Published in final edited form as:
Statistics in Medicine (2021), **40**(16): 3823–3842
DOI: 10.1002/sim.9002

Semiparametric recurrent event vs time-to-first-event analyses in randomized trials: Estimands and model misspecification

YUJIE ZHONG

School of Statistics and Management,

Shanghai University of Finance and Economics, Shanghai, P.R. China

E-mail: zhong.yujie@mail.shufe.edu.cn

RICHARD J. COOK

Department of Statistics and Actuarial Science,

University of Waterloo, Waterloo, ON, N2L 3G1, Canada

Summary

Insights regarding the merits of recurrent event and time-to-first-event analyses are needed to provide guidance on strategies for analyzing intervention effects in randomized trials involving recurrent event responses. Using established asymptotic results we introduce a framework for studying the large sample properties of estimators arising from semiparametric proportional rate function models and Cox regression undermodel misspecification. The asymptotic biases and power implications are investigated for different data generating models, and we study the impact of dependent censoring on these findings. Illustrative applications are given involving data from a cystic fibrosis trial and a carcinogenicity experiment, following which we summarize findings and discuss implications for clinical trial design.

Keywords: Cox regression, estimands, multiplicative rate functions, power, recurrent events, robustness

1 Introduction

For many chronic diseases featuring recurrent clinical events it is natural to evaluate the effect of therapeutic interventions on the basis of event occurrence. Examples include studies of a respiratory disease featuring repeated exacerbations of symptoms (Cazzola et al., 2012), febrile seizure prevention trials (Strengell et al., 2009), neurological trials evaluating prophylactic interventions for epileptic seizures (Nakamura et al., 2017), and clinical trials of bone strengthening drugs for the prevention of fractures in osteoporosis (Ettinger et al., 1999). Traditionally attention has been restricted to delaying the occurrence of the first event using Cox regression, but over the last three decades there has been increasing appreciation that use of information on event occurrence after the first event will provide a more comprehensive reflection of the effect of treatment.

There has been considerable discussion among statisticians involved in pharmaceutical research and regulatory agencies about the use of recurrent events for the evaluation of therapeutic interventions in clinical trials. Methods based on rate functions with multiplicative covariate effects are among the most widely appealing methods (Cook and Lawless, 2007) and are seeing increased use. There are two primary challenges with these analyses. First for valid inference, treatment comparisons based on rate functions must address between-individual heterogeneity in the propensity for events. Mixed Poisson models which incorporate individual-level random effects (Lawless, 1987a; Therneau and Grambsch, 2000) offer one approach for achieving this. Alternatively if robust methods are of interest, multiplicative models based on marginal rate functions can be fitted using Poisson estimating equations if robust sandwich variance estimates are used (Andersen and Gill, 1982; Lawless and Nadeau, 1995; Lin et al., 2000). The second challenge is that event occurrence informs treating physicians that current therapy may be ineffective for particular individuals, which can lead to early study withdrawal. This creates a type of event-dependent censoring which can affect properties of standard estimators (Strawderman, 2000; Cook et al., 2009). For parametric or semiparametric likelihood-based analyses, this form of dependent censoring is ignorable in the sense that estimators, while less efficient due to shorter follow-up, remain consistent if the response model assumptions are valid. When the goal is to carry out robust analyses methods are typically based on incompletely specified models with estimation and inference based on estimating functions. In this case, event-dependent censoring leads to inconsistent estimators unless inverse probability of censoring weighted estimating functions are used (Cook et al., 2009, 2010; Akacha and Ogundimu, 2016). These and other approaches (Wei et al., 1989; Prentice et al., 1981) have been studied in terms of the large sample properties of estimators (Boher and Cook, 2006; Zhong and Cook, 2019) and via simulation (Kelly and Lim, 2000), but there remains uncertainty about the advantages, limitations, and interpretation of findings from recurrent event analyses in comparison to Cox regression for the time to the first event. In particular there has been increased recent attention on the nature and interpretation of estimands in the context of recurrent events (Akacha and Ogundimu, 2016; Akacha, 2019; Stark, 2018; Lee and Cook, 2019; Roger et al., 2019). Some view recurrent event analyses as leading to estimators which are more efficient than estimators based on Cox regression for the time to the first event; this was shown to be in the case for Poisson recurrent event processes (Cook, 1995). More generally, however, the limiting value of estimators from Cox regression model and marginal rate-based models are different; we highlight this by exploring the determinants of these limiting values. We consider several particular recurrent event process models within a broad family, and evaluating the limiting values of estimators from Cox regression and marginal semiparametric rate-based models (Andersen and Gill, 1982). Clear interpretation of estimators is of central importance, but clinical researchers and regulators often make decisions based on the results of hypothesis tests so we also use large sample theory

to investigate the power implications on tests of the null hypothesis of no treatment effect in these frameworks. The work is motivated in part by the need for recommendations on the analysis of recurrent events in clinical trials (Agency, 2019).

The remainder of this article is organized as follows. In Section 2, we describe a general class of models for generating a censored recurrent event process and review the estimating functions and large sample theory for the semiparametric Andersen-Gill model (Andersen and Gill, 1982) and Cox regression Cox, 1972. We describe a framework to study estimands from recurrent event and Cox analyses in Section 3.1 and consider the limiting behaviour of estimators from the rate-based analysis and Cox regression for special cases including Poisson process (Section 3.2.1) and mixed Poisson processes (Section 3.2.2). Markov processes are considered in Sections 3.3. In Section 4, we discuss the impact of dependent right-censoring. Illustrative applications are given in Section 5 and we conclude with a summary of the findings from this work and make recommendations on the design and analysis of future studies in Section 6.

2 Notation and Model Formulation

2.1 Formulation of a Joint Recurrent Event and Censoring Process

Let T_k denote the time of the kth event in a recurrent event process and let $\{N(s), 0 < s\}$ be a right-continuous counting process recording the number of events over time $s > 0$ so that $dN(t) = 1$ if an event occurs at time t and $dN(t) = 0$ otherwise. Figure 1 (a) gives a multistate representation of a recurrent event process with transitions to the right taking place upon event occurrence. Here, T_k is the entry time to state k and $N(t)$ records the state occupied at time $t > 0$. We consider the setting of a clinical trial and so let X denote a fixed covariate vector; in many of the investigations that follow we let X be a binary scalar variable indicating whether an individual received the experimental treatment ($X = 1$) or standard of care ($X = 0$), but we present the results for the case that X is a vector in what follows. While they can be useful when formulating suitable models for life history processes and dynamic path analysis (Aalen et al., 2008), we do not consider time-dependent covariates here as they do not play a central role in the assessment of randomized interventions - our focus here.

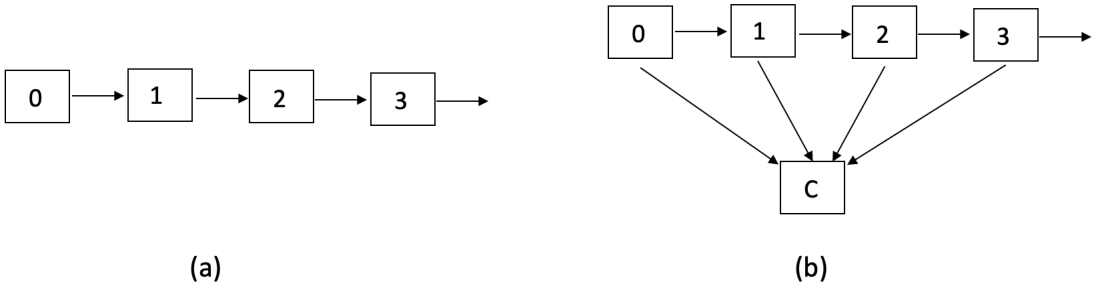

Figure 1: Multistate models depicting a recurrent event process (panel (a)) and a joint recurrent event and random censoring process (panel (b)).

We also let U denote an auxiliary scalar covariate with $E(U) = 1$, $\text{var}(U) = \phi$, and $U \perp X$. Then we let $\mathcal{H}(t) = \{[N(s), 0 < s < t], X, U\}$ denote the history of the process which includes the number and times of events over $(0, t)$ and the covariates X and U. The intensity function for the event process is given by

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta N(t) = 1 | \mathcal{H}(t))}{\Delta t} = \lambda(t|\mathcal{H}(t)) \ , \quad t > 0 \ , \tag{1}$$

where $\Delta N(t) = N(t + \Delta t^-) - N(t^-)$ records the number of events over $[t, t + \Delta t)$ and $\lim_{\Delta t \downarrow 0} \Delta N(t) = dN(t)$. The intensity function fully defines the event generating process (Cook and Lawless, 2007).

In clinical trials failure time responses are subject to administrative and random censoring. Let $A > 0$ be an administrative censoring time defined by the dates of accrual and the date the study will be closed. These administrative censoring times often vary across individuals when participants are recruited serially over time. Individuals may also withdraw from a study and we let R denote a random withdrawal time giving $C = \min(R, A)$ as the net censoring time. Then $Y(t) = I(t \leq C)$ indicates whether the individual is on study and at risk of failure at time $t > 0$. We let $d\bar{N}(t) = Y(t)dN(t) = 1$ if an event occurs *and is observed* at time t, with $d\bar{N}(t) = 0$ otherwise. $\bar{N}(t) = \int_0^t d\bar{N}(s)$ counts the cumulative number of events observed over $(0, t]$, and $\Delta\bar{N}(t) = \bar{N}(t + \Delta t^-) - \bar{N}(t^-)$.

Figure 1 (b) contains a multistate diagram depicting the joint recurrent event and censoring process which has a countably infinite set of states recording the number of observed recurrent events, and an absorbing state entered upon random censoring. We let $\bar{\mathcal{H}}(t) = \{[Y(s), \bar{N}(s), 0 < s < t], X, U\}$ be the history of the joint process involving the censored recurrent event process and the random censoring time. The intensity for the recurrent event process in the context of this observation scheme is

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta\bar{N}(t) = 1|\bar{\mathcal{H}}(t))}{\Delta t} = \bar{\lambda}(t|\bar{\mathcal{H}}(t)) \tag{2}$$

If the random censoring process is conditionally independent of the recurrent event process then we can write

$$\bar{\lambda}(t|\bar{\mathcal{H}}(t)) = Y(t)\lambda(t|\mathcal{H}(t)) \,, \tag{3}$$

and the intensities governing the $j \to j+1$ transitions in Figure 1 (b) are aligned with those in Figure 1 (a). This is essential for standard analysis of the data obtained under right-censoring scheme to yield information about the recurrent event process in the absence of censoring (Lawless and Cook, 2019).

Likewise we let $C^R(t) = I(R \leq t)$, so the counting process $\{C^R(s), 0 < s\}$ records the occurrence of random censoring; the counting process for random censoring is itself censored by the administrative censoring time A. The intensity for the random censoring time, corresponding to entry to State C in Figure 1 (b), is then

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta C^R(t) = 1|\bar{\mathcal{H}}(t))}{\Delta t} \;\; = \;\; Y(t)\gamma(t|\bar{\mathcal{H}}(t))$$

where $\Delta C^R(t) = C^R(t + \Delta t^-) - C^R(t^-)$. For parametric or semiparametric likelihood-based analyses this intensity can be dependent on the cumulative number of events and consistent results will be obtained for the recurrent event process (Cook et al., 2010). When robust marginal rate-based analyses are carried out inconsistent estimates will be obtained unless inverse probability of censoring weights are used. We therefore stress that the term "robust" is meant to convey robustness of the analysis to misspecification of the recurrent event process; these robustness properties do not hold when it comes to event-dependent withdrawal.

2.2 Large Sample Results for Semiparametric Regression for Recurrent Events

Statistical properties of the semiparametric proportional rate function model for recurrent events was developed by Andersen and Gill (1982) under a "working Poisson process" assumption with a rate function of the form

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta N(t) = 1|X)}{\Delta t} = \rho(t|X) = \rho_0(t)\exp(\beta X) \,, \tag{4}$$

where $\rho_0(t)$ is an unspecified baseline rate function and $\exp(\beta X)$ is a multiplicative term reflecting the effect of the covariate X. Lawless (1987b) gives the partial likelihood and associated estimating equations for the semiparametric setting and Andersen and Gill (1982) derive the large sample theory; the semiparametric model (4) is sometimes called the Andersen-Gill model. Lin et al. (2000) provide a rigorous derivation of the limiting behaviour of estimators with an emphasis on robust variance estimation.

