $j = 1, 2$. Then letting $\bar{N}_{ij}(t) = \int_0^t \bar{Y}_{ij}(s)dN_{ij}(s)$, the observed bivariate counting process is $\{\bar{N}_i(s), 0 < s\}$ where $\bar{N}_i(t) = (\bar{N}_{i1}(t), \bar{N}_{i2}(t))'$. The complete history of the observation and event processes is then denoted by $\bar{H}_i(t) = \{\bar{N}_i(s), Y_i(s), 0 < s < t, X_i\}$ and the complete intensity for $j \to 3 - j$ transitions is

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta \bar{N}_{ij}(t) = 1 | \bar{H}_i(t))}{\Delta t} = \bar{Y}_{ij}(t)\lambda_{ij}(t|H_i(t)), \qquad j = 1, 2, \tag{1}$$

under independent censoring (Cook and Lawless, 2018).

The probability of a particular sample path (Cook and Lawless, 2007) for individual $i$ is

$$\prod_{k=1}^{N_{i1}(C_i)} \lambda_{i1}(s_{ik}|H_i(s_{ik})) \exp\left(-\int_0^{C_i} \bar{Y}_{i1}(u)\lambda_{i1}(u|H_i(u))du\right)$$
$$\times \prod_{l=1}^{N_{i2}(C_i)} \lambda_{i2}(t_{il}|H_i(t_{il})) \exp\left(-\int_0^{C_i} \bar{Y}_{i2}(u)\lambda_{i2}(u|H_i(u))du\right)$$

conditional on the censoring time. While likelihoods can be constructed based on such expressions our interest lies in robust assessment of treatment effects in clinical trials with the goal of preventing the onset of episodic symptomatic periods; in such settings, the aim is to reduce the occurrence of $1 \to 2$ transitions in Figure 1 rather than model the full process. We also note that intensity-based methods are less amenable to the assessment of randomized interventions in clinical trials since they involve conditioning on internal features of life history processes (Kalbfleisch and Prentice, 2011). We emphasize therefore that the two-state model is described here in order to provide a basis for

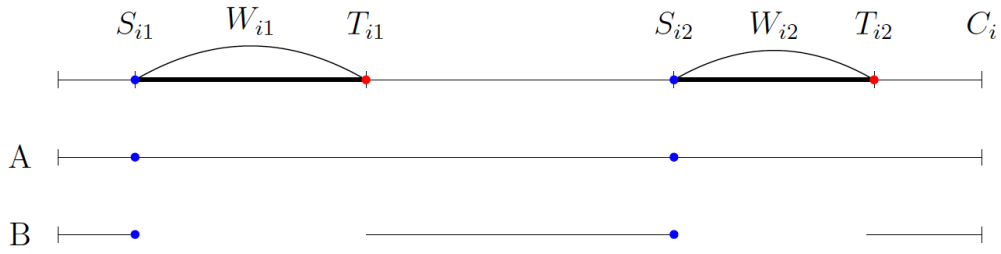



Figure 2: A schematic of a hypothetical timeline diagram with risk set definition (RSD) A and B

studying the limiting behaviour of estimators from semiparametric rate-based analyses commonly used in clinical trials.

In what follows, we consider two risk-set definitions depicted in Figure 2. In risk set definition A (RSD-A), individuals are included in the risk set for transitions from state 1 to state 2 during symptomatic periods (i.e. we set the at risk indicator to $\bar{Y}_i^A(t) = Y_i(t)$); this represents a misspecification since individuals in the midst of an episode are not at risk for the onset of an episode. In risk set definition B (RSD-B), individuals are not considered at risk during symptomatic periods (i.e. we set $\bar{Y}_i^B(t) = \bar{Y}_{i1}(t) = Y_i(t)Y_{i1}(t)$). In many ways, RSD-B seems sensible since it is aligned with how these periods are treated in the intensity-based analysis described earlier. Rate-based analyses in randomized clinical trials, however, are best directed at estimation of marginal features, and exclusion of individuals from the risk set based on their status after randomization (which itself is potentially influenced by the treatment received) induces confounding (Cook and Lawless, 2018). In causal inference terminology, the state of being "at risk for the onset of a new episode" is a collider (i.e. in the causal path) for the effect of treatment on the occurrence of exacerbations (Cole et al., 2009; Hernán and Robins, 2016). Thus, while excluding individuals from the risk set during episodes has intuitive appeal, it precludes the ability to make direct causal statements about intervention effects on marginal features. Retaining individuals in the risk set during episodes is unnatural since they are not really at risk, but this enables one to make causal inferences. These points motivate us to explore the limiting behaviour of estimators obtained from the two approaches for defining the risk sets in marginal rate-based analyses to gain insight into the determinants of the resulting estimands. We carry out this investigation in the context of the underlying two-state process of this section and a generalization of it we give in Section 4.2.

## 3 Standard Recurrent Event Analyses

### 3.1 A Semiparametric Multiplicative Rate Function Model

The semiparametric marginal rate-based model (Andersen and Gill, 1982) involves the assumption that covariates act multiplicatively a baseline rate function. Here, we temporarily consider a setting where $\lambda_{i2}(t|H_i(t)) \to \infty$ so the resulting data can be viewed as arising from a point process. We write the multiplicative model assumption as

$$E(dN_{i1}(t)|X_i = x_i) = dR_{i1}(t) = dR_{01}(t)\exp(x_i'\gamma_1)\,, \tag{2}$$

where the baseline rate function $dR_{01}(t)$ has no specific parametric form. With data $\{Y_i(s), d\bar{N}_{i1}(s), 0 < s; X_i, i = 1, \ldots, m\}$ from a sample of $m$ independent individuals, the estimating functions are

$$\sum_{i=1}^{m} Y_i(t)\{dN_{i1}(t) - dR_{i1}(t)\} = 0\,, \quad t > 0\,, \tag{3}$$

for the baseline rate function and

$$\sum_{i=1}^{m} \int_0^{\infty} Y_i(t)\{dN_{i1}(t) - dR_{i1}(t)\}x_i = 0\,, \tag{4}$$

for the regression coefficients. Solving (3) with fixed $\gamma_1$ gives the profile "Breslow" estimate

$$d\tilde{R}_{01}(t;\gamma_1) = \frac{\sum_{i=1}^m Y_i(t)dN_{i1}(t)}{\sum_{i=1}^m Y_i(t)\exp(x_i'\gamma_1)}\,, \tag{5}$$

and substituting (5) into (4) gives an estimating function for $\gamma_1$ as

$$U(\gamma_1) = \sum_{i=1}^{m} \int_0^{\infty} Y_i(s)\left\{x_i - \frac{\sum_{i=1}^m Y_i(t)\exp(x_i'\gamma_1)x_i}{\sum_{i=1}^m Y_i(t)\exp(x_i'\gamma_1)}\right\}dN_{i1}(t)\,. \tag{6}$$

We obtain $\hat{\gamma}_1$ by solving $U(\gamma_1) = 0$. The baseline mean function is given by $R_{01}(t) = \int_0^t dR_{01}(s)ds = E\{N_{i1}(t)|x_i = 0\}$ and estimated by substituting $\hat{\gamma}_1$ into (5) to compute

$$\hat{R}_{01}(t) = \int_0^t \frac{\sum_{i=1}^m Y_i(u)dN_{i1}(u)}{\sum_{i=1}^m Y_i(u)\exp(x_i'\hat{\gamma}_1)}\,.$$

These results correspond to the setting for which rate-based methods are intended, where the events are instantaneous and so have no duration associated with them. Under the assumption of multiplicative covariate effects on a baseline rate when events are instantaneous, robust inferences are possible if "sandwich" variance estimates are used (Lawless and Nadeau, 1995; Lin et al., 2000); see Section 3.2 .

To distinguish between the estimating equations based on the different risk set definitions, we can replace $Y_i(t)$ with $\bar{Y}_i^A(t)$ in this derivation and denote the resulting estimates as $\hat{\gamma}_1^A$ and $\hat{R}_{01}^A(t)$. Under the alternative risk set definition (RSD-B), we proceed in the same fashion by replacing $Y_i(t)$ with $\bar{Y}_i^B(t)$ in (3) and (4) and labelling the corresponding estimators $\hat{\gamma}_1^B$ and $\hat{R}_{01}^B(t)$. In what follows, we consider the interpretation of these limiting values (estimands) in a variety of settings.

### 3.2 Large Sample Behaviour under Model Misspecification

The estimating equations under (3) and (4) are justified based on the assumption that the events arise from a Poisson process but the resulting estimator is consistent for $\gamma_1$ more generally if the proportional rate function assumption is satisfied. This will not typically be the case when the events are onset times of episodes in a two-state process; we explore the large sample behaviour of the estimators in this setting here.

To unify the treatment of the different risk set definitions, we rewrite (6) with the risk set indicator denoted by $\bar{Y}_i^h(t)$ to obtain

$$U^h(\gamma_1) = \sum_{i=1}^{m} \int_0^{\infty} \bar{Y}_i^h(t)\left\{x_i - \frac{S^{(1,h)}(\gamma_1,t)}{S^{(0,h)}(\gamma_1,t)}\right\}dN_{i1}(t)\,, \tag{7}$$

with $S^{(k,h)}(\gamma_1,t) = \sum_{i=1}^m \bar{Y}_i^h(t)\exp(x_i'\gamma_1)x_i^{\otimes k}$ for $k = 0,1,2$, $h = A,B$, where $a^{\otimes 2}$ means $aa'$, $a^{\otimes 1} = a$, and $a^{\otimes 0}$ represents a scalar 1. We let $\hat{\gamma}_1^h$ be the solution to $U^h(\gamma_1) = 0$ and $\gamma_1^h$ its limiting value, which is the solution to

$$\int_0^{\infty} \left\{s^{(1,h)}(u) - \frac{s^{(1,h)}(\gamma_1,u)}{s^{(0,h)}(\gamma_1,u)}\, s^{(0,h)}(u)\right\} = 0 \tag{8}$$

where $s^{(k,h)}(\gamma_1, u) = E\{S^{(k,h)}(\gamma_1, u)\}$ and $s^{(k,h)}(u) = E\{\bar{Y}_i(u)X_i^{\otimes k}dN_{i1}(u)\}$, $k = 0, 1$, $h = A, B$, and all expectations are taken with respect to the underlying true model. We let

$$R_{01}^h(t) = \int_0^t \frac{E\{\bar{Y}_i^h(u)dN_{i1}(u)\}}{S^{(0,h)}(\gamma_1^h, u)} \tag{9}$$

be the limiting value of the corresponding baseline mean function estimate (Boher and Cook, 2006).

Note that if $E\{dN_{i1}(t)|x_i, \bar{Y}_i^h(t) = 1\} = dR_{01}^h(t)\exp(x_i'\gamma_1^h)$, then the estimating function in (7) can be written (Lin et al., 2000) as

$$U^h(\gamma_1) = \sum_{i=1}^m \int_0^\infty \left\{x_i - \frac{S^{(1,h)}(\gamma_1, t)}{S^{(0,h)}(\gamma_1, t)}\right\} dM_{i1}^h(t)$$

where $dM_{i1}^h(t) = \bar{Y}_i^h(t)\{dN_{i1}(t) - dR_{01}^h(t)\exp(x_i'\gamma_1^h)\}$. Since $\bar{Y}_i^h(t)$ is a predictable process (Andersen et al., 1993), $m^{-1/2}U^h(\gamma_1)$ is asymptotically $N(0, \mathcal{B}(\gamma_1))$ in distribution where

$$\mathcal{B}(\gamma_1) = E\left[\left(\int_0^\infty \left\{x_i - \frac{s^{(1,h)}(\gamma_1, s)}{s^{(0,h)}(\gamma_1, s)}\right\} dM_{i1}^h(s)\right)\left(\int_0^\infty \left\{x_i - \frac{s^{(1,h)}(\gamma_1, t)}{s^{(0,h)}(\gamma_1, t)}\right\}' dM_{i1}^h(t)\right)\right].$$

By Taylor series expansion, $m^{1/2}(\hat{\gamma}_1^h - \gamma_1^h) \simeq A^{-1}(\gamma_1^h)\, m^{-1/2}\, U(\gamma_1^h)$ so $m^{1/2}\,(\hat{\gamma}_1^h - \gamma_1^h)$ converges to $MVN(0, \mathcal{A}^{-1}(\gamma_1^h)\mathcal{B}(\gamma_1^h)\mathcal{A}^{-1}(\gamma_1^h))$ in distribution where

$$\mathcal{A}(\gamma_1) = E\left[\int_0^\infty \bar{Y}_i^h(t)\left\{\frac{s^{(2,h)}(\gamma_1, t)}{s^{(0,h)}(\gamma_1, t)} - \frac{s^{(1,h)}(\gamma_1, t)^{\otimes 2}}{s^{(0,h)}(\gamma_1, t)^2}\right\} dN_{i1}(t)\right].$$

The robust variance $\mathcal{A}^{-1}(\gamma_1^h)\mathcal{B}(\gamma_1^h)\mathcal{A}^{-1}(\gamma_1^h)$ is empirically estimated by $\hat{A}^{-1}(\hat{\gamma}_1^h)\hat{B}(\hat{\gamma}_1^h)\hat{A}^{-1}(\hat{\gamma}_1^h)$ in finite samples where

$$\hat{A}(\hat{\gamma}_1^h) = \frac{1}{m}\sum_{i=1}^m \left(\int_0^\infty \bar{Y}_i^h(t)\left\{\frac{S^{(2,h)}(\gamma_1, t)}{S^{(0,h)}(\gamma_1, t)} - \frac{S^{(1,h)}(\gamma_1, t)^{\otimes 2}}{S^{(0,h)}(\gamma_1, t)^2}\right\} dN_{i1}(t)\right)\Bigg|_{\gamma_1 = \hat{\gamma}_1^h},$$

$$\hat{B}(\hat{\gamma}_1^h) = \frac{1}{m}\sum_{i=1}^m \left(\int_0^\infty \left\{x_i - \frac{S^{(1,h)}(\gamma_1, t)}{S^{(0,h)}(\gamma_1, t)}\right\} d\hat{M}_{i1}^h(t)\right)\left(\int_0^\infty \left\{x_i - \frac{S^{(1,h)}(\gamma_1, t)}{S^{(0,h)}(\gamma_1, t)}\right\}' d\hat{M}_{i1}^h(t)\right)\Bigg|_{\gamma_1 = \hat{\gamma}_1^h},$$

and $d\hat{M}_{i1}^h(t) = \bar{Y}_i^h(t)\{dN_{i1}(t) - d\hat{R}_{01}^h(t)\exp(x_i'\hat{\gamma}_1^h)\}$ for $h = A, B$. In the analysis of sample data, the estimate

$$\widehat{asvar}(\sqrt{m}(\hat{\gamma}_1^h - \gamma_1^h)) = \hat{A}^{-1}(\hat{\gamma}_1^h)\,\hat{B}(\hat{\gamma}_1^h)\,\hat{A}^{-1}(\hat{\gamma}_1^h)$$

is used as a basis for inference.