Consider a sample of n independent individuals where the subscript i labels data from individual i, $i = 1, \ldots, n$. Under the assumptions that censoring is independent and non-informative (Cook and Lawless, 2007), the log partial likelihood is

$$\sum_{i=1}^{n} \int_0^\infty Y_i(t) \left\{ \log \rho(t|X_i) dN_i(t) - \rho(t|X_i) dt \right\} \ . \tag{5}$$

In the semiparametric setting of (4) we let $d\mu_0(t) = \rho_0(t)dt$ $(t > 0)$ so that $d\mu_0(\cdot)$ can be viewed as an infinite dimensional parameter. Differentiating the terms in (5) with respect to $d\mu_0(t)$ we obtain the estimating equations

$$\sum_{i=1}^{n} Y_i(t) \left\{ dN_i(t) - d\mu_0(t) \exp(\beta X_i) \right\} = 0 \ , \quad 0 < t \ . \tag{6}$$

The profile Breslow-type estimator $d\tilde{\mu}_0(t;\beta) = d\bar{N}_\cdot(t) / \sum_{i=1}^n Y_i(t) \exp(\beta X_i)$ is the solution where $d\bar{N}_\cdot(t) = \sum_{i=1}^n Y_i(t) dN_i(t)$. Differentiating (5) with respect to β and replacing $d\mu_0(t)$ with $d\tilde{\mu}_0(t;\beta)$ gives the profile partial score equation

$$U(\beta) = \sum_{i=1}^{n} U_i(\beta) = \sum_{i=1}^{n} \int_0^\infty Y_i(t) \left\{ X_i - \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)} \right\} dN_i(t) = 0 \ , \tag{7}$$

where $S^{(r)}(\beta, t) = n^{-1} \sum_{i=1}^n Y_i(t) \exp(\beta X_i) X_i^{\otimes r}$ with $X_i^{\otimes 0} = 1$, $X_i^{\otimes 1} = X_i$ and $X_i^{\otimes 2} = X_i X_i'$. The solution to (7) is denoted by $\widehat{\beta}$.

If the proportional rate function assumption in (4) is valid and censoring is independent given X then $\widehat{\beta}$ is consistent for β but more generally it is consistent for $\beta^\dagger$, the solution to

$$\int_0^\infty \left\{ s^{(1)}(t) - \frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)} s^{(0)}(t) \right\} dt = 0 \tag{8}$$

where $s^{(r)}(\beta, t) = E\{S^{(r)}(\beta, t)\}$ and $s^{(r)}(t)dt = E\{Y(t) X^{\otimes r} dN(t)\}$, $r = 0, 1, 2$ with the expectation $E\{ \cdot \}$ taken with respect to the true recurrent event, censoring, and covariate processes (Andersen and Gill, 1982). Lin et al. (2000) showed that

$$\sqrt{n}(\widehat{\beta} - \beta^\dagger) \to \mathrm{N}\left(0, \mathcal{A}^{-1}(\beta^\dagger) \mathcal{B}(\beta^\dagger) [\mathcal{A}^{-1}(\beta^\dagger)]'\right) \ , \tag{9}$$

where $\mathcal{A}(\beta) = E[-\partial U_i(\beta)/\partial\beta]$ and $\mathcal{B}(\beta) = E[U_i(\beta) U_i'(\beta)]$.

2.3 Large Sample Results for Cox Regression

Let $N^1(t) = I(T_1 \leq t)$ indicate that the first event has occurred by time t, $\{N^1(s), 0 < s\}$ denote the corresponding counting process, and $\mathcal{H}^1(t) = \{[N^1(s), 0 < s < t], X, U\}$ denote the history of the failure process with a fixed covariate X. The hazard for the first event is

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta N^1(t) = 1 | \mathcal{H}^1(t))}{\Delta t} = Y^1(t) \lambda^1(t | \mathcal{H}^1(t)) \ , \quad t > 0 \ , \tag{10}$$

where $\Delta N^1(t) = N^1(t + \Delta t^-) - N^1(t^-)$ and $Y^1(t) = I(t \leq T_1)$ indicates the first event has not occurred before time t^-.

If $Y(t) = I(t \leq C)$ as before and we let $\bar{Y}(t) = Y(t)Y^1(t)$, then $d\bar{N}^1(t) = \bar{Y}(t)dN^1(t) = 1$ indicates that the first event occurs *and is observed* at time t, with $d\bar{N}^1(t) = 0$ otherwise. Then $\bar{N}^1(t) = \int_0^t d\bar{N}^1(s)$ indicates that the first event has been observed over $(0, t]$ and $\bar{\mathcal{H}}^1(t) = \{[Y(s), \bar{N}^1(s), 0 < s < t], X, U\}$ is the corresponding history. The hazard for the observed event in the presence of right censoring is then

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta \bar{N}^1(t) = 1 | \bar{\mathcal{H}}^1(t))}{\Delta t} = \bar{\lambda}^1(t | \bar{\mathcal{H}}^1(t)) \ , \tag{11}$$

where $\Delta \bar{N}^1(t) = \bar{N}^1(t + \Delta t^-) - \bar{N}^1(t^-)$. If censoring is independent of the failure process given X, $\bar{\lambda}^1(t|\bar{\mathcal{H}}^1(t)) = \bar{Y}(t)\lambda^1(t|\mathcal{H}^1(t))$ and it is possible to learn about the underlying event intensity (or hazard) function.

The Cox model involves the further assumption that

$$\lambda^1(t|\mathcal{H}_i^1(t)) = \lambda_0(t) \exp(\eta X_i) \ ,$$

but we point out that the regression coefficient η has a different interpretation than the regression coefficient β in (4). The partial score equation under the Cox model is

$$U_1(\eta) = \sum_{i=1}^n U_{i1}(\eta) = \sum_{i=1}^n \int_0^\infty \bar{Y}_i(t) \left\{ X_i - \frac{R^{(1)}(\eta, t)}{R^{(0)}(\eta, t)} \right\} dN_i^1(t) = 0 \ , \tag{12}$$

where $R^{(\ell)}(\eta, t) = n^{-1} \sum_{i=1}^n \bar{Y}_i(t) \exp(\eta X_i) X_i^{\otimes \ell}$. The solution to (12) is denoted by $\widehat{\eta}$.

If the proportional rate function assumption in (4) is valid and censoring is independent then $\widehat{\eta}$ is consistent for η but otherwise it is consistent for $\eta^\dagger$, the solution to

$$\int_0^\infty \left\{ r^{(1)}(t) - \frac{r^{(1)}(\eta, t)}{r^{(0)}(\eta, t)} r^{(0)}(t) \right\} dt = 0 \tag{13}$$

where $r^{(\ell)}(\eta, t) = E\{R^{(\ell)}(\eta, t)\}$ and $r^{(\ell)}(t)dt = E\{\bar{Y}(t)X^{\otimes \ell} dN^1(t)\}$, $\ell = 0, 1, 2$ with the expectation $E\{\cdot\}$ taken with respect to the true failure time, censoring, and covariate processes. Moreover, Lin and Wei (1989) showed that

$$\sqrt{n}(\widehat{\eta} - \eta^\dagger) \to \mathrm{N}\left(0, \mathcal{A}_1^{-1}(\eta^\dagger)\mathcal{B}_1(\eta^\dagger)[\mathcal{A}_1^{-1}(\eta^\dagger)]'\right) \ , \tag{14}$$

where $\mathcal{A}_1(\eta) = E[-\partial U_{i1}(\eta)/\partial \eta]$ and $\mathcal{B}_1(\eta) = E[U_{i1}(\eta)U_{i1}'(\eta)]$.

In Section 3, we explore the behaviour of the limiting values $\beta^\dagger$ and $\eta^\dagger$ by computing their values in some specific settings within a class of models. The purpose is to contrast the interpretation of the estimands and give remarks to help guide the selection of target estimands in the design of clinical trials which we do in Section 6.1.

3 Estimands for Some Particular Underlying Models

3.1 A Framework to Study Estimands from Recurrent Event and Cox Analyses

While the general intensity functions were defined in Section 2.1, here we consider a specific formulation which facilitates the study of particular recurrent event models and censoring processes in order to discuss the process features which influence the interpretation of estimands

from standard analyses. We also use this general model to highlight the fact that only in very special cases are the estimands from a Cox regression model and a recurrent event analysis the same. We focus on the case here with X a scalar binary indicator of assignment to the experimental treatment arm versus a control treatment.

We consider a general framework in which $U > 0$ is a scalar random variable with $E(U) = 1$, $\text{var}(U) = \phi$, and (1) has the form

$$\lambda(t|N(t^-) = k, \mathcal{H}(t)) = \lambda_k(t) \exp(\beta_k X + \nu_1 \log U) \ , \quad k = 0, 1, \ldots . \tag{15}$$

This general formulation accommodates a conditional intensity with event-dependence by allowing $\lambda_k(t) \neq \lambda_{k-1}(t)$, heterogeneity in the intensity between individuals, and complex effects of treatment when the regression coefficients vary according to the cumulative number of events. We next discuss some particular models.

3.1.1 Mixed-Poisson Processes

Note that with $\lambda_k(t) = \lambda_{k-1}(t)$, $k = 1, 2, \ldots$ the baseline intensity is a rate function sometimes denoted by $\rho_0(t)$, and does not depend on the cumulative number of events. If in addition $\beta_k = \beta$, $k = 1, 2, \ldots$, then there is also a common treatment effect. Under the further restriction that $\nu_1 = 0$ (or $\phi = 0$) this model reduces to a Poisson process (Lawless, 1987b) and if $\nu_1 \neq 0$ and $\phi > 0$ then a mixed Poisson model is obtained; for a negative binomial process (Lawless, 1987a), we require $\nu_1 = 1$ with U gamma distributed with mean 1 and variance ϕ.

3.1.2 Recurrent Events Arising from Mixed-Markov Processes

If $\nu_1 = 0$ (or $\phi = 0$) but $\lambda_k(t) \neq \lambda_{k-1}(t)$ for $k = 1, \ldots$, then a Markov model is obtained; a common treatment effect arises if $\beta_k = \beta_0$, $k = 1, 2, \ldots$; this model is compatible with the assumptions of the so-called Prentice-Williams-Peterson analysis which was described in Prentice et al. (1981) outside of the context of randomized trials. Zhong and Cook (2019) show that the validity of this model for causal inference in randomized clinical trials hinges on the assumption that $\nu_1 = 0$ (or $\phi = 0$), since if there are any omitted covariates confounding is induced through this stratification even in the presence of baseline randomization. The most general model within this framework occurs when $\lambda_k(t) \neq \lambda_{k-1}(t)$ for some $k \in \{1, \ldots, K\}$ and $\beta_k \neq \beta_{k-1}$ for some $k \in \{1, \ldots, K\}$ and $\phi > 0$ with $\nu_1 \neq 0$.

3.1.3 Processes Involving Dependent Censoring

When viewed in the context of the multistate model in Figure 1 (b), we can also consider a general class of censoring models with intensities of the form

$$\gamma(t|\bar{N}(t^-) = k, \bar{\mathcal{H}}(t)) = Y(t)\gamma_k(t) \exp(\psi X + \nu_2 \log U) \ , \quad k = 0, 1, \ldots , \tag{16}$$

where if $\gamma_k(t) = \gamma(t)$, $k = 1, 2, \ldots$, and $\nu_2 = 0$ then censoring is independent of the recurrent event process given X. If either $\gamma_k(t) \neq \gamma(t)$ for some k or ν_1 and ν_2 are non-zero with $\phi > 0$, then the random censoring time is not independent of the recurrent event process. In the former setting the intensities for censoring in Figure 1 (b) differ depending on the cumulative number of events observed, while for the latter case the dependence arises from shared time-fixed attributes; only the latter are at play when analyses are based on the time to the first event.

The model in (15) is quite general so in order to help in the discussion of particular models we add some structure. Specifically when baseline intensity functions and covariate effects depend on the cumulative number of events we suppose:

(i) $\beta_k = b_k\beta_{k-1}$, $k = 1, \ldots,$ and

(ii) $\lambda_k(t) = r_k\lambda_{k-1}(t)$, $k = 1, \ldots.$

If $b_k = 1$ for $k = 1, 2, \ldots,$ then the treatment effect is independent of the cumulative number of events, and if $r_k = 1$ for $k = 1, 2, \ldots,$ then we obtain a common conditional baseline rate function which we may write as $\lambda_k(t) = \rho_0(t)$, $k = 1, 2, \ldots,$. We consider this setting in Section 3.2. Regarding the censoring intensities we let:

(iii) $\gamma_k(t) = g_k\gamma_{k-1}(t)$, $k = 1, 2, \ldots,$

so that if $g_k = 1$ for $k = 1, 2, \ldots,$ there is no systematic dependence of the censoring rate on the cumulative number of events; censoring is completely independent if in addition $\phi = 0$ or $\nu_2 = 0$.

3.2 Investigations Under the Proportional Rate Model

Suppose $\beta_k = \beta$, $k = 1, 2, \ldots$ so that there is a common treatment effect, and moreover that $\lambda_k(t) = \lambda_{k-1}(t) = \rho_0(t)$, $k = 1, 2, \ldots$ so that the baseline intensity does not depend on the cumulative number of events.

3.2.1 Relative Efficiency Investigations for Poisson Processes

If $\nu_1 = 0$ or $\phi = 0$ then (15) reduces to the intensity for a Poisson process (Lawless, 1987b) and $\beta^\dagger = \eta^\dagger = \beta$ so it is reasonable to consider the efficiency gains from an Andersen-Gill analysis (Andersen and Gill, 1982) over Cox regression (Cox, 1972). Note that because it is a Poisson process $\mathcal{A}(\beta) = \mathcal{B}(\beta)$ in (9) and we have $\sqrt{n}(\widehat{\beta} - \beta) \to \mathrm{N}(0, \mathcal{A}^{-1}(\beta))$. Likewise for the Cox model $\mathcal{A}_1(\beta) = \mathcal{B}_1(\beta)$ in (14) so based on Cox regression $\sqrt{n}(\widehat{\beta} - \beta) \to \mathrm{N}(0, \mathcal{A}_1^{-1}(\beta))$. The asymptotic relative efficiency of the Andersen-Gill versus Cox estimator is defined here as $ARE(\beta) = \mathcal{A}^{-1}(\beta)/\mathcal{A}_1^{-1}(\beta)$ which will take on values less than 1 representing the improvement in precision from the Andersen-Gill estimator over the Cox estimator.

We set $A = 1$ to be a common administrative censoring time and consider a random R_i which is an independent exponential random variable with hazard rate γ_0 such that $P(R_i < A) = 0.20$ or 0.40 giving a 20% or 40% early withdrawal rate over the course of the study. We let $X_i \sim \mathrm{Bern}(0.5)$ and set $\rho_0(t) = \rho_0$ for a time homogeneous rate and let $\beta = \log 0.75$ or $\log 0.50$ to correspond to a 25% or 50% reduction in the rate of events with treatment. Figure 2 contains a plot of the ARE where the horizontal axis is the expected number of events in the control arm by the administrative censoring time (i.e. $E\{N(A)|X = 0\}$) with values ranging from 0.5 to 5. Different lines are plotted for the 20% and 40% early withdrawal rate and for the two different values of β. It is apparent that the Andersen-Gill analysis yields a more efficient estimator of treatment effect compared to Cox analysis based on the time to first event, with the efficiency gain becoming quite substantial as the expected number of events increases. Moreover, for a given point on the horizontal axis the efficiency gain from an Andersen-Gill analysis becomes attenuated as the magnitude of the treatment effect increases and as the censoring rate increases; theses have the effect of decreasing the total expected number of *observed* events in the trial.

3.2.2 Investigations Under Mixed Poisson Processes

If $\nu_1 = 1$ and $\phi > 0$ a mixed Poisson process is obtained; furthermore if U is gamma distributed then a negative binomial process is obtained (Lawless, 1987a). Here $\beta^\dagger = \beta$ but $\eta^\dagger \neq \eta$ since

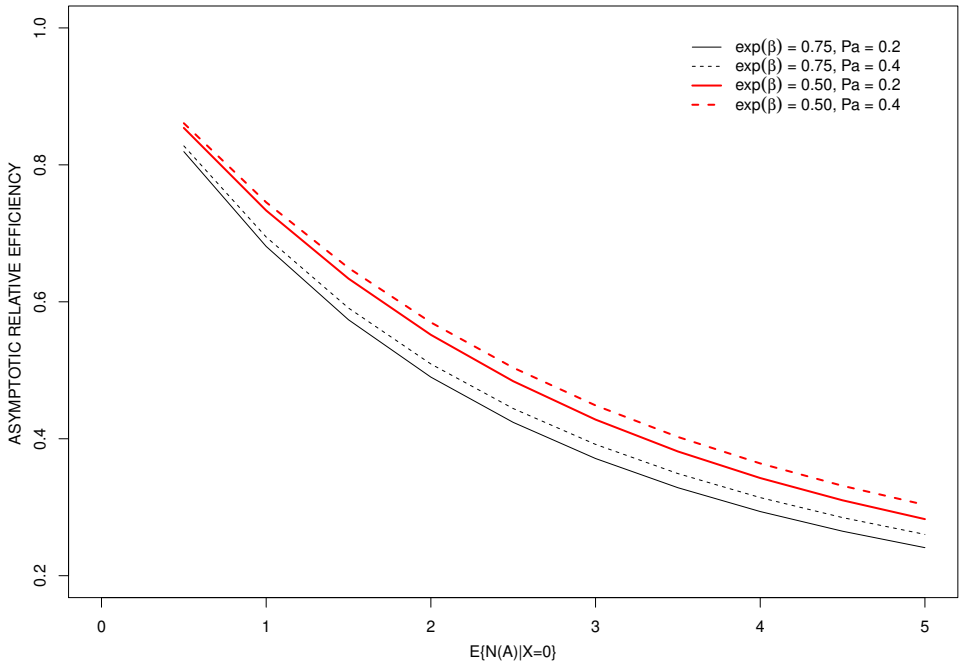

Figure 2: The asymptotic relative efficiency (ARE) for an estimator from an Andersen-Gill analysis versus a Cox analysis for different independent random censoring rates (20% and 40%) and treatment effects ($\exp(\beta) = 0.75$ and 0.50) as a function of $E\{N(A)|X = 0\}$.

the Cox model is misspecified, and $\beta \neq \eta$ the two analyses have incompatible estimands. Since $X \perp U$ due to randomization the survivor function for $T_1|X = x$ is

$$P(T_1 \geq t|X = x) = P(N(t) = 0|X = x) = \int_0^\infty P(N(t) = 0|X = x, U = u)dG(u) \ .$$

If U is gamma distributed then this is given by

$$\mathcal{F}(t|x) = \int_0^\infty \exp(-u\mu(t|x))\frac{u^{\phi^{-1}-1}\exp(-u/\phi)}{\Gamma(\phi^{-1})\phi^{\phi^{-1}}}du = \left(\frac{1}{1+\phi\mu(t|x)}\right)^{\phi^{-1}} \ ,$$

where $\mu(t|x) = \int_0^t \rho_0(s)ds\exp(\beta x) = \mu_0(t)\exp(\beta x)$ is the mean function for the recurrent event process and $\mu_0(t)$ is the baseline mean function. The cumulative hazard function is $H(t|x) = -\log\mathcal{F}(t|x) = \phi^{-1}\log(1+\phi\mu_0(t)\exp(\beta x))$, the hazard function is $h(t|x) = \rho_0(t)\exp(\beta x)/(1+\phi\mu_0(t)\exp(\beta x))$, and the hazard ratio is

$$\frac{h(t|x=1)}{h(t|x=0)} = \frac{\rho_0(t)e^\beta/(1+\phi\mu_0(t)e^\beta)}{\rho_0(t)/(1+\phi\mu_0(t))} = \left[\frac{1+\phi\mu_0(t)}{1+\phi\mu_0(t)e^\beta}\right]\exp(\beta) \ ,$$

which is a function of time unless $\beta = 0$ or we are in the setting of Section 3.2.1 where $\phi = 0$. As a result the Cox model is misspecified and we rely on the results of Struthers and Kalbfleisch (1986) and Lin and Wei (1989) to determine the limiting value $\eta^\dagger$ via (13). As noted by Wu and Cook (2012) and Rufibach (2019) when a failure time model is misspecified the limiting value will depend on the censoring distribution. We next explore the other determinants of the limiting value of the estimator from the Cox model.

Without loss of generality we set $A = 1$ and let $\beta = \log 0.5$. Figure 3 (a) contains plots of the limiting value of the Cox model estimator $\eta^\dagger$ against the variance of the random effect ϕ

with independent random censoring with a 40% probability of early withdrawal for $\mu(A|X = 0) = 0.5$, 1, 2 and 4. To help understand the role of random censoring in determining the limiting value of the estimator, we display the probability of observing the first event by the administrative censoring time by evaluating the cumulative incidence function for the first event at time A when treating random censoring as a competing event; this probability is displayed in Figure 3 (b) as a function of the expected number of events and the extent of extra-Poisson variation. The plots in Figure 3 (a) show how the estimands for the treatment effect under the Cox model become more attenuated the greater the extra-Poisson variation, with the extent of attenuation greater when the event rate is bigger and the effective censoring rates are lower.

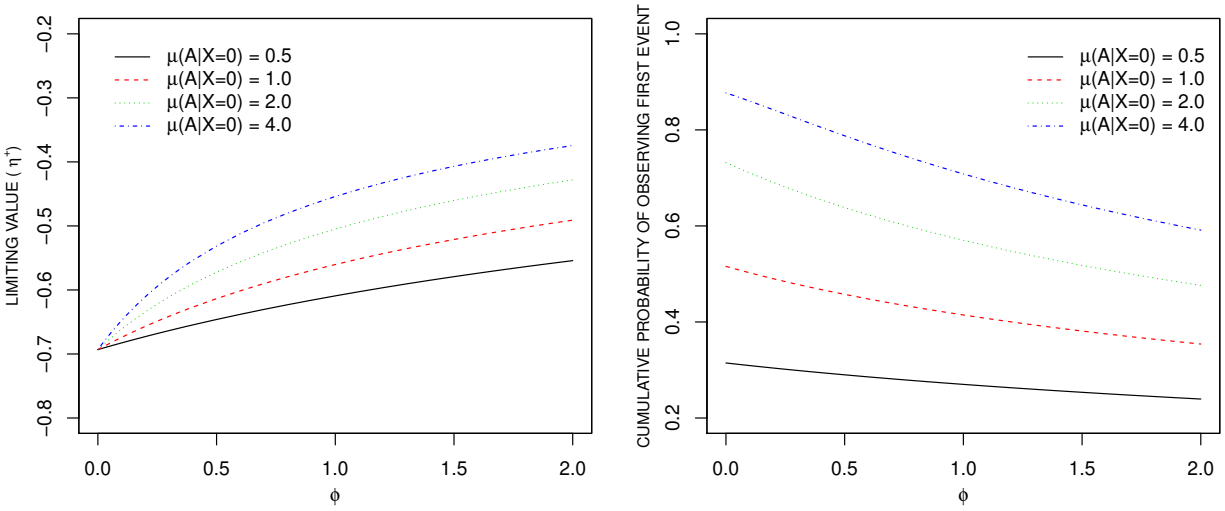

Figure 3: The limiting value of regression coefficient from a Cox regression model when the data are generated by a mixed Poisson process considering different degrees of extra-Poisson variation reflected by ϕ, different expected number of events by the administrative censoring time in the control arm ($\mu(A|X = 0)$) (panel (a)) and the cumulative probability that the first event is observed (panel (b)).

Wald tests can be carried out based on estimates from the Andersen-Gill and Cox models — the power of these tests are affected by both the limiting values and the asymptotic variances of the respective estimators. While Wald tests are most appropriate when based on estimators with a clear interpretation, hypothesis testing is a core part of evaluating an experimental intervention, and tests for effects in superiority trials are often prescribed in a protocol before any data are available and model assumptions can be assessed. In general, however, the null hypothesis of no treatment effect will be rejected based on a Wald test at a rate compatible with the nominal level, provided it is based on robust variance estimates (Boher and Cook, 2006). Finally, while we consider Wald-based tests, these will have frequency properties similar to the analogous pseudo-score tests routinely applied.

Consider Wald tests for the parameter of interest, ξ, which we used to represent β under the Andersen-Gill model and η under the Cox model. If the null hypothesis is $H_0 : \xi = 0$ and the alternative is $H_A : \xi \neq 0$, then the "asymptotic" power of the Wald tests at significance

level ω for samples of size m can be calculated as

$$\begin{aligned}\text{power} &= P\left(\left|\hat{\xi}/\sqrt{\text{asvar}_0(\hat{\xi})}\right| > z_{\omega/2}\middle| H_A\right) \\ &= 1-\Phi\left(\frac{z_{\omega/2}\sqrt{\text{asvar}_0(\hat{\xi})}-\sqrt{m}\,\xi^\dagger}{\sqrt{\text{asvar}_A(\hat{\xi})}}\right)+\Phi\left(\frac{-z_{\omega/2}\sqrt{\text{asvar}_0(\hat{\xi})}-\sqrt{m}\,\xi^\dagger}{\sqrt{\text{asvar}_A(\hat{\xi})}}\right), \qquad (17)\end{aligned}$$

where $\text{asvar}_0(\hat{\xi})$ and $\text{asvar}_A(\hat{\xi})$ are the asymptotic variances of the estimates under the null and alternative hypothesis, which can be evaluated by (9) under the Andersen-Gill model and by (14) under the Cox model, respectively. Moreover, $\xi^\dagger$ is the limiting value of ξ which we have denoted by $\beta^\dagger$ or $\eta^\dagger$ under the Andersen-Gill or Cox regression models, respectively. We therefore evaluate the relative power of Andersen-Gill and Cox regression analyses by plotting the power curves based on Wald tests using the asymptotic variance formula for samples of size m where m is determined to give 80% power to detect a 25% or 50% reduction in the risk of events with $E\{N(A)|X=0\}=2$ and $\phi=2$. Figure 4 shows that when the true value of treatment effect is the same as the one we used to determine the sample size, the power of Wald tests under the Andersen-Gill analysis with robust variance estimates can achieve the desired 80% power. Naturally as the treatment effect decreases, the power decreases for both analyses but the power of the Wald tests based on the Andersen-Gill model is always bigger than the corresponding test under the Cox model. The power of Wald test under the Andersen-Gill model with a naive variance estimate is greater than the tests based on the other analyses because it under-estimates the variation in the data and hence features an inflated type I error rate; robust variance estimation is therefore recommended for analyses based on rate functions.