## 4 Mean Function and Regression Coefficient Estimands

Here, we investigate the limiting properties of estimators under independent censoring when analysis is based on a marginal rate-based model. We consider settings involving a Markov/semi-Markov model for the onset and duration of recurrent episodes (Section 4.1) and a mixed model with dependent bivariate random effects modulating the baseline transition intensities (Section 4.2). Under each scenario, we obtain the asymptotic bias of the estimated mean function in a one-sample setting and then the regression coefficient of a treatment indicator under the model assuming multiplicative covariate effects.

### 4.1 Misspecification under a Markov/semi-Markov Data Generating Process

When the onset of exacerbations is governed by a Markov model, intensity (1) reduces to

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta \bar{N}_{i1}(t) = 1 | \bar{H}_i(t))}{\Delta t} = \bar{Y}_{i1}(t)\lambda_{i1}(t)\,, \tag{10}$$

where $\lambda_{i1}(t)$ depends only on $x_i$ and the time $t$ since the origin of the process. If the resolution of an episode is governed by a semi-Markov intensity, then

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta \bar{N}_{i2}(t) = 1 | \bar{H}_i(t))}{\Delta t} = \bar{Y}_{i2}(t)\lambda_{i2}(B_i(t))\,, \tag{11}$$

which, besides $\bar{Y}_{i2}(t)$, $\lambda_{i2}(t)$ depends only on $x_i$ and the time $B_i(t) = t - S_{N_{i1}(t^-)}$ since the onset of the episode. Hu et al. (2011) examined the asymptotic properties and proved consistency of estimators from analyses using RSD-B in this setting, so we focus here on RSD-A.

#### 4.1.1 Marginal Rate and Mean Function Estimation

We consider the setting with no covariates first in which case we let $\lambda_{i1}(t) = \lambda_1(t)$ in (10) and $\lambda_{i2}(t) = \lambda_2(B_i(t))$ in (11). We consider the rate function $E\{dN_{i1}(t)\} = dR_1(t)$ and a simplified (Cook and Lawless, 2007) version of (3) with at risk indicator $\bar{Y}_i^A(t)$:

$$\sum_{i=1}^{m} \bar{Y}_i^A(t)\{dN_{i1}(t) - dR_1(t)\}\,.$$

Taking the expectation of the contribution from individual $i$ under completely independent censoring, we obtain

$$E[\bar{Y}_i^A(t)\{dN_{i1}(t) - dR_1(t)\}] = E[\bar{Y}_i^A(t)\{P(Y_{i1}(t) = 1)\lambda_1(t)\,dt - dR_1(t)\}] = 0 \tag{12}$$

where the expectations and probabilities are computed based on the full model given in Section 2 under the assumptions in (10) and (11). Equation (12) has solution

$$dR_1^A(t) = P(Y_{i1}(t) = 1)\lambda_1(t)\,dt \tag{13}$$

and we note that the estimand $R_1^A(t) = \int_0^t dR_1^A(s)$ is the true mean function for the counting process $\{N_{i1}(u), 0 < u\}$. Hence, the standard Nelson-Aalen estimator with RSD-A is consistent for the cumulative mean function under independent right censoring (Lawless and Nadeau, 1995; Nelson, 1995). We note however that the estimator $\widehat{R}_1^A(t)$ will be asymptotically biased (conservatively) for the cumulative intensity (rate) function $\Lambda_1(t) = \int_0^t \lambda_1(s)ds$ and if $P(Y_{i1}(t) = 1)$ is small (i.e. if there is a high probability of being in the exacerbation state), the bias may be appreciable.

For illustration, we consider a particular parametric setting with $\lambda_1(t) = 2$ and independently and identically distributed gap times $W_{ik} \sim \text{Gamma}(2, \lambda_2)$ with $E(W_{ik}) = 2/\lambda_2$. We derive an expression of $P(Y_{i1}(t) = 1)$ in Appendix A, where we show that as $t \uparrow \infty$, the probability $P(Y_{i1}(t) = 1)$ converges to $\lambda_2/(2\lambda_1 + \lambda_2)$. Figure 3 displays the cumulative intensity $\Lambda_1(t) = 2t$ and the limiting value of the Nelson-Aalen estimator $R_1^A(t)$ and illustrates that the asymptotic bias becomes larger (more negative) with increasing $t$ (left panel) and as the mean sojourn time in the exacerbation state increases (right panel).

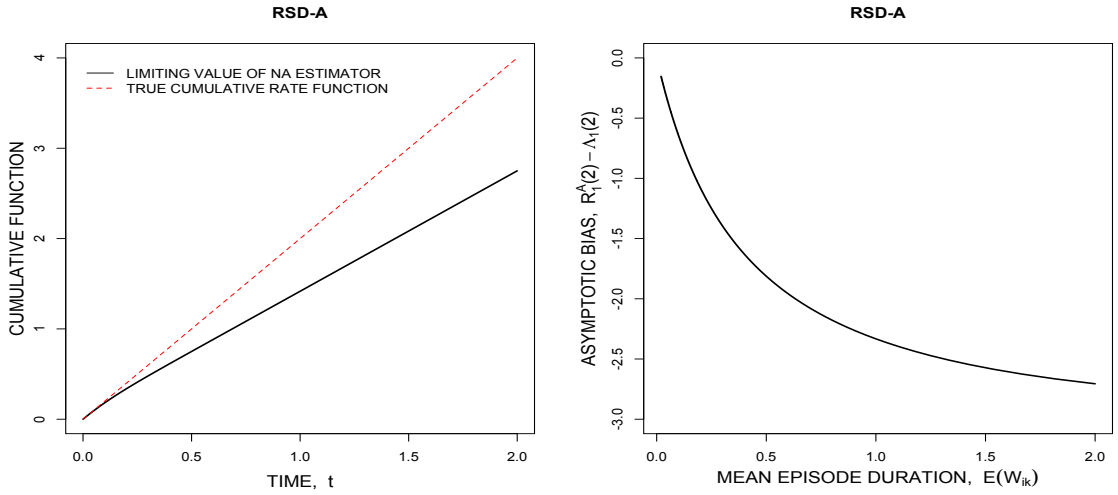



Figure 3: The limiting values and the asymptotic bias of Nelson-Aalen estimator under the RSD-A setting as a function of t with $E(W_{ik}) = 0.25$ (left panel) and as a function of $E(W_{ik})$ at t=2 (right panel) at fixed values of $\lambda_{01} = 2$, C=2, and 20% random censoring

### 4.1.2 Estimation in the Regression Setting

Here, we consider an underlying model given by (10) and (11) with $\lambda_{i1}(t) = \lambda_{01}(t)\exp(x_i\beta_1)$. Note that we use $\beta_1$ to represent the coefficient in the two-state model to distinguish it from the parameter $\gamma_1$ in the working rate-based model. Since the proportional rate model with RSD-A does not account for the duration of the exacerbation episodes, the limiting value of $\gamma_1^A$ will in general differ from $\beta_1$ and we explore this here.

We consider the setting of a randomized clinical trial where $X_i$ is a Bernoulli treatment indicator with $P(X_i = 1) = 0.5$. To compute the limiting value $\gamma_1^A$ based on (8), we need to evaluate the associated functions. Based on this formulation, $s^{(k,A)}(u) = E\{\bar{Y}_i^A(u)X_i^k dN_{i1}(u)\}$ and $s^{(k,A)}(\gamma_1, u) = E\{\bar{Y}_i^A(u)X_i^k \exp(X_i\gamma_1)\}$ for $k = 0, 1$. The explicit forms are

$$s^{(0,A)}(u) = \sum_{x=0}^{1} P(\bar{Y}_{i1}(u) = 1|X_i = x)P(X_i = x)\lambda_{01}(u)\,du\,\exp(x\beta_1) \tag{14}$$

and

$$s^{(1,h)}(u) = P(\bar{Y}_{i1}(u) = 1|X_i = 1)P(X_i = 1)\lambda_{01}(u)\,du\,\exp(\beta_1) \tag{15}$$

as well as

$$s_1^{(0,A)}(\gamma_1, u) = \sum_{x=0}^{1} P(Y_i(u) = 1)P(X_i = x)\exp(x\gamma_1)$$

and

$$s_1^{(1,A)}(\gamma_1, u) = P(X_i = 1)P(Y_i(u) = 1)\exp(\gamma_1)\,.$$

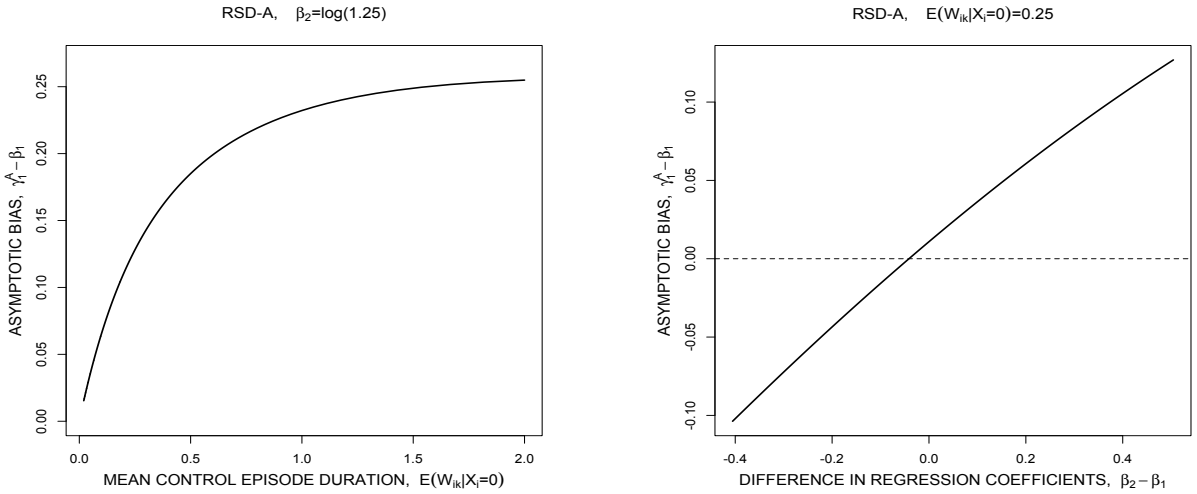
If we solve (8) based on this specification, we obtain

$$\gamma_1^A = \beta_1 + \log\left(\frac{\int_0^\infty P(\bar{Y}_{i1}(u) = 1|X_i = 1)\,du}{\int_0^\infty P(\bar{Y}_{i1}(u) = 1|X_i = 0)\,du}\right), \tag{16}$$

where the particular result will depend on further details of the model specification.