3.3 Recurrent Events Arising from Markov Models

3.3.1 Limiting Value of the Estimator from an Andersen-Gill Model

If $\lambda_k(t) \neq \lambda_{k-1}(t)$ for some $k=1,\ldots,$ but $\beta_k=\beta$ for $k=1,2,\ldots,$ when $\nu_1=0$, then a Markov model is obtained with a common treatment effect. We let $r_k=e^\alpha$, for $k=0,\ldots,K+1$ so that the occurrence of each event increases the baseline rate of the next event up until the $(K+1)$st event and set $r_k=1$ for $k=K+2,\ldots,K_m$ so that the risk does not increase beyond the $(K+1)$st event. We consider a maximum number of recurrent events of K_m, so data are generated based on the multistate model with state 0 and K_m event states; we choose K_m to be sufficiently large that the probability of entering the absorbing state K_m over the planned period of observation is close to zero.

Time-homogeneous transition intensities are obtained by letting $\lambda_0(t)=\rho_0$. We let $\mathbb{Q}$ denote the $(K_m+1)\times(K_m+1)$ transition intensity matrix with (j,j) entries $-\lambda_{j-1}$, $(j,j+1)$ entries λ_{j-1} above the diagonal, and all other entries zero. The Chapman-Kolmogorov equations then give,

$$\mathbb{P}(s,s+t|X=0)=\exp(\mathbb{Q}t)\,, \qquad (18)$$

where $\mathbb{P}(s,s+t|X=0)=\mathbb{P}(0,t|X=0)$ and $P_{jl}(0,t|X=0)=P(N(t)=l|N(0)=j,X=0)$ (Cox and Miller, 1965). Here, $\eta^\dagger=\eta=\beta$ but $\beta^\dagger\neq\beta$ since the Andersen-Gill model ignores the state-dependent transition intensity; the Cox model for the time to the first event is still valid.

We let $\alpha=\log 1.2$ so there is a 20% increase in the risk of an event each time an event occurs up to $K=5$, and let $K_m=20$. We focus on $\beta=\log 0.5$ and $\log 0.75$, and consider 20%

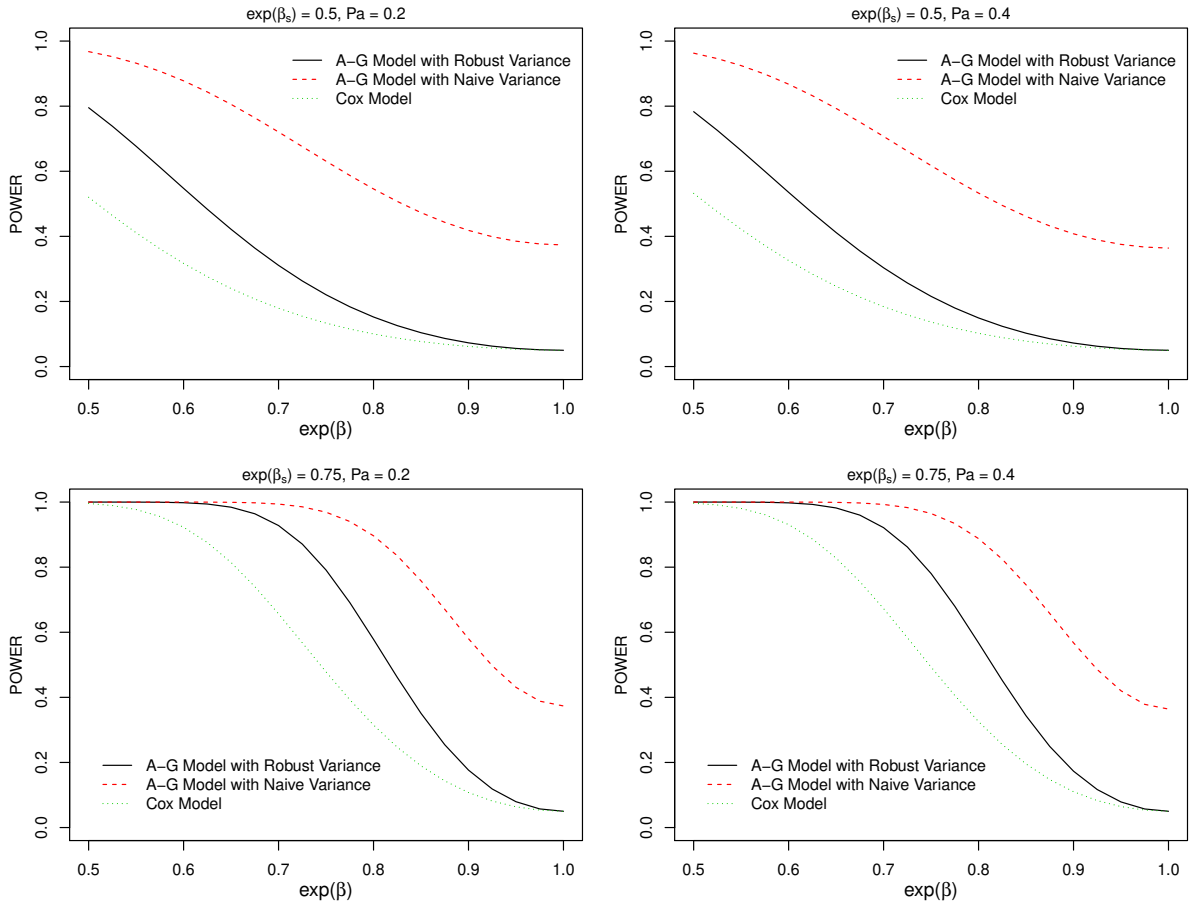

Figure 4: Power of Wald tests under Andersen-Gill and Cox regression models for samples of size m to give 80% power to detect a 25% or 50% reduction in the risk of events with $E\{N(A)|X=0\}=2$ and $\phi=2$ under a gamma-Poisson (negative binomial) data generating process.

and 40% early withdrawal. We determine ρ_0 so that $\mu(A|X=0)=0.5,1,2$ and 4, where

$$\mu(t|X=0)=\sum_{k=0}^{K_m} k\cdot P(N(t)=k|N(0)=0,X=0) \tag{19}$$

is the expected number of events at time t given $X=0$. Figure 5 shows the limiting bias of the Andersen-Gill model estimator $\beta^{\dagger}$ against the increase in risk of event each time (i.e. $\exp(\alpha)$). The lines illustrate how the estimands of the treatment effect of the Andersen-Gill model varies as a function of the increase of transition intensity each time an event occurs, the expected number of events in the control arm, treatment effect and the censoring rate. The magnitude of asymptotic bias of the estimates from the Andersen-Gill analysis increase when $\exp(\alpha)$ is far from 1 and the expected number of event increases. This is worse when the magnitude of treatment effect is large.

3.3.2 Power Considerations

We also consider the power of Wald tests based on Andersen-Gill and Cox regression analyses when the events follow a Markov process by plotting the power curves based on Wald tests using asymptotic variance formula for samples of size m to give 80% power to detect a 25% or 50% reduction in the risk of events with $E\{N(A)|X=0\}=2$ under the mixed Poisson model.

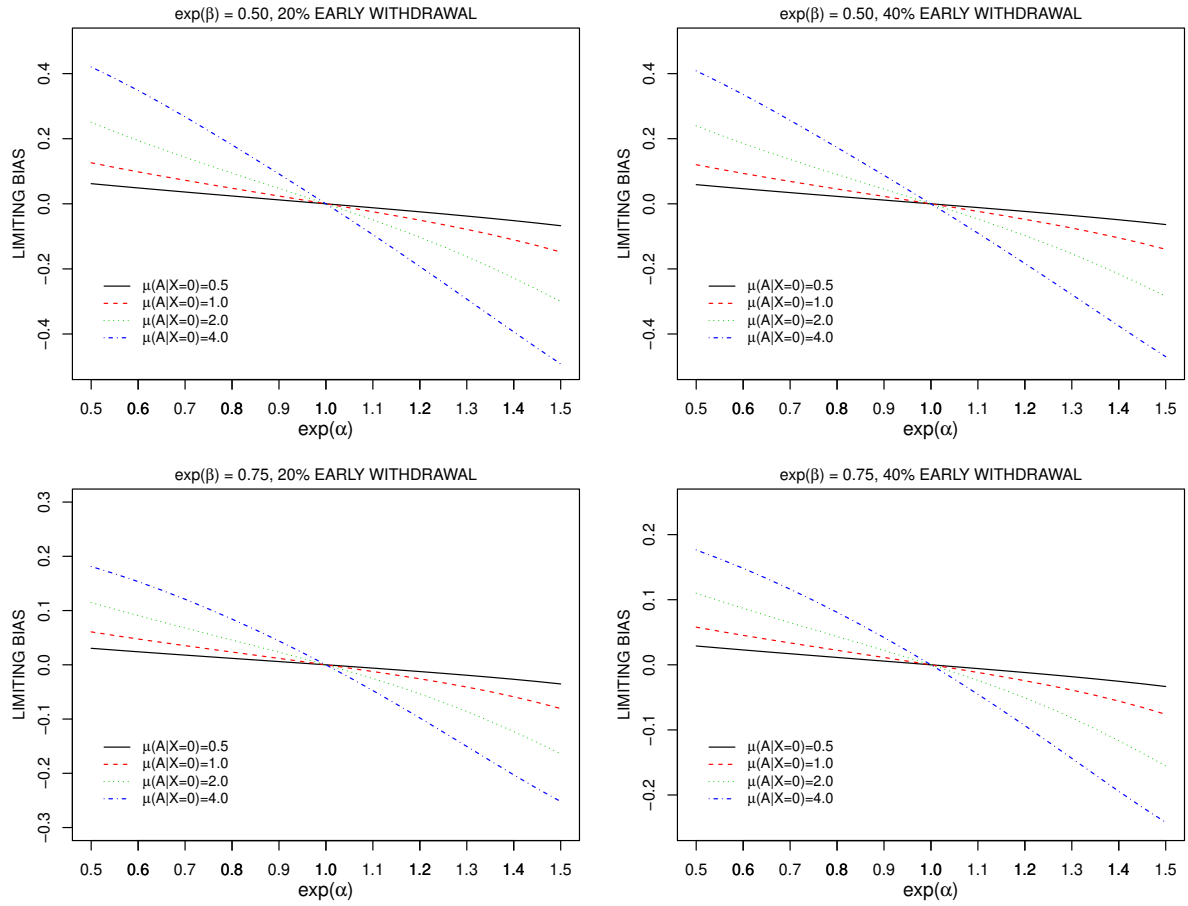

Figure 5: The limiting bias of regression coefficient from Andersen-Gill model as a function of the increase of transition intensity when the data are generated by a Markov process considering different expected number of events in the control arm ($\mu(A|X=0)$), treatment effect and early withdrawal rates.

Figure 6 shows that although the Andersen-Gill model ignores the state-dependent transition intensity the robust variance ensures it yields valid tests of the null hypothesis of no treatment effect in the sense that the type I error rate is compatible with the nominal 5% level. Although Cox regression gives a consistent estimate of the treatment effect, its asymptotic variance is quite large compared to the robust variance under the Andersen-Gill model, so the power of Cox regression analysis is less than that of Andersen-Gill analysis.

4 Implications of Dependent Censoring

Event-dependent censoring is at higher risk of occurring in settings with recurrent event outcomes since event occurrence can influence the way physicians treat patients, and certain kinds of treatment changes may lead to study withdrawal. Strategies for dealing with dependent censoring include joint modeling of the censoring and event processes through shared or correlated frailty parameters (Cook et al., 2010), joint multistate modeling (Cook et al., 2009), or use of inverse probability of censoring weights (Cook et al., 2009). Ghosh and Lin (2003) alternatively considered scale-change models for the event and dependent censoring processes, and leave the dependence structure unspecified. Artificial censoring was considered in their paper as another approach to mitigate bias arising from dependent censoring, this method had also been extended by Hsieh et al. (2011). The purpose here is not to explore how such methods

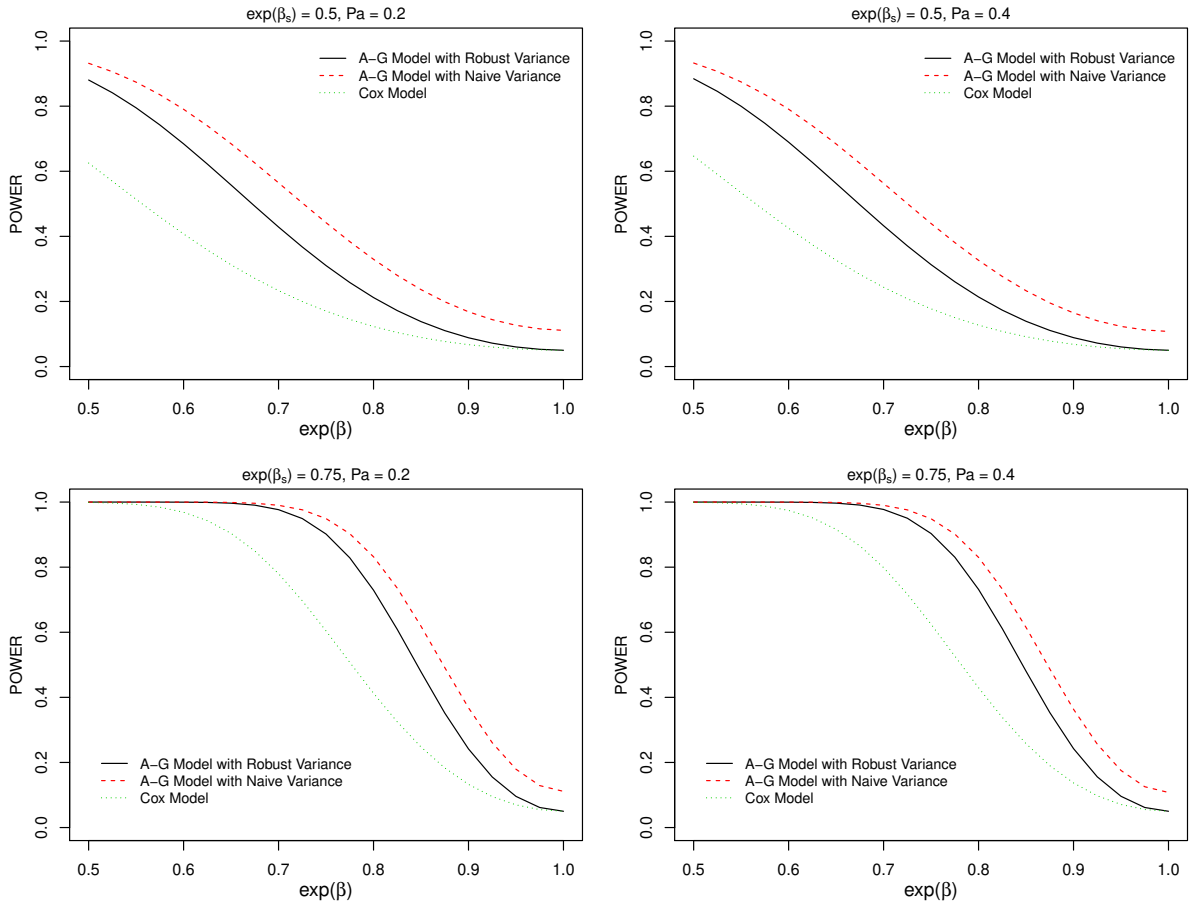

Figure 6: Power of Wald tests under Andersen-Gill and Cox regression models when events follow a Markov process; samples of size m is determined by giving 80% power to detect a 25% or 50% reduction in the risk of events with $E\{N(A)|X=0\}=2$ under mixed Poisson model.

can improve estimators but rather to investigate how dependent censoring may influence the limiting values of estimators arising from mixed Poisson and Markov processes.

4.1 Mixed Poisson Model and Correlated Random Effects

Here, we introduce a slightly more general formulation to facilitate a discussion about dependent censoring. We relabel the mean 1 gamma distributed random effect U in (15) as U_1 and denote its variance by ϕ_1 and we set $\nu_1 = 1$. We also relabel the mean 1 random effect U in (16) as U_2 and denote its variance by ϕ_2, and we set $\nu_2 = 1$. We assume that given the random terms U_1 and U_2, the recurrent event process and the censoring process is conditionally independent, and the recurrent event process given the random term U_1 is a Poisson process with rate $U_1\lambda(t)\exp(\beta X)$, while the random censoring has conditional intensity $U_2\gamma(t)\exp(\zeta X)$. Here, we consider time non-homogeneous processes with $\lambda(t) = \lambda^{\kappa_1}\kappa_1 t^{\kappa_1-1}$ and $\gamma(t) = \gamma^{\kappa_2}\kappa_2 t^{\kappa_2-1}$. We then use a Clayton copula to induce a dependence between the random terms U_1 and U_2 indexed by parameter θ so the joint cumulative distribution function for (U_1, U_2) is

$$F(u_1, u_2) = (F_1^{-\theta}(u_1\,;\phi_1) + F_2^{-\theta}(u_2\,;\phi_2) - 1)^{-1/\theta}\,,$$

where $F_1(u_1\,;\phi_1)$ and $F_2(u_2\,;\phi_2)$ are the marginal cumulative distribution function for U_1 and U_2, respectively. The strength of the association between U_1 and U_2 is reflected by Kendall's τ defined in terms of θ as $\tau = \theta/(\theta+2)$ (Nelson, 2006).

When $\zeta = 0$, even when $\{N(s), 0 < s\}$ and R are associated because Kendall's $\tau \neq 0$, fitting the Andersen-Gill model yields a consistent estimate of the treatment effect (i.e. $\beta^\dagger = \beta$). Note, however, that $\beta^\dagger \neq \beta$ if $\zeta \neq 0$ *and* Kendall's $\tau \neq 0$. The estimator from a Cox regression analysis is also inconsistent in general for reasons discussed in Section 3.2.2 but the dependent censoring changes the limiting value. Figure 7 shows the limiting value of the estimators from an Andersen-Gill model and Cox model as a function of Kendall's τ. We set the expected number of events in the control group to be $\mu(A|X = 0) = 2$, $\beta = \log 0.75$ and set the random censoring rate to give 20% early withdrawal. As noted above when $\zeta = 0$, the Andersen-Gill model gives consistent estimate of the treatment effect, but it leads to biased estimates when $\zeta = \log 1.5$ and when the censoring process and event process are dependent. The bias increases with the strength of the dependence between these two processes. Interestingly, while the values of κ_1 and κ_2 influence the properties of the estimators of the baseline rate or hazard functions, they do not impact the limiting value of the regression coefficients from the Andersen-Gill or Cox analyses. The Cox regression analyses gives inconsistent estimate of the treatment effect, but the bias becomes smaller as the dependence between the event and censoring processes increases; and the bias is bigger when $\zeta = \log 1.5$ compared with those when $\zeta = 0$ under the Cox model.

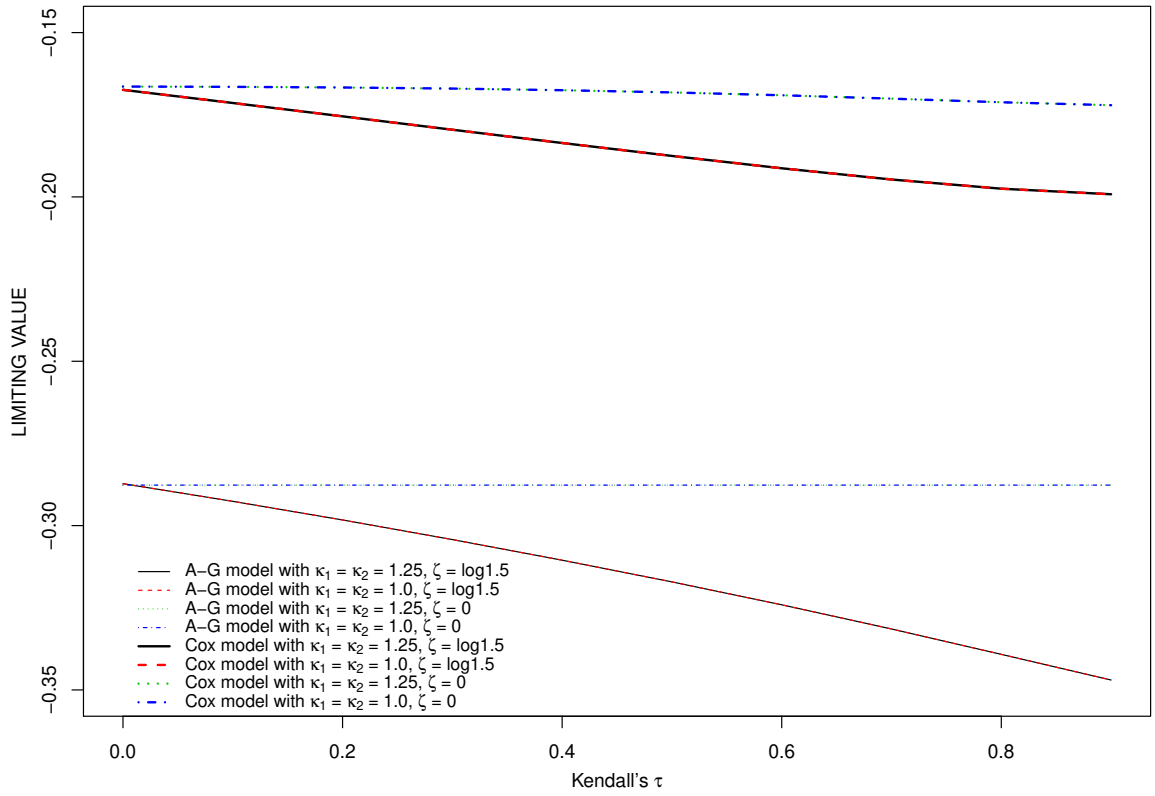

Figure 7: Limiting value of estimators from the Andersen-Gill and Cox models under dependent censoring with events generated according to a mixed Poisson process

4.2 Markov Models with State-Dependent Censoring Process

We now consider another scenario where the dependent censoring arise because its intensity depends on the previous cumulative number of events, that is, event-dependent censoring. Let

the event process is a Markov process with transition intensity $\lambda_k(t)$ from state k to state $k+1$, $\lambda_k(t) = \lambda(t|\bar{N}(t^-) = k, \bar{\mathcal{H}}(t)) = \lambda_k \exp(\beta X)$. We consider $\lambda_k = \exp(k\alpha)$ for $k = 0, 1, \ldots, K+1$ and then $\lambda_k = \exp((K+2)\alpha)$ for $k = K+2, \ldots, K_m$; we consider $K = 5$, $K_m = 20$. The intensity for censoring process from state k to a censoring state is $\gamma(t|\bar{N}(t^-) = k, \bar{\mathcal{H}}(t)) = Y(t)\gamma_0 g^k$, where g reflects the increase of risk of censoring when the cumulative number of events increase.