Again, we consider a particular parametric setting in more detail for illustration. We assume a time homogeneous a multiplicative intensity-based regression model with $\lambda_{i1}(t|x_i) = \lambda_{01}\exp(x_i\beta_1)$ for

the onset of episodes, and that $W_{ik}|x_i \sim \text{Gamma}(2, \lambda_{02}\exp(x_i\beta_2))$ where $E(W_{ik}|x_i) = 2/(\lambda_{02}\exp(x_i\beta_2))$ is the mean episode duration. In addition to the baseline functions, the magnitude of $\beta_1$ and $\beta_2$ will determine $\gamma_1^A$ and the robust covariance matrix $\mathcal{A}^{-1}(\gamma_1^A)\mathcal{B}(\gamma_1^A)\mathcal{A}^{-1}(\gamma_1^A)$; see Appendix B. Here, we set $\lambda_{01} = 2$, $\beta_1 = \log(0.75)$ and $\beta_2 = \log(1.25)$. We set $A = 2$ and take $D_i$ to be exponentially distributed to give 20% random censoring. Figure 4(a) shows that the asymptotic bias $\gamma_1^A - \beta_1$ increases as the mean duration of the exacerbation episodes increases. The decreasing and negligible asymptotic bias associated with shorter mean durations of episodes arises since this approaches the setting in which the events are instantaneous and the estimating functions are valid. The limiting bias is plotted as a function of $\beta_2 - \beta_1$ in Figure 4(b) in the setting where $E(W_{ik}|X_i = 0) = 0.25$, $\lambda_{02} = 20$ and $\beta_1 = \log(0.75)$. It is apparent that the sign of bias for the coefficients depends on the difference between $\beta_1$ and $\beta_2$. If treatment reduces the risk of an episode and shortens the duration of symptoms, the misspecification of the risk set will lead to an underestimation of the treatment effect on the true risk of symptomatic episodes.




(a) (b)

Figure 4: The asymptotic bias of a coefficient under the AG model with RSD-A as a function of $E(W_{ik}|X_i = 0)$ (panel a), and $\beta_2 - \beta_1$ (panel b) at fixed values of $\lambda_{01} = 2$, and $\beta_1 = \log(0.75)$ with administrative censoring time $A = 2$ and 20% random censoring

### 4.1.3 Simulation Studies

Simulation studies were conducted to examine the empirical bias and the performance of the robust variance estimator for rate-based analyses under misspecification of the risk set. The data are generated according to the two-state model of Section 2. We are primarily interested in the performance of the estimators from a marginal rate-based model with the original formulation (see Section 3.1) using RSD-A and RSD-B to correspond to the common *ad hoc* approach for dealing with the duration of the episodes. We let $X_i$ represent a treatment indicator which is Bernoulli distributed with $P(X_i = 1) = 0.5$. We set $\lambda_{01} = 2$ and $\beta_1 = \log(0.75)$, where $\lambda_{i1}(t) = \lambda_{01}\exp(x_i\beta_1)$, $E\{W_{ik}|X_i = 0\} = 0.1$, 0.25, or 0.5 where $W_{ik}|x_i \sim \text{Gamma}(2, \lambda_{02}\exp(x_i\beta_2))$. We set the administrative censoring time to $A = 2$ and generate an exponentially distributed drop-out time $D_i$ such that $P(D_i < A) = 0.20$; the net censoring time is $C_i = \min(A, D_i)$. We generate $S_{i1}|x_i$ as exponential with hazard $\lambda_{01}\exp(x_i\beta_1)$). If $S_{i1} > C_i$, then this individual completed their follow-up without

experiencing any exacerbation episodes; otherwise, we generate $W_{i1}|x_i \sim \text{Gamma}(2, \lambda_{02}\exp(x_i\beta_2))$ and compute $T_{i1} = S_{i1} + W_{i1}$ as the termination time of the first episode. If $T_{i1} > C_i$, then the process is terminated and the resolution time of the first episode for this subject is censored at $C_i$ and otherwise we take $T_{i1} = S_{i1} + W_{i1}$. For each $k > 1$ with $T_{i,k-1} < C_i$, we generate the start time for the $k$th episode as $S_{ik}|x_i, S_{ik} > T_{i,k-1}$ as exponential with rate $\lambda_{01}\exp(x_i\beta_1)$ and left-truncation time $T_{i,k-1}$. If $S_{ik} > C_i$, we censor the onset time of the $k$th episode at $C_i$ but if $S_{ik} < C_i$, we simulate the duration of the $k$th episode as $W_{ik}|x_i \sim \text{Gamma}(2, \lambda_{02}\exp(x_i\beta_2))$ so that the termination time of this episode is $T_{ik} = S_{ik} + W_{ik}$. If $T_{ik} < C_i$, then the simulation process continues for subsequent episodes, but otherwise the process terminates and the resolution time of the $k$th episode is censored at $C_i$. We generate $nsim = 1000$ samples of size $m = 1000$ each and report the results in Table 1 for analyses under both RSD-A and RSD-B.

The results show very good agreement between the asymptotic and empirical biases in all settings. Estimators based on RSD-B perform uniformly well in terms of bias and empirical coverage. Note that this model is compatible with the working Markov assumption and so the robust variance is not required; the naive and robust standard errors agree very well with the empirical standard error. As expected from the asymptotic calculations, the empirical bias of the estimated regression coefficients under RSD-A is positive when $\beta_1 < \beta_2$. The average treatment effect under RSD-A is attenuated by the positive treatment effect for the resolution of episodes, and when $\beta_1 = \beta_2$, the empirical bias is low. Interestingly, the use of a robust standard error can induce lower empirical coverage probabilities than a naive standard errors that are used since, here, robust standard errors may be smaller than the naive standard error; see Appendix C for an explanation.

### 4.2 Misspecification of the Risk Set Under Heterogeneity and Dependence

Here, we generalize the underlying two-state process to accommodate unexplained between-individual variation in the risk of symptomatic episodes and their duration through the introduction of random effects. Suppose $U_i = (U_{i1}, U_{i2})'$ is a bivariate random effect for an alternating process so that under the assumption of independent censoring, the conditional intensity functions (1) take the form

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta \bar{N}_{i1}(t) = 1 | \bar{H}_i(t), U_i = u_i)}{\Delta t} = u_{i1}\bar{Y}_{i1}(t)\lambda_{i1}(t) \,,$$

and

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta \bar{N}_{i2}(t) = 1 | \bar{H}_i(t), U_i = u_i)}{\Delta t} = u_{i2}\bar{Y}_{i2}(t)\lambda_{i2}(B_i(t)) \,,$$

where $U_{ij}$ is gamma distributed with $E(U_{ij}) = 1$ and $\text{Var}(U_{ij}) = \phi_j$, for $j = 1, 2$, and a bivariate density function $g(U_i)$ is obtained through a copula model (Joe, 1997).

#### 4.2.1 Marginal Rate and Mean Function Estimation

In the absence of covariates, we assume $\lambda_{i1}(t|H_i(t), U_i = u_i) = u_{i1}\lambda_1(t)$ and retain the semi-Markov form of the $2 \to 1$ intensity given $U_i = u_i$, with $W_{ik}|u_{i2} \sim \text{Gamma}(2, u_{i2}\lambda_2)$. Considering the contribution from a single individual to the estimating equation

$$\sum_{i=1}^{m} \bar{Y}_i^h(t)\{dN_{i1}(t) - dR_1(t)\} = 0 \,,$$

we write

$$E[\bar{Y}_i^h(t)E\{dN_{i1}(t) - dR_1(t)|\bar{Y}_i^h(t) = 1\}] = 0 \,.$$

Table 1: Frequency properties of regression estimator from naive use of the Andersen-Gill model; events simulated according to a two-state processes with $\lambda_{i1}(t|H_i(t)) = \lambda_{01}\exp(x_i\beta_1)$ with $\lambda_{01} = 2, \exp(\beta_1) = 0.75, W_{ik}|X_i \sim GAM(2, \lambda_{02}\exp(X_i\beta_2))$ over $(0, 2]$ with 20% random censoring, $M = 1000$, $nsim = 1000$

| | RSD-A ($\bar{Y}_i^A(t) = Y_i(t)$) | | | | | | | | RSD-B ($\bar{Y}_i^B(t) = Y_i(t)Y_{i1}(t)$) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $E[W_{ik}\|x_i = 0]$ | ABIAS[a] | EBIAS[a] | SE[b] | ESE | ASE[c] | ASE[d] | ECP[c] | ECP[d] | ABIAS[a] | EBIAS[a] | SE[b] | ESE | ASE[c] | ASE[d] | ECP[c] | ECP[d] |
| Treatment lengthens mean episode duration; $\exp(\beta_2) = 0.75$ | | | | | | | | | | | | | | | | |
| 0.10 | 0.003 | 0.001 | 0.034 | 0.034 | 0.040 | 0.033 | 0.980 | 0.946 | -0.0003 | -0.0013 | 0.039 | 0.040 | 0.040 | 0.040 | 0.945 | 0.944 |
| 0.25 | 0.011 | 0.011 | 0.031 | 0.031 | 0.043 | 0.031 | 0.993 | 0.933 | 0.0007 | -0.0003 | 0.043 | 0.043 | 0.043 | 0.043 | 0.961 | 0.958 |
| 0.50 | 0.030 | 0.030 | 0.030 | 0.030 | 0.048 | 0.030 | 0.984 | 0.830 | -0.0003 | -0.0003 | 0.048 | 0.047 | 0.048 | 0.048 | 0.960 | 0.959 |
| Treatment shortens mean episode duration; $\exp(\beta_2) = 1.25$ | | | | | | | | | | | | | | | | |
| 0.10 | 0.066 | 0.065 | 0.034 | 0.034 | 0.039 | 0.034 | 0.622 | 0.511 | -0.0003 | -0.0013 | 0.039 | 0.039 | 0.039 | 0.039 | 0.942 | 0.940 |
| 0.25 | 0.128 | 0.128 | 0.032 | 0.033 | 0.042 | 0.032 | 0.091 | 0.023 | -0.0003 | -0.0003 | 0.042 | 0.043 | 0.042 | 0.042 | 0.951 | 0.951 |
| 0.50 | 0.185 | 0.185 | 0.031 | 0.030 | 0.046 | 0.031 | 0.001 | 0.000 | -0.0003 | -0.0003 | 0.046 | 0.044 | 0.046 | 0.046 | 0.953 | 0.954 |

[a] ABIAS = $\gamma_1^h - \beta_1$ and EBIAS is the empirical bias defined as $\bar{\hat{\gamma}}_1^h - \bar{\beta}_1$ where $\bar{\hat{\gamma}}_1^h$ is the mean estimate over all simulations.
[b] The limiting value of robust standard error.
[c] Naive results, ECP[1]% is the empirical coverage probability for $\beta_1$ of a nominal 95% confidence intervals using the naive standard errors.
[d] Robust results, ECP[2]% is the empirical coverage probability for $\beta_1$ of a nominal 95% confidence intervals using the robust standard errors.

If $h = A$, then $\bar{Y}_i^A(t) = Y_i(t)$ and since censoring is conditionally independent of the event process, $dR_1^A(t)$ solves

$$E[Y_i(t)\{E(U_{i1}|Y_{i1}(t) = 1)P(Y_{i1}(t) = 1)\lambda_1(t)dt - dR_1(t)\}] = 0$$

to give $dR_1^A(t) = E(U_{i1}|Y_{i1} = 1)P(Y_{i1}(t) = 1)\lambda_1(t)dt$. If $h = B$, on the other hand, since $\bar{Y}_i^B(t) = Y_i(t)Y_{i1}(t)$, we write

$$E[Y_i(t)\{E(U_{i1}|Y_{i1}(t) = 1)P(Y_{i1}(t) = 1)\lambda_1(t)\,dt - P(Y_{i1} = 1)dR_1(t)\}] = 0$$

with solution $dR_1^B(t) = E(U_{i1}|Y_{i1}(t) = 1)\lambda_1(t)dt$. Note that

$$P(Y_{i1}(t) = 1) = \int_0^\infty \int_0^\infty P(Y_{i1}(t) = 1|u_i)\,dG(u_i)$$

with $dG(u_i) = g(u_i)du_{i1}du_{i2}$, where we obtain $P(Y_{i1}(t) = 1|u_i)$ from (A.2) or (A.3) in Appendix A by replacing $\lambda_{01}$ and $\lambda_{02}$ with $u_{i1}\lambda_1$ and $u_{i2}\lambda_2$, respectively, and considering only the case with $X_i = 0$. We then compute $E(U_{i1}|Y_{i1}(t) = 1)$ as

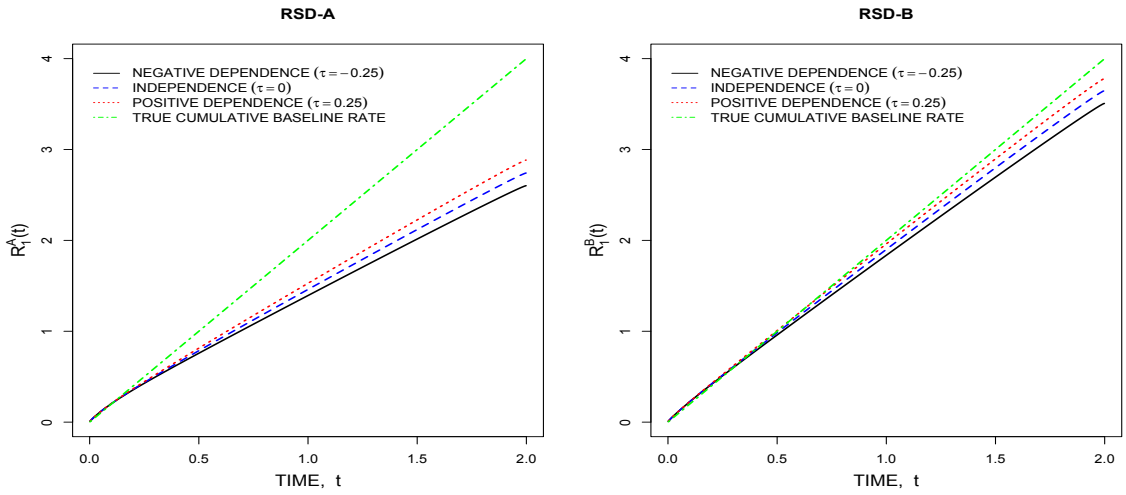
$$\int_0^\infty \int_0^\infty u_{i1}g(u_i|\bar{Y}_{i1}(t) = 1)du_{i1}du_{i2} = \int_0^\infty \int_0^\infty u_{i1}\frac{P(Y_{i1}(t) = 1|u_i)}{P(Y_{i1}(t) = 1)}\,dG(u_i)\,.$$