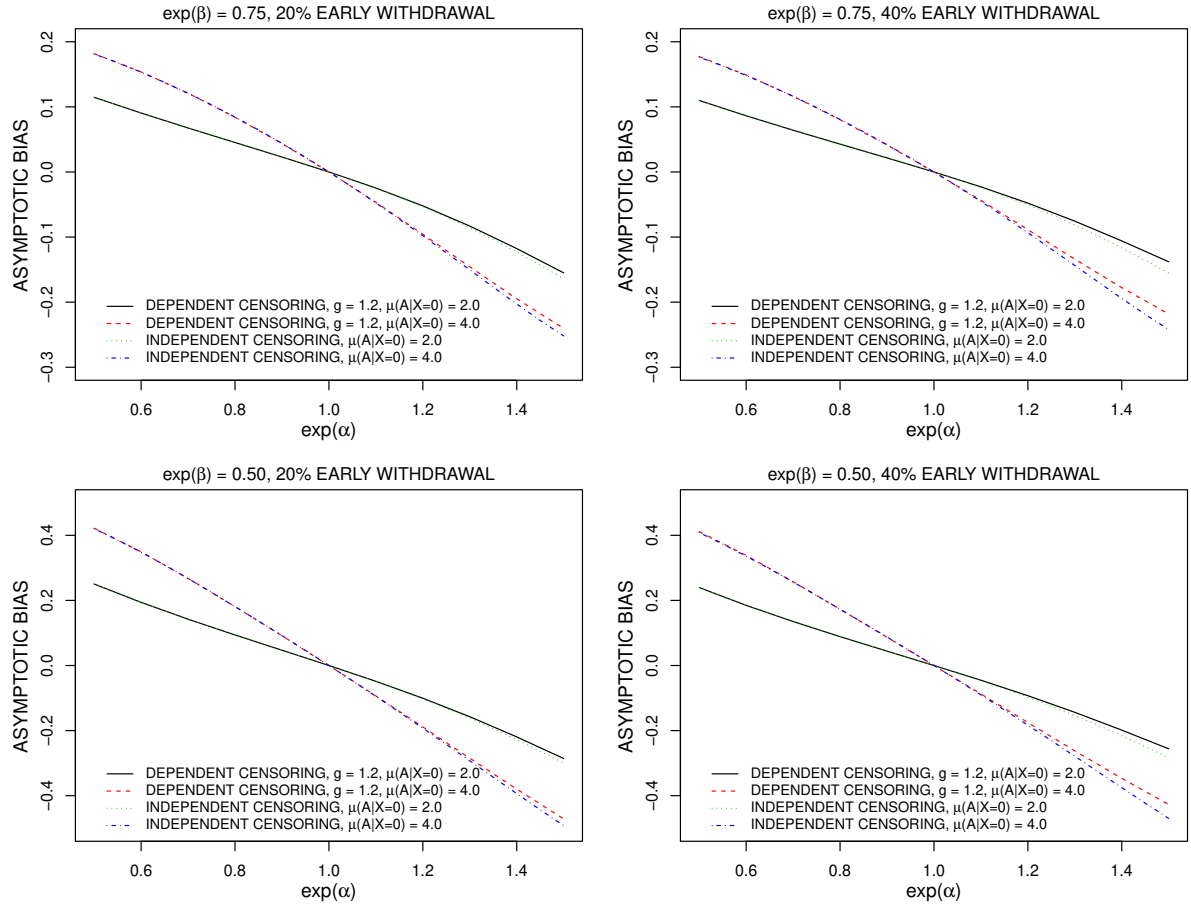

Figure 8: Limiting bias of estimators under Andersen-Gill model as a function of $\exp(\alpha)$ when the event process is a Markov process

When the censoring intensity depends on the cumulative number of events, the conditionally independent censoring assumption required for the Andersen-Gill model is not satisfied, but the estimator from the Cox regression analysis is consistent since this form of event dependent censoring is not manifest before the first event. We derive the limiting value of estimators under the Andersen-Gill model and study the limiting behavior of estimates from both models. Figure 8 illustrates the limiting bias under an Andersen-Gill model as a function of $\exp(\alpha)$, the parameter that reflects the increase in risk upon the occurrence of each event. This trend is given for different values of the mean number of events in the control arm ($\mu(A|X = 0)$), treatment effects given by $\beta = \log 0.75$ and $\log 0.5$, and different early withdrawal rates for both independent ($g = 1.0$) add dependent ($g = 1.2$) censoring. We see that as $\exp(\alpha)$ becomes further from 1, the magnitude of the limiting bias of the Andersen-Gill estimator increases under dependent and independent censoring. When the expected number of events or the magnitude of treatment effect increase, the asymptotic biases increase. When the censoring process and event process are associated the bias is slightly smaller when $\exp(\alpha) > 1.2$; this arises because the effective censoring rate is higher and less information (to be used inappropriately) is provided by each individual. The two panels of Figure 9 show the asymptotic variance of es-

timators under the Andersen-Gill and Cox analyses as a function of $\exp(\alpha)$ in the same setting but for $\beta = \log 0.75$ and the early withdrawal rate is 20%. We see that when α increases and the expected number of events decreases, the variance of estimates increase. When the censoring process depends on the cumulative number of events ($g = 1.2$), the asymptotic variance of estimators is getting smaller. Also the asymptotic variance of estimators under Andersen-Gill model is smaller than those under Cox model when $\exp(\alpha) \leq 1.4$.

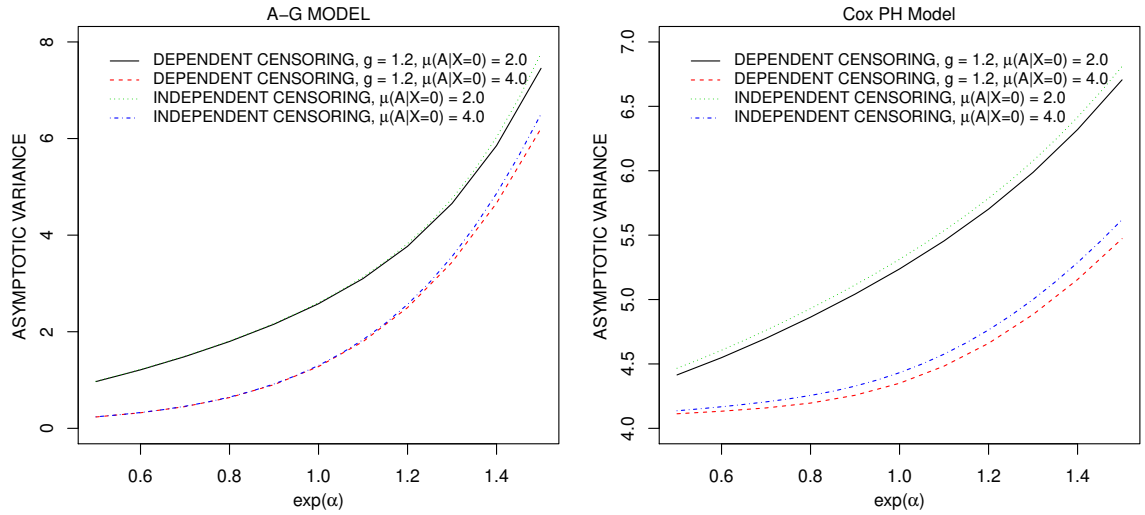

Figure 9: Asymptotic variance of estimators under Andersen-Gill model and Cox model as a function of $\exp(\alpha)$ when the event process is a Markov process

In the Supplementary Material, we report additional results for estimators from the marginal rate-based analysis when the true data generating process is Markov and censoring is state-dependent. There we find that the limiting bias is larger when the expected number of events is greater; the asymptotic bias is also sensitive to size of the treatment effect and relatively insensitive to event-dependent censoring for the modest range of dependence values (g) we consider for the censoring intensity; see Figure S.1. The impact of asymptotic variance of these estimators and the power of tests are also considered with the curves given in Figure S.2 and Figure S.3, respectively.

5 Illustrative Applications

5.1 Analysis of Recurrent Exacerbations in Cystic Fibrosis

Cystic fibrosis is a respiratory disease with airway obstruction caused by the accumulation of mucus in the lungs due to extracellular DNA; this results in recurrent pulmonary exacerbations. When delivered to the lungs in an aerosolized form, a highly purified recombinant form of DNase I called rhDNase cuts extracellular DNA, reducing the viscoelasticity of airway secretions and improving clearance. In a randomized double-blind trial reported by Fuchs et al. (1994), 321 individuals were assigned to receive rhDNase and 324 were assigned to receive a placebo treatment. The primary purpose of this study was to investigate the effect of rhDNase on the suppression of exacerbations so to this end the onset times of exacerbations were recorded over the study period of approximately 169 days. In the control arm 139 individuals had at least one exacerbation, 42 had at least two exacerbations, and 18 had at least three exacerbations; in the rhDNase arm these numbers were 104, 39 and 9, respectively. This study was reported on in Therneau and Hamilton (1997) and the data are available at the website for Cook and Lawless (2007).

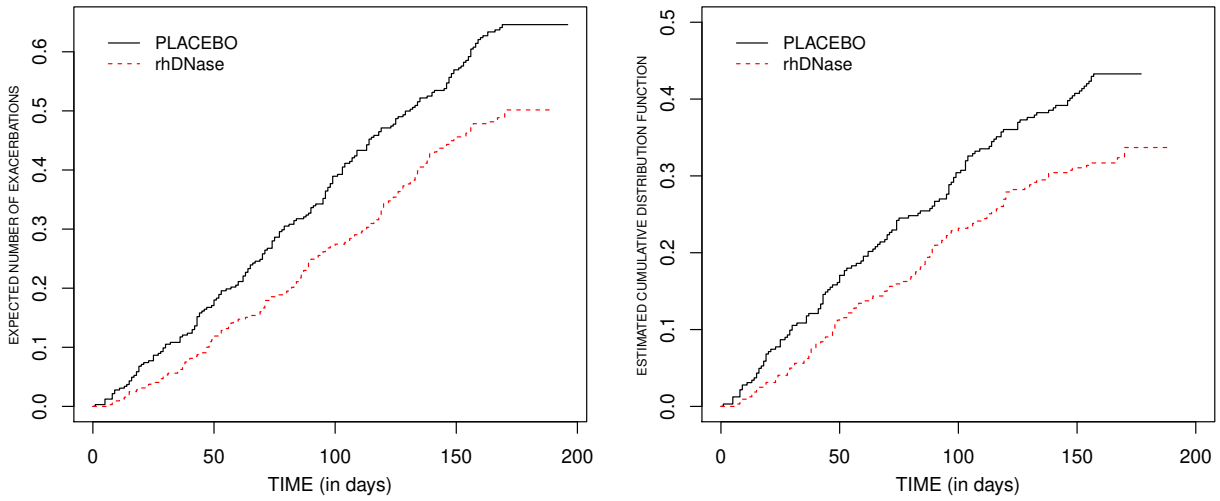

Figure 10: Nelson-Aalen estimates of the mean function for recurrent events and Kaplan-Meier estimates of cumulative probability function for the time to first exacerbations in the rhDNase trial by Fuchs et al. (1994).

Table 1: Estimates of treatment effects using the Andersen-Gill, partially conditional (stratified Andersen-Gill), and semiparametric negative binomial model for the recurrent events along with a Cox model for the time-to-first-event for the cystic fibrosis study reported by Fuchs et al. (1994).

	EST	RR	S.E.	p-value	ϕ
	rhDNase in Cystic Fibrosis				
Andersen-Gill Model	-0.271	0.763	0.124	0.029	-
Partially Conditional Rate Model	-0.234	0.791	0.108	0.030	-
Semiparamertric Negative Binomial	-0.271	0.763	0.125	0.030	0.67
Cox Model	-0.365	0.694	0.130	0.005	-

Figure 10 (a) gives plots of the Nelson-Aalen estimates of the mean functions for the rhDNase and control arms while Figure 10 (b) gives plots of the Kaplan-Meier estimates of the cumulative probability function for the time to first exacerbation. Table 1 gives the summary statistics from fitting the Andersen-Gill model, the stratified Andersen-Gill model, a semiparametric mixed Poisson model with a gamma frailty, and a Cox regression model for the time to the first exacerbation; the variance ϕ of the random effect is estimated as $\widehat{\phi} = 0.67$. The point estimates from the Andersen-Gill and semiparametric negative binomial analyses are identical, and the corresponding standard errors are very close, differing only in the third decimal place. When fitting stratified Andersen-Gill model, $k+1$ time-dependent strata were defined based on no events ($N_i(t^-) = 0$), 1 event ($N_i(t^-) = 1$), 2 events ($N_i(t^-) = 2$), upto $k-1$ events ($N_i(t^-) = k-1$), and then $\geq k$ events ($N_i(t^-) \geq k$). Since few people experienced more than 3 events, we set $k = 3$ and consider 4 time-dependent strata. The stratified Andersen-Gill model yields smaller evidence of a treatment benefit with a relative rate of 0.791 compared to 0.763 obtained from the Andersen-Gill and negative binomial analyses, and the standard error is also smaller since conditioning on the cumulative number of events explains some of the variation.

This analysis is not recommended however, with the limitations discussed in Zhong and Cook (2019). The Cox model yields a point estimate yielding a hazard ratio of 0.694 and with an only slightly larger standard error of 0.130 it gives a $p-$value for the test of no treatment effect at 0.005 which is smaller than those of the recurrent event analyses.

5.2 A Carcinogenicity Experiment Involving Mammary Tumours in Rats

Here, we consider data from a carcinogenicity experiment on the times to the development of mammary tumours in 48 female rats (Gail et al., 1980). All rats were exposed to a carcinogen following which they were assigned to one of two groups; 23 and 25 rats were randomized to the treatment and control groups, respectively. Rats were examined every 2 to 3 days during the 122 day follow-up period, and the days on which new tumours were discovered were recorded. The main objective of this study is to compare the onset rate of tumours in rats assigned to the treatment and control groups to assess the prophylatic nature of the treatment under study. In the control group all rats developed at least one tumour, 21 developed at least two tumours, and 9 rats developed more than 7 tumours; in the treatment arm 21 rats had at least one tumour, 14 rats had at least two tumours, and no rat developed more than 7 tumours.

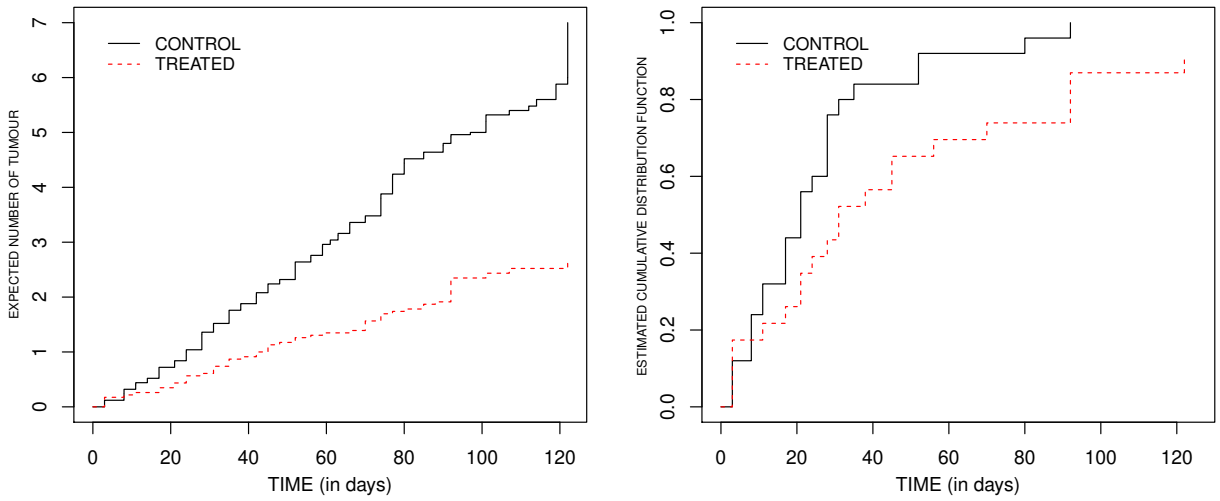

Figure 11: Nelson-Aalen estimates of the mean function for recurrent tumours and Kaplan-Meier estimates of cumulative probability function for the time to first tumour development using data from carcinogenicity study of Gail et al. (1980).

Figure 11 give plots of the Nelson-Aalen estimates of the mean functions and Kaplan-Meier estimates of the cumulative probability function for the time that the first tumour for the treatment and control rats, respectively. Table 2 gives the summary statistics from fitting the Andersen-Gill model, the stratified Andersen-Gill model with 8 strata, a semiparametric mixed Poisson model with a gamma frailty, and a Cox regression model for the time to the first tumour development; the variance ϕ of the random effect is estimated to be much lower here at $\widehat{\phi} = 0.27$. The point estimates from the Andersen-Gill and semiparametric negative binomial analyses are again identical at -0.816 giving a relative rate of 0.442, with the corresponding standard error slightly smaller for the Andersen-Gill analysis. The stratified Andersen-Gill model again yielded a more conservative estimate of the treatment effect with a point estimate of -0.535; the smaller standard error was again evident but we reiterate that this analysis is not recommended. The Cox model gives an intermediate estimate of treatment benefit with the point estimate of the hazard ratio given as 0.503.

Table 2: Estimates of treatment effects using the Andersen-Gill, partially conditional (stratified Andersen-Gill), and semiparametric negative binomial model for the recurrent events along with a Cox model for the time-to-first-event for carcinogenicity study of Gail et al. (1980).

	EST	RR	S.E.	p-value	ϕ
	Tumour onset in rats (Gail et al., 1980)				
Andersen-Gill Model	-0.816	0.442	0.198	<0.001	-
Partially Conditional Rate Model	-0.535	0.586	0.133	<0.001	-
Semiparametric Negative Binomial	-0.816	0.442	0.211	<0.001	0.27
Cox Model	-0.686	0.503	0.312	0.028	-

6 Remarks and Discussion

6.1 Summary Remarks

Here we make some summary remarks on the findings of the asymptotic findings and simulation studies we report on here, and provide some additional guiding comments.

1. If data are generated according to a Poisson process then the Andersen-Gill model will yield a more efficient estimator of the multiplicative treatment effect. Section 3.2.1, Figure 2

2. Within the class of data generating mixed Poisson models, estimands corresponding to a Cox model and Andersen-Gill model are in general incompatible with the differences influenced by the extent of heterogeneity and the expected number of events. Section 3.2.2, Figure 3

3. If the mean function specification is valid, robust variance estimates are needed for valid inference with the Andersen-Gill model. Section 3.2.2, Figure 4

4. Robust methods based on rate functions are robust in the sense that only the functional form of the marginal mean function must be correct for valid inference. With random censoring however, consistent estimators are only obtained if censoring is independent given the covariates controlled for in the rate function model. Section 3.3

5. If there is an event-dependence in the rate function as characterized by the Markov model of Section 3.3.1, the estimand from a multiplicative rate function model can be conservative or anticonservative depending on the expected number of events, and whether the event rate increases or decreases with event occurrence. Section 3.3.1, Figure 5

6. Within the class of mixed Poisson models the Andersen-Gill model will tend to lead to more powerful tests of treatment effects compared to a Cox model; more generally this may not be true if the treatment simply affects the risk of the first event. Figure 4

7. If the censoring intensity depends on the treatment arm and there is an association between the censoring and the recurrent event processes, biased estimates will be obtained from the Andersen-Gill analysis. Section 4

8. Inverse probability of censoring weights can be used to mitigate the bias arising from some forms of dependent censoring. (Cook et al., 2009)

9. Issues involving recurrent events and dependent terminal events are much more challenging to deal with and causal inference may best be carried out using utility-based analyses rather than hazard or intensity-based analyses.

6.2 Discussion

Discussions in the pharmaceutical industry and among regulatory agencies continue about the merit and limitations of recurrent event analyses. Central to these discussions are the interpretation of estimands in the various approaches to recurrent event analysis, the robustness of inferences to different kinds of complications that may arise, and the power of competing analyses. Semiparametric analyses based on partially specified models are appealing in their robustness to misspecification of the data generating process for the recurrent events, but they are at risk of biases from event-dependent withdrawal or censoring. Likelihood or partial likelihood analysis based on mixed Poisson models are insensitive to event-dependent censoring but these do require full models assumptions; the inferences regarding multiplicative treatment effects appear relatively robust to misspecification of the mixing distribution so these remain a viable approach to analysis. In fact mixed Poisson models offer a natural basis for the design of clinical trials based on recurrent events.

We have highlighted the fact that the rate-based recurrent event modeling framework is generally incompatible with the Cox regression model routinely applied in analyses of the time to the first event. When the recurrent events are generated by a Poisson process then multiplicative rate based analyses and the Cox model yield estimates which are consistent for the same estimand and the former will be more efficient in general. However, there is typically extra-Poisson variation which will mean the corresponding estimands differ and the interpretations of findings are therefore different. In terms of power, no single method consistently dominates and our illustrative analyses show that Cox regression may lead to smaller or larger $p-$values than recurrent event analyses in any given setting.

A variety of methods can be used to assess the appropriateness of the multiplicative rate or hazard function assumptions in recurrent event and time to first event analyses. Nelson-Aalen estimates (Cook and Lawless, 2007) of the cumulative baseline mean or hazard functions can be plotted, which should be roughly proportional to one another across the two treatment groups. To check the functional form of treatment effects Schoenfeld residual (Therneau and Grambsch, 2000) plots can also be examined; while we have not emphasized general regression modeling here these plots are perhaps most useful for assessing effects of continuous covariates. Finally model expansion is sufficiently straightforward with modern software that fitting expanded models and carrying out tests of the need for such model expansion is straightforward. Diagnostics regarding the assumption of independent censoring are less developed, but by formulating multistate models such the one depicted in Figure 1 (b), plots of Nelson-Aalen estimates of the cumulative $j \rightarrow C$ transition rates can be inspected (Cook et al., 2009); these estimates should have comparable slopes if censoring is truly independent of event occurrence. If there is evidence to suggest dependent censoring joint models for the event and censoring processes via frailty (Cook et al., 2010; Wang et al., 2001), or inverse probability of censoring weights (Cook et al., 2009; Miloslavsky et al., 2004) can be adopted. Robust methods based on scale-change models and using artificial censoring to remove bias have also been developed (Ghosh and Lin, 2003; Hsieh et al., 2011).

A quite different and considerably great challenge arises when a recurrent event process is terminated by another event. Examples are ubiquitous and include the analysis of recurrent graft rejection episodes in transplantation studies where episodes are terminated by total graft rejection, recurrent metastases in patients with advanced cancer where the metastatic process is terminated by death, and recurrent exacerbations in chronic obstructive pulmonary disease

in elderly patients at risk of terminating event of death. Important work on semiparametric analysis of data from such processes has taken place over the past 20 years including Ghosh and Lin (2000) and Ding et al. (2009) among others. Causal inference in this setting is particularly challenging but building on the increasing popularity of tests based on restricted mean lifetimes, utility-based analyses offer a promising avenue for development in this setting. These challenges are beyond the scope of the present work, where we focus on the performance of common estimators of treatment effects in randomized trials.

Much of the work of the Estimands Working Group in the survival setting has been directed at the development of methodology for time to event analyses when the proportional hazards assumption is violated (Akacha et al., 2017; Rufibach, 2019) where weighted log-rank tests (León et al., 2020) have received considerable attention for use on their own or as part of supremum-based tests. Estimates with a simple causal interpretation are of course a priority following rejection of a null hypothesis of no difference; see Fay et al. (2018). Two degree of freedom tests accommodating more general departures from the null hypothesis of common hazards have been developed for the failure time setting and analogous generalizations are possible for recurrent events (Cook et al., 1996).

Acknowledgements

This research was supported by grants from National Natural Science Foundation of China (NSFC-11901376), Shanghai Pujiang Program (2019PJC051), and SUFE Innovation Funding (2019110051) to Y Zhong, a Discovery Grant and Supplement Award from the Natural Science and Engineering Research Council of Canada to RJ Cook (RGPIN 155849 and RGPIN 04207) along with a grant from the Canadian Institutes for Health Research (FRN 13887). RJ Cook is a Faculty of Mathematics Research Chair, University ofWaterloo.

Conflict of Interest

The authors declare no potential conflict of interests.

Data Availability Statement

The R code that support the findings of this study are available upon request from the corresponding author.

References

Aalen, O., Borgan, O., and Gjessing, H. (2008). *Survival and event history analysis: a process point of view*. Springer Science & Business Media.

Agency, E. M. (2019). ICH Guidelines. https://www.ema.europa.eu/en/human-regulatory/research-development/scientific-guidelines/ich-guidelines. Accessed: 2019-08-26.

Akacha, M. (2019). Choosing measures of treatment benefit: estimands and beyond. *CHANCE*, 32(4):12–17.

Akacha, M., Bretz, F., and Ruberg, S. (2017). Estimands in clinical trials–broadening the perspective. *Statistics in Medicine*, 36(1):5–19.

Akacha, M. and Ogundimu, E. O. (2016). Sensitivity analyses for partially observed recurrent event data. *Pharmaceutical Statistics*, 15(1):4–14.

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

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

Cazzola, M., Anapurapu, S., and Page, C. P. (2012). Polyvalent mechanical bacterial lysate for the prevention of recurrent respiratory infections: a meta-analysis. *Pulmonary Pharmacology and Therapeutics*, 25(1):62–68.

Cook, R. J. (1995). The design and analysis of randomized trials with recurrent events. *Statistics in Medicine*, 14(19):2081–2098.

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

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

Cook, R. J., Lawless, J. F., and Lee, K.-A. (2010). A copula-based mixed Poisson model for bivariate recurrent events under event-dependent censoring. *Statistics in Medicine*, 29(6):694–707.

Cook, R. J., Lawless, J. F., and Nadeau, C. (1996). Robust tests for treatment comparisons based on recurrent event responses. *Biometrics*, 52(2):557–571.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)*, 34:187–220.

Cox, D. R. and Miller, H. D. (1965). *The Theory of Stochastic Processes*. Methuen & Co, London.

Ding, A.-A., Shi, G., Wang, W., and Hsieh, J.-J. (2009). Marginal regression analysis for semi-competing risks data under dependent censoring. *Scandinavian Journal of Statistics*, 36(3):481–500.