Figure 5(a) shows the limiting value of cumulative rate function estimate under RSD-A (left panel) and RSD-B (right panel) over $(0, 2]$ with 20% random censoring, $\lambda_1(t) = \lambda_1 = 2$ and $E(W_{ik}) = 0.25$; the cumulative baseline rate is $\Lambda_1(t) = \lambda_1 t$. Figure 5(b) shows the asymptotic bias of the cumulative rate function estimates at $t = 2$ as a function of the mean sojourn time in the exacerbation state. In both settings, we assume that $U_{ij}$ is gamma distributed with mean 1 and variance $\phi_j = 0.4$ for $j = 1, 2$ and we link $U_{i1}$ and $U_{i2}$ with the Gaussian copula (Nelsen, 2006) having Kendall's $\tau = -0.25, 0$, and 0.25. Here, the cumulative mean function is not equal to the cumulative intensity function due to symptom duration. We note that $R_1^A(2)$ and $R_1^B(2)$ are both smaller than the true value of the cumulative rate function $\Lambda_1(2)$ and the bias decreases as Kendall's $\tau$ increases; a strong positive association between $U_{i1}$ and $U_{i2}$ implies that the duration of exacerbations tends to decrease as the risk of exacerbations increases. As the mean sojourn time for exacerbations increases, the bias increases in both RSD-A and RSD-B, and RSD-B yields estimators with smaller bias than RSD-A. The Nelson-Aalen estimate with RSD-B shows a little departure from the true cumulative baseline rate where the bias arises because of the heterogeneity and dependence between random effects for the alternating two-state process.

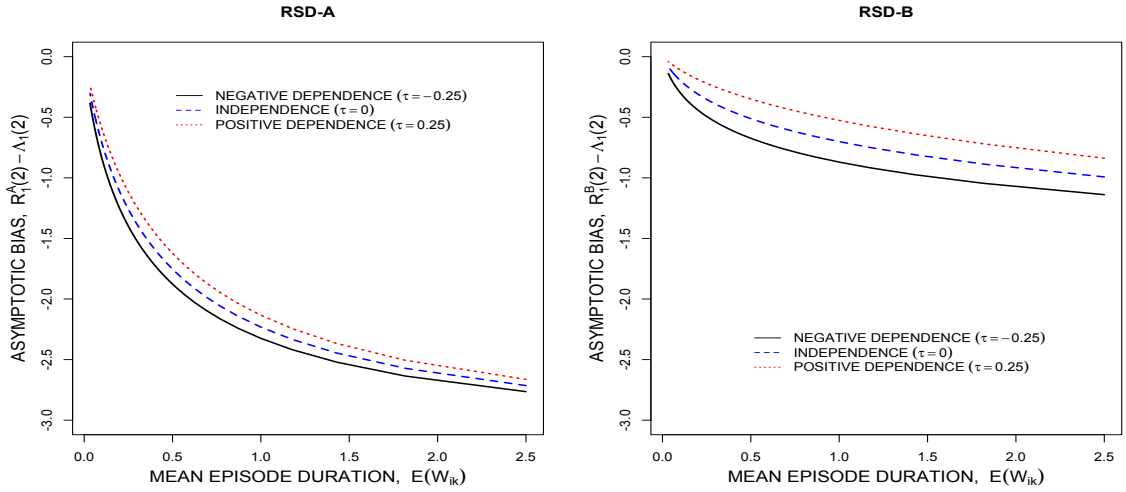
### 4.2.2 Estimation in the Regression Setting

Here, we consider the regression setting when the underlying two-state model features heterogeneity and process dependence through a copula. We assume $\lambda_{i1}(t|u_{i1}, x_i) = u_{i1}\lambda_{01}\exp(x_i\beta_1)$ and $W_{ik}|u_{i2}, x_i \sim \text{Gamma}(2, u_{i2}\lambda_{02}\exp(x_i\beta_2))$. We again consider a randomized clinical trial where $X_i$ is Bernoulli with $P(X_i = 1) = 0.5$. Then, under a working model $dR_{i1}(t) = dR_{01}(t)\exp(x_i\gamma_1)$, the limiting values of the regression coefficient estimators are again obtained by solving (8). The calculations required are analogous to those of Section 4.1.2 and 4.2.1 and are omitted here. We note, however, that we can simplify $\gamma_1^A$ here as

$$\gamma_1^A = \beta_1 + \log\left(\frac{\int_0^\infty P(\bar{Y}_{i1}(u) = 1|X_i = 1)E(U_{i1}|Y_{i1}(u) = 1, X_i = 1)du}{\int_0^\infty P(\bar{Y}_{i1}(u) = 1|X_i = 0)E(U_{i1}|Y_{i1}(u) = 1, X_i = 0)du}\right).$$



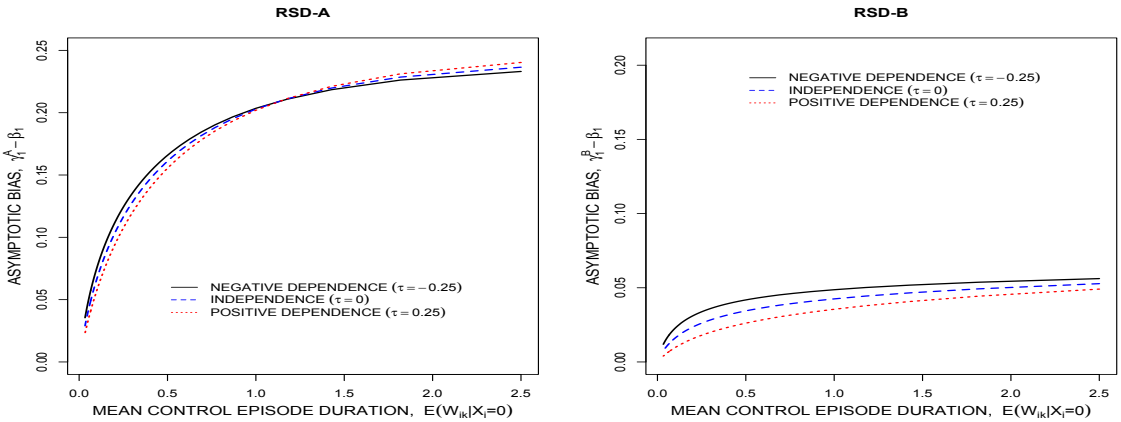

(a) Setting: $\lambda_{01} = 2$, $E[W_{ik}] = 0.25$, $\phi_1 = \phi_2 = 0.4$ with the Gaussian copula and $C = 2$, and 20% random censoring.



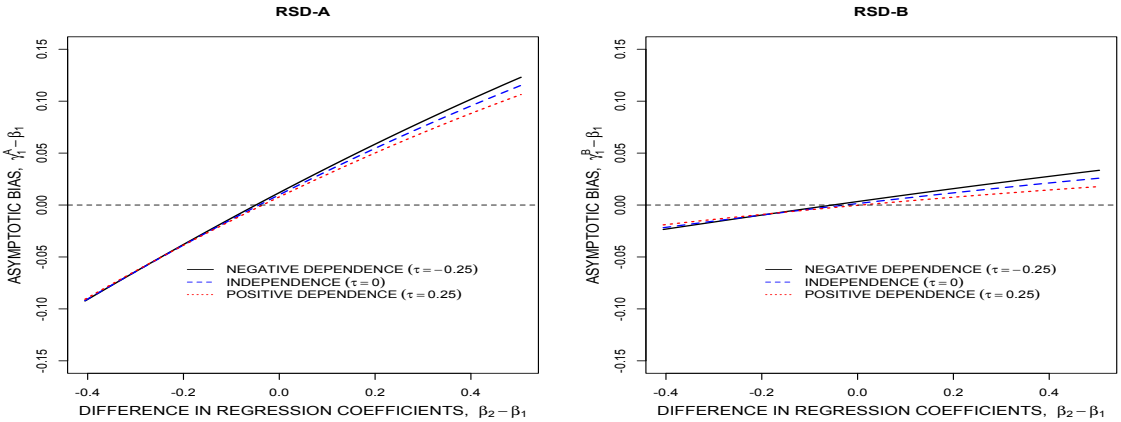

(b) Setting: $\lambda_{01} = 2$, $\phi_1 = \phi_2 = 0.4$ with the Gaussian copula and $C = 2$, and 20% random censoring.

Figure 5: The limiting value of the Nelson-Aalen estimate and the true cumulative baseline hazard under dependence models arising from correlated random effects.

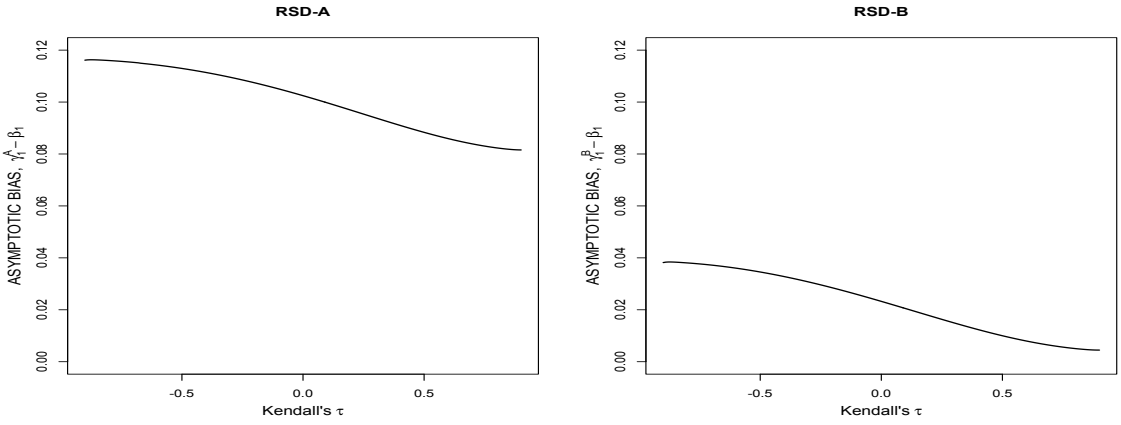
The limiting value $asvar(\sqrt{m}(\hat{\gamma}_1^h - \beta_1)) = \mathcal{A}^{-1}(\gamma_1^h)\,\mathcal{B}(\gamma_1^h)\,\mathcal{A}^{-1}(\gamma_1^h)$ which is estimated by $\hat{A}^{-1}(\hat{\gamma}_1^h)\,\hat{B}(\hat{\gamma}_1^h)\,\hat{A}^{-1}(\hat{\gamma}_1^h)$ as in Section 3.2. Figure 6 displays the asymptotic bias of regression coefficient $\gamma^A$ and $\gamma^B$ when the marginal rate-based model is fitted. We set $\lambda_{01} = 2$, $\beta_1 = \log(0.75)$, consider $\phi_1 = \phi_2 = 0.4$, and let $A = 2$ denote an administrative censoring time with an exponential random censoring time giving 20% early loss to follow up. In Figure 6(a), we see that the larger the mean sojourn time in the exacerbation state, the larger the resulting bias with the estimator from RSD-A incurring the larger bias. In Figure 6(b), the asymptotic bias is expressed as a function of $\beta_2 - \beta_1$, the difference of the treatment effects of the conditional transition intensities for the case where $E(W_{ik}|X_i = 0) = 0.25$. Here, we see that the resulting bias can be positive or negative depending on the magnitude of $\beta_2 - \beta_1$; again, RSD-A yields a estimator which is more sensitive to this type of misspecification. In Figure 6(c), where $E(W_{ik}|X_i = 0) = 0.25$ and $\beta_2 = \log 1.25$, we see that in the setting considered there is a modest impact of the association between the random effects (as reflected by Kendall's $\tau$) on the asymptotic bias; the use of RSD-A again yields the more sensitive estimator.




(a) Asymptotic bias as a function of mean (control) episode duration $E(W_{ik}|X_i = 0)$; $\beta_2 = \log(1.25)$.




(b) Asymptotic bias as a function of difference $\beta_2 - \beta_1$; $E(W_{ik}|X_i = 0) = 0.25$ and $\tau = 0.25$.




(c) Asymptotic bias as a function of Kendall's $\tau$; $E(W_{ik}|X_i = 0) = 0.25$ and $\beta_2 = \log(1.25)$.

Figure 6: Plots of the asymptotic biases $\gamma_1^A - \beta_1$ (left panels) and $\gamma_1^B - \beta_1$ (right panels) under marginal rate-based analyses; we set $\lambda_{01} = 2$, $\beta_1 = \log(0.75)$, $\phi_1 = \phi_2 = 0.4$ and consider administrative censoring at $A = 2$ with 20% random censoring due to early withdrawal.

Here, we describe further simulation studies based on the model incorporating correlated random effects and we investigate finite sample properties of estimators based on the different risk set definitions, extent of the heterogeneity and the dependence between the alternating processes. The data are generated as in Section 4.1.3 but when generating $X_i$, we also generate $U_i = (U_{i1}, U_{i2})'$ using a copula with marginal gamma distributions with $\mathrm{E}(U_{ij}) = 1$ and $\mathrm{VAR}(U_{ij}) = \phi_j = 0.40$, $j = 1, 2$. We use the Gaussian copula (Nelsen, 2006) to model the dependence between the random effects, and set Kendall's $\tau$=-0.25, 0.00, and 0.25. We set $\lambda_{01} = 2$ and $\beta_1 = \log(0.75)$, where $\lambda_{i1}(t|H_i(t), u_i) = u_{i1}\lambda_{01}\exp(x_i\beta_1)$, and $E(W_{ik}|X_i = 0) = 0.1$, 0.25, and 0.5, where $W_{ik}|u_{i2}, x_i \sim$ Gamma$(2, u_{i2}\lambda_{02}\exp(x_i\beta_2))$. For the $k$th episode, we generate the onset time as $S_{ik}|u_{i1}, x_i, S_{ik} > T_{i,k-1}$ which was taken to be a truncated exponential random variable with rate $u_{i1}\lambda_{01}\exp(x_i\beta_1)$, and the duration is generated as $W_{ik}|u_{i2}, x_i \sim$ Gamma$(2, u_{i2}\lambda_{02}\exp(x_i\beta_2))$ resulting in potential resolution time $T_{ik} = S_{ik} + W_{ik}$. The administrative censoring time was set at $A = 2$ and as in Section 4.1.3 an exponential drop-out time $D_i$ was generated to give 20% random censoring; again, $C_i = \min(A, D_i)$ was the right censoring time. We set $m = 1000$ and generated a total of 1000 samples. The results are reported in Table 2 under the RSD-A and RSD-B setting.

Table 2 shows that the means of estimated coefficients are almost equal to their limiting values. In this setting, the impact of using an incorrect definition of the risk set can be appreciable, consistent with the result in Table 1, but there is little effect of a dependence between the random effects under RSD-B. As we observed in Figure 6, the bias decreases as Kendall's $\tau$ increases, and the longer the mean sojourn time and the farther $\beta_2$ is from $\beta_1$, the bigger the bias. There are differences between the naive standard errors and robust standard errors due to the model misspecification, but there is good agreement between the empirical standard error and the average robust standard error compared to the agreement between the average naive standard error and the empirical standard error. Under RSD-B, the robust variance estimates performed fairly well compared to the naive variance estimates although the empirical coverage probabilities are not acceptable when $E(W_{ik}|X_i = 0)$ is appreciable. Under RSD-A, serious bias and low coverage probabilities are obtained. As a result, we conclude that, in regression settings, it is important to take into account the duration of symptoms when specifying the risk set.

## 5 Impact of the Episode Duration Distribution on Power

Here, we explore the impact of misspecification of the risk set on study power. We consider the design of a randomized trial with recurrent events where at the design stage we assume the events are generated by a mixed Poisson process and that the sample size is computed based on the specification $\lambda_{i1}(t|H_i(t), u_{i1}) = u_{i1}\lambda_{01}\exp(x_i\beta_1)$ with $U_{i1}$ gamma distributed with mean 1 and variance $\phi_1$ (Cook and Lawless, 2007). If we wish to test if an intervention has an effect on event occurrence, we typically test $H_0$: $\beta_1 = \beta_{10} = 0$ vs. $H_A$: $\beta_1 \neq \beta_{10} = \beta_{1A}$ with $\beta_{1A}$ the effect of interest. Here, we study the impact on power of the duration of exacerbation episodes as well as the association between the onset of exacerbation episodes and their duration. As previous sections, we consider a semiparametric rate-based model with a robust standard error under RSD-A and RSD-B with analysis here based on a two-sided Wald test.

We set $\lambda_{01} = 2$, $\phi_1 = 0.4$, $\beta_{10} = 0$ and $\beta_{1A} = \log 0.75$ and consider a two-sided test with size 5% and planned a study with 80% power to pick up the effect of interest; these specifications correspond to type I and II error rates denoted by $\alpha_1 = 0.05$ and $\alpha_2 = 0.20$, respectively. With balanced randomization, we have $P(X_i = 1) = P(X_i = 0) = 0.5$. The sample size is computed

Table 2: Frequency properties of regression estimator from naive use of Andersen-Gill model; events simulated according to the conditional intensity-based model of Section 2 where $\lambda_{i1}(t|H_i(t), u_i) = u_{i1}\lambda_{01}\exp(x_i\beta_1)$ with $\lambda_{01} = 2, \exp(\beta_1) = 0.75, W_{ik}|X_i \sim GAM(2, \lambda_{02}\exp(X_i\beta_2))$, $\phi_1 = \phi_2 = 0.4, \tau = (-0.25, 0.00, 0.25)$ over $(0, 2]$ with 20% random censoring

| | RSD-A ($\bar{Y}_i^A(t) = Y_i(t)$) | | | | | | | | RSD-B ($\bar{Y}_i^B(t) = Y_i(t)Y_{i1}(t)$) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\tau$ | ABIAS[a] | EBIAS[a] | SE[b] | ESE | ASE[c] | ASE[d] | ECP[c] | ECP[d] | ABIAS[a] | EBIAS[a] | SE[b] | ESE | ASE[c] | ASE[d] | ECP[c] | ECP[d] |
| | | | | | | $E[W_{ik}\|X_i = 0] = 0.10$; $\exp(\beta_1) = 0.75$ ($\beta_1 = -0.2877$); $\exp(\beta_2) = 0.75$ | | | | | | | | | |
| -0.25 | 0.004 | 0.004 | 0.048 | 0.046 | 0.041 | 0.048 | 0.921 | 0.956 | 0.002 | 0.001 | 0.055 | 0.052 | 0.041 | 0.055 | 0.881 | 0.958 |
| 0.00 | 0.003 | 0.005 | 0.050 | 0.050 | 0.040 | 0.050 | 0.886 | 0.953 | 0.003 | 0.001 | 0.056 | 0.055 | 0.040 | 0.056 | 0.845 | 0.959 |
| 0.25 | 0.002 | 0.005 | 0.052 | 0.051 | 0.039 | 0.052 | 0.870 | 0.955 | 0.001 | 0.004 | 0.056 | 0.055 | 0.039 | 0.056 | 0.829 | 0.960 |
| | | | | | | $E[W_{ik}\|X_i = 0] = 0.25$; $\exp(\beta_1) = 0.75$ ($\beta_1 = -0.2877$); $\exp(\beta_2) = 0.75$ | | | | | | | | | |
| -0.25 | 0.012 | 0.014 | 0.045 | 0.045 | 0.045 | 0.045 | 0.934 | 0.940 | 0.004 | 0.006 | 0.057 | 0.057 | 0.045 | 0.057 | 0.864 | 0.944 |
| 0.00 | 0.010 | 0.014 | 0.048 | 0.046 | 0.043 | 0.047 | 0.920 | 0.945 | 0.002 | 0.005 | 0.058 | 0.056 | 0.043 | 0.058 | 0.862 | 0.953 |
| 0.25 | 0.008 | 0.011 | 0.050 | 0.050 | 0.042 | 0.050 | 0.899 | 0.944 | -0.0003 | 0.002 | 0.058 | 0.059 | 0.042 | 0.058 | 0.844 | 0.949 |
| | | | | | | $E[W_{ik}\|X_i = 0] = 0.50$; $\exp(\beta_1) = 0.75$ ($\beta_1 = -0.2877$); $\exp(\beta_2) = 0.75$ | | | | | | | | | |
| -0.25 | 0.027 | 0.029 | 0.043 | 0.043 | 0.049 | 0.044 | 0.939 | 0.912 | 0.007 | 0.010 | 0.060 | 0.059 | 0.049 | 0.060 | 0.885 | 0.955 |
| 0.00 | 0.024 | 0.026 | 0.046 | 0.046 | 0.048 | 0.046 | 0.928 | 0.913 | 0.007 | 0.003 | 0.061 | 0.062 | 0.048 | 0.061 | 0.869 | 0.944 |
| 0.25 | 0.020 | 0.023 | 0.048 | 0.047 | 0.046 | 0.048 | 0.910 | 0.925 | -0.0003 | 0.002 | 0.061 | 0.060 | 0.046 | 0.061 | 0.873 | 0.955 |
| | | | | | | $E[W_{ik}\|X_i = 0] = 0.10$; $\exp(\beta_1) = 0.75$ ($\beta_1 = -0.2877$); $\exp(\beta_2) = 1.25$ | | | | | | | | | |
| -0.25 | 0.076 | 0.077 | 0.049 | 0.047 | 0.040 | 0.049 | 0.516 | 0.646 | 0.023 | 0.025 | 0.055 | 0.052 | 0.040 | 0.054 | 0.810 | 0.940 |
| 0.00 | 0.067 | 0.069 | 0.051 | 0.050 | 0.040 | 0.051 | 0.555 | 0.727 | 0.017 | 0.019 | 0.056 | 0.055 | 0.040 | 0.056 | 0.818 | 0.941 |
| 0.25 | 0.058 | 0.065 | 0.053 | 0.051 | 0.038 | 0.052 | 0.617 | 0.807 | 0.010 | 0.013 | 0.056 | 0.055 | 0.038 | 0.056 | 0.824 | 0.948 |
| | | | | | | $E[W_{ik}\|X_i = 0] = 0.25$; $\exp(\beta_1) = 0.75$ ($\beta_1 = -0.2877$); $\exp(\beta_2) = 1.25$ | | | | | | | | | |
| -0.25 | 0.125 | 0.126 | 0.046 | 0.047 | 0.043 | 0.046 | 0.188 | 0.225 | 0.034 | 0.036 | 0.056 | 0.057 | 0.043 | 0.056 | 0.784 | 0.903 |
| 0.00 | 0.117 | 0.121 | 0.048 | 0.047 | 0.042 | 0.048 | 0.208 | 0.296 | 0.027 | 0.031 | 0.057 | 0.056 | 0.042 | 0.057 | 0.806 | 0.927 |
| 0.25 | 0.108 | 0.111 | 0.050 | 0.050 | 0.041 | 0.050 | 0.277 | 0.416 | 0.018 | 0.022 | 0.058 | 0.056 | 0.041 | 0.058 | 0.813 | 0.937 |
| | | | | | | $E[W_{ik}\|X_i = 0] = 0.50$; $\exp(\beta_1) = 0.75$ ($\beta_1 = -0.2877$); $\exp(\beta_2) = 1.25$ | | | | | | | | | |
| -0.25 | 0.166 | 0.168 | 0.044 | 0.045 | 0.047 | 0.044 | 0.050 | 0.037 | 0.042 | 0.044 | 0.059 | 0.060 | 0.047 | 0.059 | 0.762 | 0.880 |
| 0.00 | 0.162 | 0.162 | 0.046 | 0.046 | 0.046 | 0.046 | 0.059 | 0.064 | 0.035 | 0.036 | 0.060 | 0.059 | 0.046 | 0.059 | 0.804 | 0.912 |
| 0.25 | 0.156 | 0.157 | 0.048 | 0.049 | 0.045 | 0.048 | 0.086 | 0.113 | 0.026 | 0.027 | 0.060 | 0.060 | 0.045 | 0.060 | 0.814 | 0.924 |

[a] ABIAS = $\gamma_1^h - \beta_1$ and EBIAS is the empirical bias defined as $\bar{\gamma}_1^h - \bar{\beta}_1$ where $\bar{\gamma}_1^h$ is the mean estimate over all simulations.
[b] The limiting value of robust standard error.
[c] Naive results, ECP[1]% is the empirical coverage probability for $\beta_1$ of a nominal 95% confidence intervals using the naive standard errors.
[d] Robust results, ECP[2]% is the empirical coverage probability for $\beta_1$ of a nominal 95% confidence intervals using the robust standard errors.

under the assumption of a mixed Poisson model and so we have

$$m \geq \left\{ \frac{z_{\alpha_1/2}\sqrt{asvar_0(\sqrt{m}(\hat{\beta}_1 - \beta_{10}))} + z_{\alpha_2}\sqrt{asvar_A(\sqrt{m}(\hat{\beta}_1 - \beta_{1A})}}{\beta_{1A}} \right\}^2, \tag{17}$$

where $z_p$ is the upper $p$th percentile of the standard normal distribution and $asvar_0(\cdot)$ and $asvar_A(\cdot)$ denote the asymptotic variance under the null and alternative hypotheses respectively with

$$asvar(\sqrt{m}(\hat{\beta}_1 - \beta_1)) = \sum_{x=0}^{1} \left\{ P(X_i = x) E\left[ \frac{\lambda_{01} \exp(x\beta_1) C_i}{1 + \phi_1 \lambda_{01} \exp(x\beta_1) C_i} \right] \right\}^{-1}.$$

We next conduct simulation studies to investigate the impact of risk set misspecification and heterogeneity on power when the sample size is calculated based on a mixed Poisson model (17) with $E\{\bar{N}_{i1}(2)\} = 4$, $\phi_1 = 0.4$ and there is 20% random censoring. We simulate data sets of the corresponding sample size with each individuals data arising from a conditionally Markov/semi-Markov model with correlated random effects arising from the earlier copula model. We set the parameters of this model to ensure $E\{\bar{N}_{i1}(2)\} = 4$ and $\phi_1 = \phi_2 = 0.4$ to be roughly compatible with the design assumptions made under the mixed Poisson formulation. We consider different mean sojourn times for exacerbation episodes in the control arm with $E(W_{ik}|X_i = 0) = 0.1, 0.25$, or 0.5 and Kendall's $\tau = -0.25, 0$, or 0.25. For each data set, we fit the AG model with RSD-A and RSD-B and tested the null hypothesis of no treatment effect via a two-sided Wald test with robust standard errors. A total number of 2000 samples were generated according to the required sample size and the empirical rejection rates (REJ%), defined as the percentage of replicates leading to rejection of the null hypothesis, were computed and summarized in Table 3.

Table 3: Empirical rejection rates for rate-based tests of treatment effects on the onset of exacerbations where the sample size was calculated based on the mixed Poisson model with $A = 2$ and 20% random censoring, $E\{\bar{N}_{i1}(2)|x_i = 0\} = 4$, $\phi_1 = 0.4$, $\beta_{10} = 0$, and $\beta_{1A} = \log(0.75)$; in the two-state data generating model, we consider $E(W_{ik}|X_i = 0) = 0.10$, 0.25, and 0.50, $\phi_2 = 0.4$ and Kendall's $\tau = -0.25, 0$, or 0.25; $nsim = 2000$

| | | $\tau$=-0.25 | | $\tau$ = 0 | | $\tau$ = 0.25 | |
|---|---|---|---|---|---|---|---|
| $\beta_2$ | $E(W_{ik}\|X_i = 0)$ | RSD-A | RSD-B | RSD-A | RSD-B | RSD-A | RSD-B |
| $\log(0.75)$ | 0.10 | 92.1 | 81.5 | 88.2 | 79.4 | 85.8 | 79.1 |
| | 0.25 | 96.8 | 80.3 | 93.1 | 79.0 | 90.5 | 78.0 |
| | 0.50 | 97.2 | 72.4 | 95.4 | 72.3 | 92.3 | 73.2 |
| 0 | 0.10 | 77.4 | 77.8 | 73.5 | 77.1 | 72.6 | 77.0 |
| | 0.25 | 66.5 | 73.5 | 60.7 | 74.1 | 59.5 | 75.5 |
| | 0.50 | 34.5 | 64.4 | 34.2 | 67.7 | 30.9 | 67.8 |
| $\log(1.25)$ | 0.10 | 60.0 | 72.8 | 61.2 | 75.0 | 61.3 | 76.6 |
| | 0.25 | 28.3 | 68.7 | 29.7 | 69.5 | 30.6 | 73.1 |
| | 0.50 | 5.9 | 58.6 | 5.1 | 61.2 | 5.3 | 62.9 |

When $\beta_2 = 0$, it can be seen that the empirical power is lower than the nominal level when the mean duration of the episodes is low but it becomes appreciable as the mean durations increase. The effect of the intervention on the duration of the episodes is apparent in the rate-based analysis of the onset times where higher power is realized when $\beta_2 < 0$ and a further reduction is seen when $\beta_2 > 0$. These two effects are far greater under RSD-A than they are under RSD-B. There is a modest sensitivity of power to the value of Kendall's $\tau$ and again this appears to be greater for RSD-A compared to RSD-B. Figure 7 shows power curves with the same setting as the empirical study. The effect of Kendall's $\tau$ on power relies on the mean sojourn time in the exacerbation-free state and the value of $\beta_2$. When $\beta_1 \neq \beta_2$, the increase in the mean sojourn time in the exacerbation-state reduces power, however, when $\beta_1 = \beta_2$, power is greater than 80% with RSD-A. The loss in power with RSD-B is smaller than the one with RSD-A when $\beta_1 \neq \beta_2$.

## 6 Application to a Herpes Simplex Trial

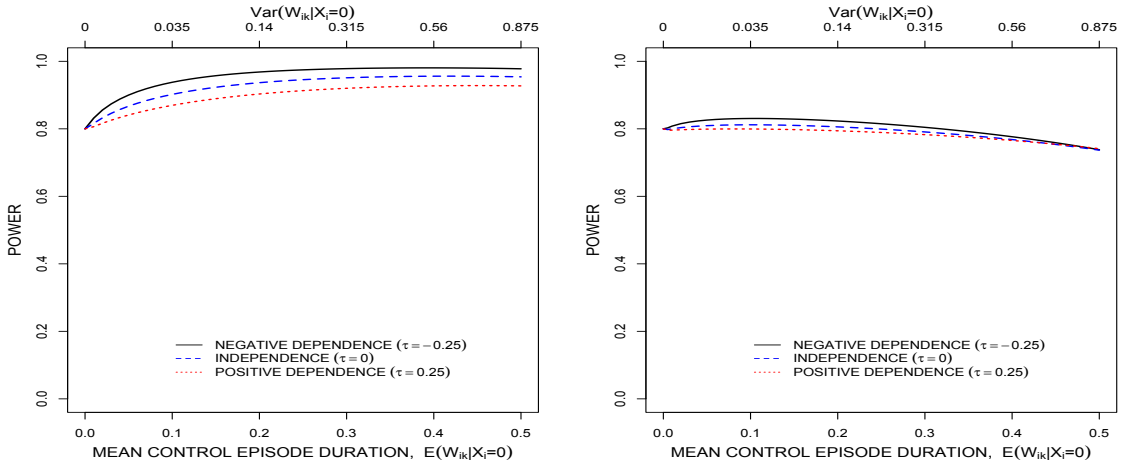
Herpes simplex infection leads to recurrent symptomatic episodes which typically last two to four weeks in duration. A multicenter open-label randomized two-period crossover trial was conducted to compare the efficacy of the use of valacyclovir defined as suppressive therapy versus episodic therapy (Romanowski et al., 2003). Suppressive therapy was valacyclovir at a dosage of 500 mg once daily and episodic therapy was valacyclovir at a dosage of 500 mg twice daily for 5 days commencing upon the outbreaks of symptoms. If herpes outbreaks occurred in the suppressive arm, patients received episodic therapy (the 500 mg twice daily) for 5 days and returned to suppressive therapy after 5 days. Out of the total of 225 patients enrolled, 202 completed the two 24-week periods of the study. After the first period of the study, patients switched another therapy so that each patient received both treatments for the 48-week study period. The mean of the total number of outbreaks for the first period is 4.02 with the standard error of 3.90. The mean symptom duration was 24.1 days with a minimum of 1 day and a maximum of 175 days. In this paper, we only consider the first 24-week study period and so each patient yielded data on either suppressive or episodic therapy. We also include gender (female vs male) and virus type (HSV1 or HSV2) as covariates in addition to treatment (episodic therapy vs suppressive therapy). In Table 4, we report on analyses of herpes simplex study using the semiparametric rate-based analysis with RSD-A and RSD-B, respectively.

Table 4: Analysis of occurrence of herpes simplex using RSD-A and RSD-B based on the Andersen-Gill model

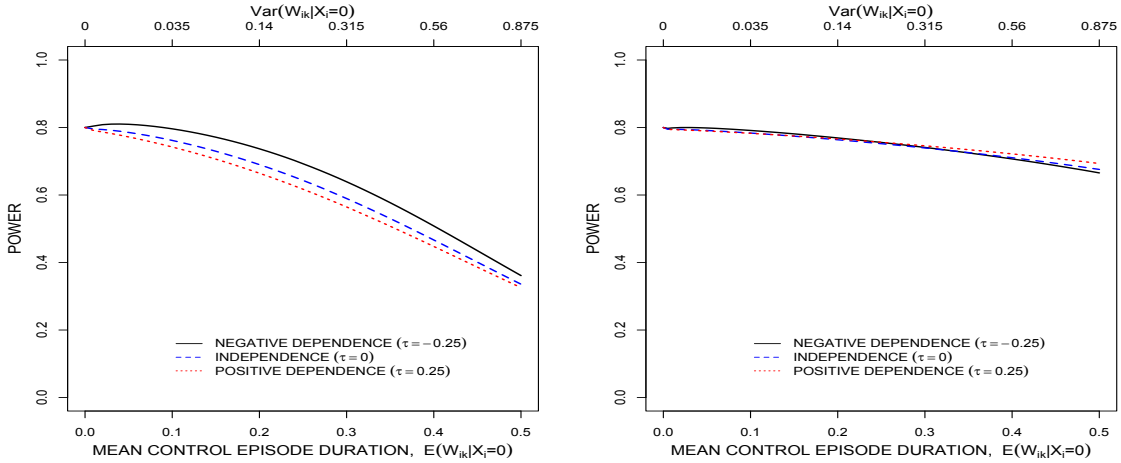
| | RSD-A ($\bar{Y}_i^A(t) = Y_i(t)$) | | | | RSD-B ($\bar{Y}_i^B(t) = Y_i(t)Y_{i1}(t)$) | | | |
|---|---|---|---|---|---|---|---|---|
| Covariate | EST | SE[a] | SE[b] | p[c] | EST | SE[a] | SE[b] | p[c] |
| Treatment | -1.875 | 0.200 | 0.240 | $< 0.001$ | -1.871 | 0.145 | 0.186 | $< 0.001$ |
| Sex | -0.189 | 0.135 | 0.173 | 0.276 | -0.303 | 0.115 | 0.166 | 0.067 |
| Virus Type | 0.159 | 0.121 | 0.146 | 0.277 | 0.071 | 0.107 | 0.148 | 0.632 |

[a] Naive standard error
[b] Robust standard error
[c] p-values based on robust standard error

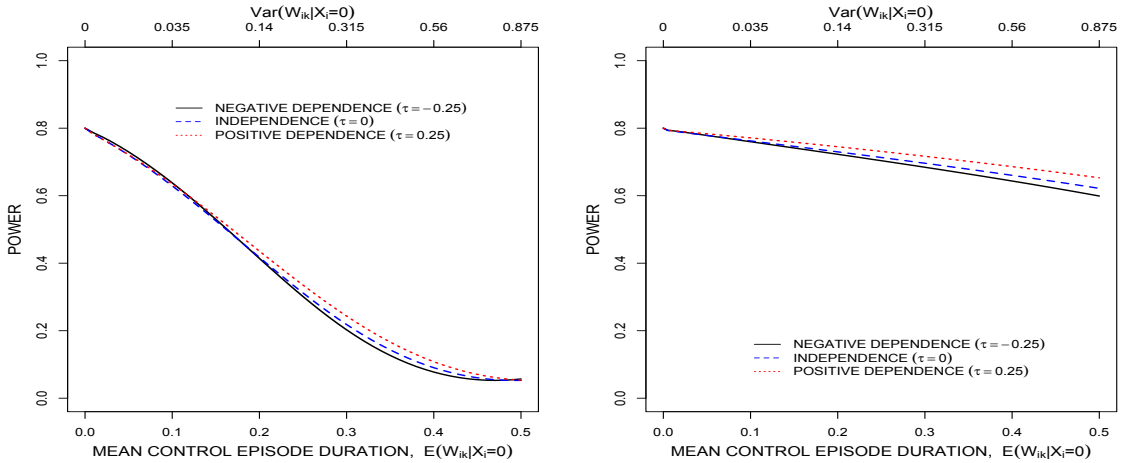
Treatment was found to have a significant effect on the occurrence of exacerbations under both RSD-A ($RR = 0.15; 95\%$ CI: 0.09, 0.25; $p < 0.001$) and RSD-B ($RR = 0.15; 95\%$ CI: 0.11, 0.22; $p < 0.001$) with the two estimates in very close agreement. The estimate of the effect of gender




(a) $\beta_2 = \log(0.75)$




(b) $\beta_2 = 0$




(c) $\beta_2 = \log(1.25)$

Figure 7: Power curves based on RSD-A (left panel) and RSD-B (right panel) with Kendall's $\tau$ -0.25, 0, and 0.25 where the sample size is calculated based on the mixed Poisson model with $E\{\bar{N}_{i1}(2)\} = 4$, $\beta_{10} = 0, \beta_{1A} = \log(0.75)$, $\phi_1 = 0.4$, $\phi_2 = 0.4$.

with RSD-A ($RR$ = 0.83; 95% CI: 0.57, 1.19; $p = 0.276$) differed from the one with RSD-B ($RR$ = 0.74; 95% CI: 0.53, 1.02; $p = 0.067$). In addition, there are differences in the estimates of the effect of virus type between RSD-A ($RR = 1.17; 95\%$ CI: 0.84, 1.65; $p = 0.277$) and RSD-B ($RR = 1.07; 95\%$ CI: 0.80, 1.43; $p = 0.632$). We also note that the naive standard errors and robust standard errors are not identical. Figure 8 contains a plot of the estimated cumulative baseline rate function from the regression model based on RSD-A and RSD-B. The slope of the cumulative baseline rate function with RSD-B is greater than the one with RSD-A, suggestive of a higher rate for the occurrence of outbreaks with RSD-B than RSD-A (Cook and Lawless, 2007). Note that, with RSD-A, the cumulative baseline rate function can be naively interpreted as an estimate of the cumulative baseline mean function.

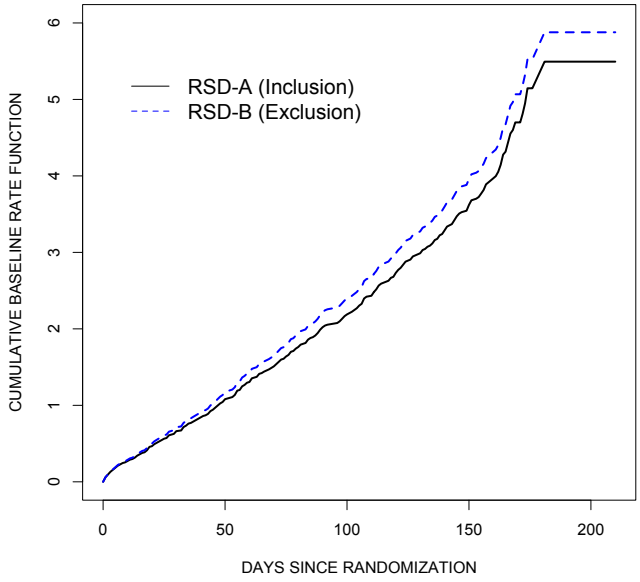



Figure 8: Cumulative baseline rate function with RSD-A (Inclusion) and RSD-B (Exclusion)

## 7 Discussion

In this paper, we have pointed out that estimators of mean function and covariate effects from the naive use of rate-based models (Andersen et al., 1993) are sensitive to the handling of risk-free periods as well as strength of the association between the onset and duration of episodic events. Misspecification of at risk indicators can lead to inconsistent estimators of regression coefficients and the use of robust standard errors does not guarantee protection against misspecification of the duration dependent processes. The biases we refer to for the mean function are specified in relation to the cumulative intensity for the onset of episodes, or the actual mean function reflected the expected number of events over time. In the regression setting, we refer to the bias of estimators of the regression coefficient for the transition intensity for the onset of episodes.

Full specification of the intensities for an alternating two-state process is challenging in practice and it is impossible to achieve robustness in this framework since correct model specification is required to ensure that the partial likelihood estimating equations are unbiased. Causal inference can be based on the expected number of events at a landmark time or based on proportional rate function models but there is a tension between the need for full specification of models to advance scientific understanding and the need for simple models supporting causal conclusions. Lee and Cook (2018)

develop a model for a mixed two-state process for characterizing recurrent episodic conditions which features a Markov time-scale for the onset of exacerbations and a semi-Markov time scale for the duration of the exacerbations. Correlated random effects enable one to assess the need to accommodate heterogeneity and allow for a dependence between the sojourn times in the exacerbation state and the risk for the onset of events.

When mortality rates are appreciable, as is the case among individuals with advanced chronic obstructive pulmonary disease, it is considerably more challenging to model the onset and duration of exacerbations and summarize the effects of interventions. In the multistate framework, an absorbing state representing death can be added, and random effects can be considered in the intensities for death. However, expressing treatment effects robustly on the onset of exacerbations is very challenging. Much work has been carried out in this area for recurrent transient events (Cook and Lawless, 1997; Ghosh and Lin, 2000, 2002) but utility-based analyses may be preferable when events have a duration associated with them (Cook et al., 2003).

An alternative approach in these more complex settings is to focus on estimation of state occupancy probabilities using nonparametric methods. Cook and Lawless (2018) discuss this for one-sample problems and consider marginal regression models for state occupancy probabilities based on direct binomial regression (Scheike et al., 2008). Utility-based analyses are also of possible value (Cook et al., 2003; Cook and Lawless, 2018). These and other marginal quantities, such as features of state entry time or sojourn time distributions, may offer a more convenient basis for causal inference since they are not defined inherently in terms of conditional probabilities. As always, the choice of the estimand must be made based on interpretation and it must be meaningful for the problem at hand. Inverse probability weighting can often be useful to correct for some selection biases and confounding, but in complex settings even use of such methods can be challenging.

## ACKNOWLEDGEMENTS

The authors thank GlaxoSmithKline for permission to use the data fromthe trial used in the application and Jerry Lawless, Kyle Raymond, and Leilei Zeng for comments on an earlier draft of this work. This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada (RGPIN 155849) and the Canadian Institutes for Health Research (FRN 13887). Richard Cook is a Canada Research Chair in Statistical Methods for Health Research.

## CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

## DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article due to confidentiality.

## REFERENCES

Andersen, P., Borgan, O., Gill, R., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

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

Boher, J. and Cook, R. (2006). Implications of model misspecification in robust tests for recurrent events. *Lifetime Data Analysis*, 12(1):69–95.

Cole, S., Platt, R., Schisterman, E., Chu, H., Westreich, D., Richardson, D., and Poole, C. (2009). Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology*, 39(2):417–420.

Cook, R. and Lawless, J. (1997). Marginal analysis of recurrent events and a terminating event. *Statistics in Medicine*, 16(8):911–924.

Cook, R. and Lawless, J. (2007). *The Statistical Analysis of Recurrent Events*. Springer, New York, NY.

Cook, R. and Lawless, J. (2018). *Multistate Models for the Analysis of Life History Data*. Chapman and Hall/CRC, New York.

Cook, R., Lawless, J., Lakhal-Chaieb, L., and Lee, K. (2009). Robust estimation of mean functions and treatment effects for recurrent events under event-dependent censoring and termination: application to skeletal complications in cancer metastatic to bone. *Journal of the American Statistical Association*, 104(485):60–75.

Cook, R., Lawless, J., and Lee, K. (2003). Cumulative processes related to event histories. *SORT-Statistics and Operations Research Transactions*, 27(1):13–30.

Cox, D. (1967). *Renewal Theory*. Methuen, London.

Cox, D. and Miller, H. (1965). *The Theory of Stochastic Processes*. Chapman and Hall, London.

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

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

Grossman, R., Mukherjee, J., Vaughan, D., Eastwood, C., Cook, R., LaForge, J., and Lampron, N. (1998). A 1-year community-based health economic study of ciprofloxacin vs usual antibiotic treatment in acute exacerbations of chronic bronchitis: the canadian ciprofloxacin health economic study group. *Chest*, 113(1):131–141.

Hernán, M. and Robins, J. (2016). *Causal Inference Book*. Boca Raton: Chapman & Hall/CRC.

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

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. Chapman and Hall, London.

Kalbfleisch, J. and Prentice, R. (2011). *The Statistical Analysis of Failure Time Data*, volume 360. John Wiley & Sons, Hoboken, NJ, USA.

Kessing, L., Olsen, E., Andersen, P., and in cooperation with the Department of Psychiatric Demography, University of Aarhus, Psychiatric Hospital, Risskov, Denmark (1999). Recurrence in affective disorder: analyses with frailty models. *American Journal of Epidemiology*, 149(5):404–411.

Klein, J. (1992). Semiparametric estimation of random effects using the cox model based on the em algorithm. *Biometrics*, pages 795–806.

Lawless, J. (1987). Negative binomial and mixed poisson regression. *Canadian Journal of Statistics*, 15(3):209–225.

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

Lee, J. and Cook, R. (2018). Heterogeneity and dependence modeling for alternating two-state processes via copulas. *Manuscript*.

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

Nelsen, R. (2006). An introduction to copulas, 2nd. *New York: Springer Science Business Media*.

Nelson, W. (1995). Confidence limits for recurrence data-applied to cost or number of product repairs. *Technometrics*, 37(2):147–157.

Ng, E. and Cook, R. (1997). Modeling two-state disease processes with random effects. *Lifetime Data Analysis*, 3(4):315–335.

Romanowski, B., Marina, R., Roberts, J., and Valtrex HS230017 Study Group (2003). Patients' preference of valacyclovir once-daily suppressive therapy versus twice-daily episodic therapy for recurrent genital herpes: a randomized study. *Sexually Transmitted Diseases*, 30(3):226–231.

Scheike, T., Zhang, M., and Gerds, T. (2008). Predicting cumulative incidence probability by direct binomial regression. *Biometrika*, 95(1):205–220.

Wienke, A. (2010). *Frailty Models in Survival Analysis*. CRC Press, Boca Raton, FL.

Xue, X. and Brookmeyer, R. (1996). Bivariate frailty model for the analysis of multivariate survival time. *Lifetime Data Analysis*, 2(3):277–289.

## APPENDIX A: COMPUTATION OF STATE OCCUPANCY PROBABILITIES

$P(Y_{i1}(t) = 1|x_i)$ is difficult to calculate under the assumption of semi-Markov model, especially when the distribution of the duration of exacerbations is not exponential. Here, we decompose state 2 into two states to exploit the property of Gamma distribution which can be expressed as a sum of exponential distribution. We define a new state process $\{\bar{Z}(t), 0 < t\}$ on the extended state space $\{1, 2A, 2B\}$ (Cook et al., 2009) and let $Z(t) = 1$ if $\bar{Z}(t) = 1$ and $Z(t) = 2$ if $\bar{Z}(t) = 2A$ or $\bar{Z}(t) = 2B$, as shown in Figure A.1 . Then, $P(Y_{i1}(t) = 1|x_i)$ can be expressed as

$$P(Z(t) = 1|Z(0) = 1, x_i) = 1 - \sum_{r=2A,2B} P(\bar{Z}(t) = r|\bar{Z}(0) = 1, x_i) = P(\bar{Z}(t) = 1|\bar{Z}(0) = 1, x_i).$$


Figure A.1 : State diagram for recurrent exacerbations with extended Markov models

The term $P(\bar{Z}(t) = 1|\bar{Z}(0) = 1, x_i)$ is calculated by the transition probability matrix $\mathcal{P}(0, t|x_i) = \mathcal{P}(t|x_i) = [p_{ij}(t|x_i)]$, for $i, j = 1, 2A, 2B$. Here, we consider the time-homogeneous case. We assume that the duration of the $k$th exacerbation is $W_{ik} = W_{ik2A} + W_{ik2B}$ where $W_{ikl} \sim$ Exponential$(\lambda_{i2})$ for $l = 2A, 2B$ and $W_{ik2A} \perp W_{ik2B}$, so $W_{ik} \sim$ Gamma$(2, \lambda_{i2})$. Under the multiplicative model, we let $\lambda_{i1} = \lambda_{01} \exp(x_i\beta_1)$ and $\lambda_{i2} = \lambda_{02} \exp(x_i\beta_2)$. It is noted that there is a common covariate for the development and resolution of exacerbations. The time-homogeneous transition intensity matrix of $\{\bar{Z}(t), 0 < t\}$ on state space $\{1, 2A, 2B\}$ is

$$Q = \begin{bmatrix} -\lambda_{i1} & \lambda_{i1} & 0 \\ 0 & -\lambda_{i2} & \lambda_{i2} \\ \lambda_{i2} & 0 & -\lambda_{i2} \end{bmatrix}.$$

We let $P_{12A}(t) = P(\bar{Z}(t) = 2A|\bar{Z}(0) = 1, x_i), P_{12B}(t) = P(\bar{Z}(t) = 2B|\bar{Z}(0) = 1, x_i)$, and $P_{11}(t) = P(\bar{Z}(t) = 1|\bar{Z}(0) = 1, x_i)$. Using the Kolmogorov forward equations (Cox and Miller, 1965), we note

$$\begin{aligned} P'_{12A}(t) &= -\lambda_{i2}P_{12A}(t) + \lambda_{i1}P_{11}(t) \\ P'_{12B}(t) &= \lambda_{i2}P_{12A}(t) - \lambda_{i2}P_{12B}(t) \\ P'_{11}(t) &= \lambda_{i2}P_{12B}(t) - \lambda_{i1}P_{11}(t) \\ P_{12A}(t) + P_{12B}(t) + P_{11}(t) &= 1, \quad P_{11}(0) = 1 \end{aligned} \tag{A.1}$$

By solving the systems of equation of (A.1), we obtain $P_{11}(t)$ if the term $\lambda_{i1} - \lambda_{i2}/4 < 0$ as

$$P_{11}(t) = \frac{(\lambda_{i2})^2}{a^2 + b^2} + \exp(-at)\cos(bt)\left(\frac{2\lambda_{i1}\lambda_{i2}}{a^2 + b^2}\right) + \exp(-at)\sin(bt)\left(\frac{2\lambda_{i1}\lambda_{i2}(\lambda_{i2} - \lambda_{i1})}{(a^2 + b^2)2b}\right), \tag{A.2}$$

where $a = \lambda_{i1}/2 + \lambda_{i2}$ and $b = \sqrt{\lambda_{i1}\lambda_{i2} - (\lambda_{i1})^2/4}$. If $\lambda_{i1} - \lambda_{i2}/4 > 0$ it can be written as follows using Euler's formula,

$$\begin{aligned} P_{11}(t) = \frac{(\lambda_{i2})^2}{a^2 - (b')^2} &+ \exp(-at)\cosh(b't)\left(\frac{2\lambda_{i1}\lambda_{i2}}{a^2 - (b')^2}\right) \\ &+ \exp(-at)\sinh(b't)\left(\frac{2\lambda_{i1}\lambda_{i2}(\lambda_{i2} - \lambda_{i1})}{(a^2 - (b')^2)2b'}\right), \end{aligned} \tag{A.3}$$

where $b' = bi$. Likewise, if $\lambda_{i1} - \lambda_{i2}/4 < 0$, $P_{21}(t)$ is given as

$$P_{21}(t) = \frac{\lambda_{i2}}{2\lambda_{i1}+\lambda_{i2}} - \frac{\lambda_{i2}}{2\lambda_{i1}+\lambda_{i2}}\exp(-at)\cos(bt) - \frac{2\lambda_{i2}^2+\lambda_{i1}\lambda_{i2}}{(2\lambda_{i1}+\lambda_{i2})2b}\exp(-at)\sin(bt)$$

else

$$P_{21}(t) = \frac{\lambda_{i2}}{2\lambda_{i1}+\lambda_{i2}} - \frac{\lambda_{i2}}{2\lambda_{i1}+\lambda_{i2}}\exp(-at)\cosh(b't) - \frac{2\lambda_{i2}^2+\lambda_{i1}\lambda_{i2}}{(2\lambda_{i1}+\lambda_{i2})2b'}\exp(-at)\sinh(b't)$$

## APPENDIX B: CALCULATION OF THE ASYMPTOTIC BIAS OF $\hat{\gamma}_1^A$

Here, we derive

$$\gamma_1^A = \beta_1 + \log\left(\frac{\int_0^\infty P(\bar{Y}_{i1}(u)=1|X_i=1)du}{\int_0^\infty P(\bar{Y}_{i1}(u)=1|X_i=0)du}\right)$$

in (16). By plugging $s^{(0,A)}(u)$, $s^{(1,A)}(u)$, $s^{(0,A)}(\gamma_1,u)$, and $s^{(1,A)}(\gamma_1,u)$ into (8), we have

$$\int_0^\infty \Big\{ P(\bar{Y}_{i1}(u)=1|X_i=1)\lambda_{01}\exp(\beta_1) \\ - \frac{\exp(\gamma_1)}{1+\exp(\gamma_1)}\left(P(\bar{Y}_{i1}(u)=1|X_i=1)\lambda_{01}\exp(\beta_1) + P(\bar{Y}_{i1}(u)=1|X_i=0)\lambda_{01}\right)\Big\}du = 0.$$

Then,

$$\frac{\exp(\gamma_1)}{1+\exp(\gamma_1)} = \frac{\int_0^\infty P(\bar{Y}_{i1}(u)=1|X_i=1)\lambda_{01}\exp(\beta_1)du}{\int_0^\infty \left(P(\bar{Y}_{i1}(u)=1|X_i=1)\lambda_{01}\exp(\beta_1) + P(\bar{Y}_{i1}(u)=1|X_i=0)\lambda_{01}\right)du}. \quad \text{(B.1)}$$

We arrange (B.1) in terms of $\gamma_1$ so that

$$\exp(\gamma_1) = \exp(\beta_1)\frac{\int_0^\infty P(\bar{Y}_{i1}(u)=1|X_i=1)du}{\int_0^\infty P(\bar{Y}_{i1}(u)=1|X_i=0)du},$$

which has the final form as (16) by taking $\log()$ for both sides.

## APPENDIX C: DERIVATION OF THE ROBUST COVARIANCE MATRIX

Let $dM_{i1}^h(t) = \bar{Y}_i^h(t)\{dN_{i1}(t) - dR_{01}(t)\exp(x_i\gamma_1)dt\}$. Then

$$\begin{aligned}
\mathcal{A}(\gamma_1) &= E\left[\int_0^\infty \bar{Y}_i^h(t)\left\{\frac{s^{(2,h)}(\gamma_1,t)}{s^{(0,h)}(\gamma_1,t)} - \frac{s^{(1,h)}(\gamma_1,t)^{\otimes 2}}{s^{(0,h)}(\gamma_1,t)^2}\right\}dN_{i1}(t)\right] \\
&= \sum_{x_i}\left[\int_0^C P(X_i=x_i)P(\bar{Y}_i^h(t)=1|x_i)\left\{\frac{s^{(2,h)}(\gamma_1,t)}{s^{(0,h)}(\gamma_1,t)} - \frac{s^{(1,h)}(\gamma_1,t)^{\otimes 2}}{s^{(0,h)}(\gamma_1,t)^2}\right\}E(dN_{i1}(t)|x_i,\bar{Y}_i^h(t)=1)\right], \\
\mathcal{B}(\gamma_1) &= E\left[\left(\int_0^\infty\left\{x_i - \frac{s^{(1,h)}(\gamma_1,s)}{s^{(0,h)}(\gamma_1,s)}\right\}dM_{i1}^h(s)\right)\left(\int_0^\infty\left\{x_i - \frac{s^{(1,h)}(\gamma_1,t)}{s^{(0,h)}(\gamma_1,t)}\right\}dM_{i1}^h(t)\right)\right] \\
&= E\left[\int_0^\infty\int_0^\infty\left\{x_i - \frac{s^{(1,h)}(\gamma_1,s)}{s^{(0,h)}(\gamma_1,s)}\right\}\left\{x_i - \frac{s^{(1,h)}(\gamma_1,t)}{s^{(0,h)}(\gamma_1,t)}\right\}dM_{i1}^h(s)dM_{i1}^h(t)\right] \\
&= B_1^h + B_2^h - 2B_3^h + B_4^h, \quad \text{(C.1)}
\end{aligned}$$

where

$$B_1^h = \sum_{x_i} \int_0^C P(X_i = x_i) P(\bar{Y}_i^h(t) = 1|x_i) \left\{ x_i - \frac{s^{(1,h)}(\gamma_1, t)}{s^{(0,h)}(\gamma_1, t)} \right\}^2 E(dN_{i1}^2(t)|x_i, \bar{Y}_i^h(t) = 1),$$

$$B_2^h = \sum_{x_i} \int_0^C \int_0^C P(X_i = x_i) P(\bar{Y}_i^h(s) = 1, \bar{Y}_i^h(t) = 1|x_i) \left\{ x_i - \frac{s^{(1,h)}(\gamma_1, s)}{s^{(0,h)}(\gamma_1, s)} \right\} \left\{ x_i - \frac{s^{(1,h)}(\gamma_1, t)}{s^{(0,h)}(\gamma_1, t)} \right\}$$
$$\times E(dN_{i1}(s) dN_{i1}(t)|x_i, \bar{Y}_i^h(s) = 1, \bar{Y}_i^h(t) = 1),$$

$$B_3^h = \sum_{x_i} \int_0^C \int_0^C P(X_i = x_i) P(\bar{Y}_i^h(s) = 1, \bar{Y}_i^h(t) = 1|x_i) \left\{ x_i - \frac{s^{(1,h)}(\gamma_1, s)}{s^{(0,h)}(\gamma_1, s)} \right\} \left\{ x_i - \frac{s^{(1,h)}(\gamma_1, t)}{s^{(0,h)}(\gamma_1, t)} \right\}$$
$$\times E(dN_{i1}(s)|x_i, \bar{Y}_i^h(s) = 1, \bar{Y}_i^h(t) = 1) dR_{01}^h(t) e^{x_i \gamma_1} ,$$

and

$$B_4^h = \sum_{x_i} \int_0^C \int_0^C P(X_i = x_i) P(\bar{Y}_i^h(s) = 1, \bar{Y}_i^h(t) = 1)$$
$$\times \left\{ x_i - \frac{s^{(1,h)}(\gamma_1, s)}{s^{(0,h)}(\gamma_1, s)} \right\} \left\{ x_i - \frac{s^{(1,h)}(\gamma_1, t)}{s^{(0,h)}(\gamma_1, t)} \right\} e^{2x_i \gamma_1} dR_{01}(s) dR_{01}(t) .$$

In Section 4.1 and under the assumption of time-homogeneous rate function for the two processes,

$$E(dN_{i1}^2(t)|x_i, \bar{Y}_i^A(t) = 1) = E(dN_{i1}(t)|x_i, \bar{Y}_i^A(t)) = P(Y_{i1}(t) = 1|x_i) \lambda_{01} \exp(x_i \beta_1),$$

and

$$E(dN_{i1}(s) dN_{i1}(t)|x_i, \bar{Y}_i^A(s) = 1, \bar{Y}_i^A(t) = 1)$$
$$= P(\bar{Z}_i(s^-) = 1|\bar{Z}_i(0) = 1, x_i) P(\bar{Z}_i(t^-) = 1|\bar{Z}_i(s) = 2, x_i) \lambda_{01}^2 \exp(2x_i \beta_1)$$

for $s < t$, where $P(Y_{i1}(t) = 1|x_i) = P(\bar{Z}_i(t^-) = 1|\bar{Z}_i(0) = 1, x_i)$, and $P(\bar{Z}_i(t^-) = 1|\bar{Z}_i(s) = 2, x_i)$ is given in Appendix A. In the setting of Section 4.2 with dependent random effects,

$$E(dN_{i1}^2(t)|x_i, \bar{Y}_i^A(t) = 1) = E(dN_{i1}(t)|x_i, \bar{Y}_i^A(t) = 1)$$
$$= \int_0^\infty \int_0^\infty u_{i1} P(Y_{i1}(t) = 1|u_i, x_i) \lambda_{01}(t) \exp(x_i \beta_1) dG(u_i),$$

and $E(dN_{i1}(s) dN_{i1}(t)|x_i, \bar{Y}_i^A(s) = 1, \bar{Y}_i^A(t) = 1)$ is given by

$$\int_0^\infty \int_0^\infty u_{i1}^2 P(\bar{Z}_i(s^-) = 1|\bar{Z}_i(0) = 1, x_i, u_i) P(\bar{Z}_i(t^-) = 1|\bar{Z}_i(s) = 2, x_i, u_i) \lambda_{01}^2 \exp(2x_i \beta_1) dG(u_i)$$

for $s < t$. Moreover

$$E(dN_{i1}^2(t)|x_i, \bar{Y}_i^B(t) = 1) = \frac{E(dN_{i1}^2(t)|x_i, \bar{Y}_i^A(t) = 1)}{P(Y_{i1}(t) = 1|x_i)} ,$$

and

$$E(dN_{i1}(s) dN_{i1}(t)|x_i, \bar{Y}_i^B(s) = 1, \bar{Y}_i^B(t) = 1) = \frac{E(dN_{i1}(s) dN_{i1}(t)|x_i, \bar{Y}_i^A(s) = 1, \bar{Y}_i^A(t) = 1)}{P(Y_{i1}(s) = 1, Y_{i1}(t) = 1|x_i)}$$

for $s < t$, where $E(dN_{i1}(s)|x_i, \bar{Y}_i^B(s) = 1, \bar{Y}_i^B(t) = 1)$ is given by

$$\frac{\int_0^\infty \int_0^\infty u_{i1} g(u_i) P(\bar{Z}_i(s^-) = 1|\bar{Z}_i(0) = 1, x_i, u_i) P(\bar{Z}_i(t^-) = 1|\bar{Z}_i(s) = 2, x_i, u_i) du_{i1} du_{i2} \lambda_{01} \exp(x_i \beta_1)}{P(Y_{i1}(s) = 1, Y_{i1}(t) = 1|x_i)}$$

for $s < t$ or

$$\frac{\int_0^\infty \int_0^\infty u_{i1} g(u_i) P(\bar{Z}_i(s^-) = 1 | \bar{Z}_i(0) = 1, x_i, u_i) P(\bar{Z}_i(t^-) = 1 | \bar{Z}_i(s) = 1, x_i, u_i) du_{i1} du_{i2} \lambda_{01} \exp(x_i \beta_1)}{P(Y_{i1}(s) = 1, Y_{i1}(t) = 1 | x_i)}$$

for $t < s$.

From the asymptotic variance formula under RSD-A, $\mathcal{A}(\gamma_1^A) = B_1^A$ and $B_3^A = B_4^A$, which means $\mathcal{B}(\gamma_1^B) = \mathcal{A}(\gamma_1^A) + B_2^A - B_4^A$. However with RSD-A $B_2^A < B_4^A$ because the term

$$E\{dN_{i1}(s) dN_{i1}(t) | x_i, \bar{Y}_i^A(s) = 1, \bar{Y}_i^A(t) = 1\}$$

in $B_2^A$ is the joint conditional probability of a transition from state 1 to 2 at $s$ and $t$ given $X_i$ and the fact that they are in the exacerbation-free state at both $s$ and $t$; as a result at least one transition is required from state 2 to 1 over $(s, t)$. In contrast, $R_{01}^A(s)$ and $R_{01}^A(t)$ in $B_4^A$ only condition on being in the exacerbation-free state at times $s$ and $t$ separately. So $\mathcal{A}(\gamma_1^A) > \mathcal{B}(\gamma_1^A)$ and as a result, the naive standard error is greater than the robust standard error. Thus, robust variance estimates ensure protection against some forms of model misspecification but not under misspecification of the risk sets.