Ettinger, B., Black, D. M., Mitlak, B. H., Knickerbocker, R. K., Nickelsen, T., Genant, H. K., Christiansen, C., Delmas, P. D., Zanchetta, J. R., Stakkestad, J., et al. (1999). Reduction of vertebral fracture risk in postmenopausal women with osteoporosis treated with raloxifene: results from a 3-year randomized clinical trial. *JAMA*, 282(7):637–645.

Fay, M. P., Brittain, E. H., Shih, J. H., Follmann, D. A., and Gabriel, E. E. (2018). Causal estimands and confidence intervals associated with wilcoxon-mann-whitney tests in randomized experiments. *Statistics in Medicine*, 37(20):2923–2937.

Fuchs, H. J., Borowitz, D., Christiansen, D., Morris, E., Nash, M., Ramsey, B., Rosenstein, B. J., Smith, A. L., and Wohl, M. E. (1994). The effect of aerosilozed recombinant human dnase on respiratory exacerbations and pulmonary function in patients with cystic fibrosis. *New England Journal of Medicine*, 331:637–642.

Gail, M. H., Santner, T. J., and Brown, C. C. (1980). An analysis of comparative carcinogenesis experiments based on multiple times to tumor. *Biometrics*, 36(2):255–266.

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

Ghosh, D. and Lin, D.-Y. (2003). Semiparametric analysis of recurrent events data in the presence of dependent censoring. *Biometrics*, 59(4):877–885.

Hsieh, J.-J., Ding, A.-A., and Wang, W. (2011). Regression analysis for recurrent events data under dependent censoring. *Biometrics*, 67(3):719–729.

Kelly, P. J. and Lim, L. (2000). Survival analysis for recurrent event data: an application to childhood infectious diseases. *Statistics in Medicine*, 19(1):13–33.

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

Lawless, J. F. (1987b). Regression methods for Poisson process data. *Journal of American Statistical Association*, 82(399):808–815.

Lawless, J. F. and Cook, R. J. (2019). A new perspective on loss to follow-up in failure time and life history studies. *Statistics in Medicine*, 38(23):4583–4610.

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

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

León, L. F., Lin, R., and Anderson, K. M. (2020). On weighted log-rank combination tests and companion cox model estimators. *Statistics in Biosciences*, 12(2):225–245.

Lin, D. Y. and Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, 84:1074–1078.

Lin, D. Y., Wei, L. J., 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 (Statistical Methodology)*, 62(4):711–730.

Miloslavsky, M., Keleş, S., van der Laan, M.-J., and Butler, S. (2004). 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.

Nakamura, K., Inokuchi, R., Daidoji, H., Naraba, H., Sonoo, T., Hashimoto, H., Tokunaga, K., Hiruma, T., Doi, K., and Morimura, N. (2017). Efficacy of levetiracetam versus fosphenytoin for the recurrence of seizures after status epilepticus. *Medicine*, 96(25).

Nelson, R. (2006). *An Introduction to Copulas*. Springer-Verlag, New York.

Prentice, R. L., Williams, B. J., and Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, 68:373–379.

Roger, J. H., Bratton, D. J., Mayer, B., Abellan, J. J., and Keene, O. N. (2019). Treatment policy estimands for recurrent event data using data collected after cessation of randomised treatment. *Pharmaceutical Statistics*, 18(1):85–95.

Rufibach, K. (2019). Treatment effect quantification for time-to-event endpoints – Estimands, analysis strategies, and beyond. *Pharmaceutical Statistics*, 18(2):145–165.

Stark, M. E. (2018). *Comparison of several analysis methods for recurrent event data for different estimands*. PhD thesis, Georg-August-University Göttingen.

Strawderman, R. L. (2000). Estimating the mean of an increasing stochastic process at a censored stopping time. *Journal of the American Statistical Association*, 95(452):1192–1208.

Strengell, T., Uhari, M., Tarkka, R., Uusimaa, J., Alen, R., Lautala, P., and Rantala, H. (2009). Antipyretic agents for preventing recurrences of febrile seizures: randomized controlled trial. *Archives of Pediatrics & Adolescent Medicine*, 163(9):799–804.

Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazards models. *Biometrika*, 74(2):363–369.

Therneau, T. M. and Grambsch, P. M. (2000). *Modelling Survival Data: Extending the Cox Model*. Springer, New York.

Therneau, T. M. and Hamilton, S. A. (1997). rhDNase as an example of recurrent event analysis. *Statistics in Medicine*, 16(18):2029–2047.

Wang, M.-C., Qin, J., and Chiang, C.-T. (2001). Analyzing recurrent event data with informative censoring. *Journal of the American Statistical Association*, 96(455):1057–1065.

Wei, L. J., Lin, D. Y., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84(408):1065–1073.

Wu, L. and Cook, R. J. (2012). Misspecification of Cox regression models with composite endpoints. *Statistics in Medicine*, 31(28):3545–3562.

Zhong, Y. and Cook, R. J. (2019). The effect of omitted covariates in marginal and partially conditional recurrent event analyses. *Lifetime Data Analysis*, 25(2):280–300.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of this article.

Published in final edited form as:
Statistics in Medicine (2021), **40**(16): 3823–3842
DOI: 10.1002/sim.9002

Supplementary material for *Semiparametric recurrent event vs time-to-first-event analyses in randomized trials: Estimands and model misspecification*

YUJIE ZHONG

School of Statistics and Management,
Shanghai University of Finance and Economics, Shanghai, P.R. China
E-mail: zhong.yujie@mail.shufe.edu.cn

RICHARD J. COOK

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

Summary

Here we report on the results of additional calculations related to the limiting value of estimators from rate based analysis when the true data generating process is a Markov process. We consider cases with state-dependent censoring.

1 Markov models and Event-Dependent Censoring

1.1 Limiting values and asymptotic variances of rate-based and Cox regression

Here we consider the limiting bias of the Andersen-Gill model as a function of g where if $g \neq 1$, the censoring process depends on the event process. We consider different expected number of events in the control arm, different treatment effects, and different rates of early withdrawal. From the plots in Figure S.1 we see that the asymptotic biases are smaller for larger values of g, and that the asymptotic biases are larger when the expected number of events and treatment effects are larger. Figure S.2 illustrate the asymptotic variance of estimators under Andersen-Gill model and Cox model as a function of g when the event is Markov process with $\alpha = \log 1.2$, $\beta = \log 0.75$ and the early withdrawal rate is 20%. We see here that when the expected number of events increase or when g increases, the asymptotic variance decreases for both methods. As before the variance of the respective estimator is larger under the Cox model compared to Andersen-Gill model.

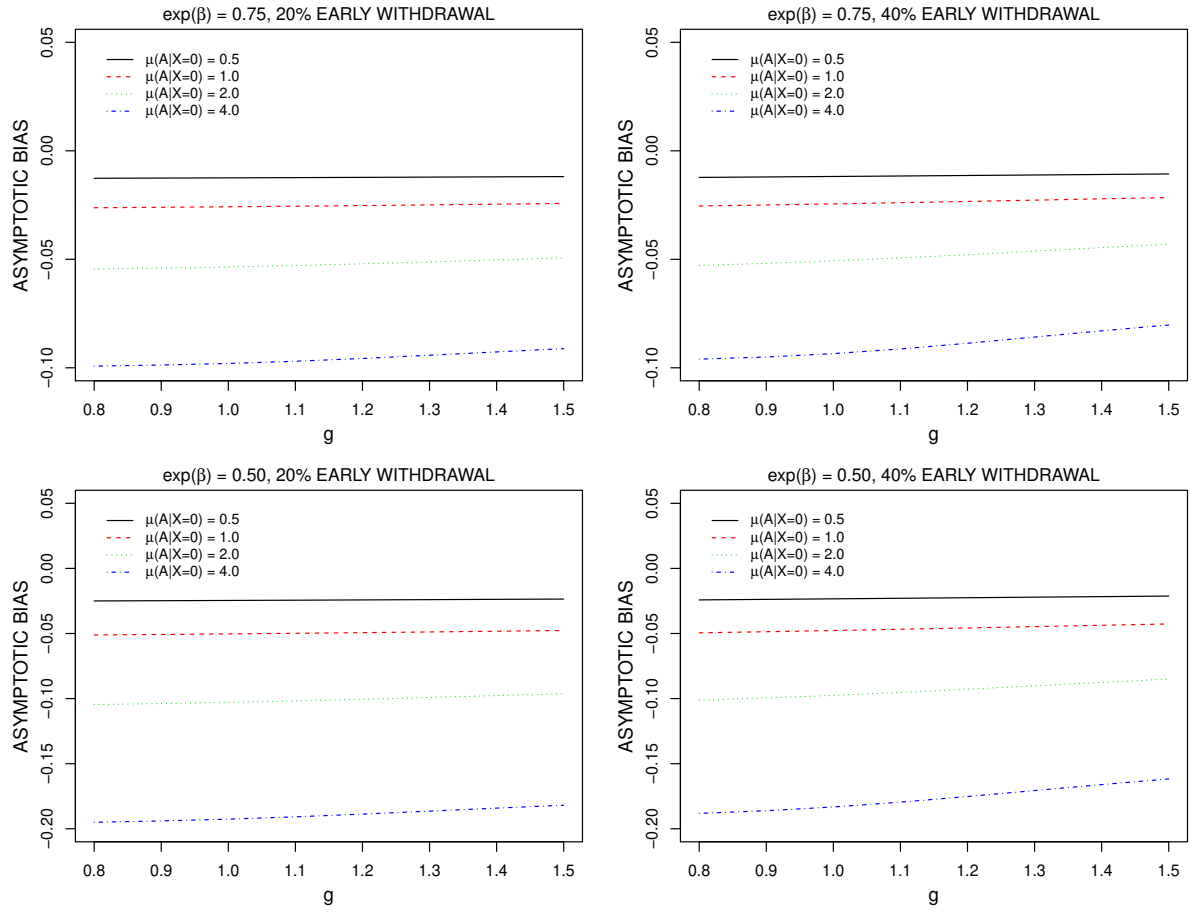

Figure S.1: Limiting bias of estimators under Andersen-Gill model as a function of g when the event process is a Markov process ($\alpha = \log 1.2$) and the intensity function for censoring process depends on the cumulative number of events

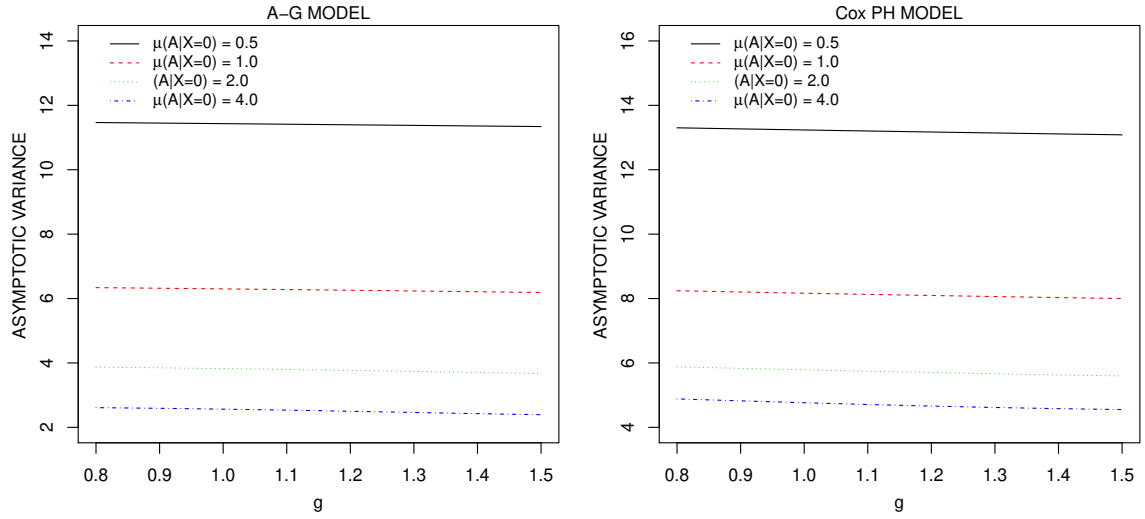

Figure S.2: Asymptotic variance of estimators under Andersen-Gill model and Cox model as a function of g when the event process is a Markov process ($\alpha = \log 1.2$) and the intensity function for censoring process depends on the cumulative number of events

1.2 Power implications

The limiting value and the asymptotic variances play a role in Wald tests of the hypotheses $H_0 : \beta = 0$ versus $H_A : \beta \neq 0$. We therefore now consider the effect of event-dependent censoring on the power of tests of treatment effects under Markov models for $g = 1.0$ and 1.5. The sample size is determined to give 80% power to detect treatment effect under Andersen-Gill model or Cox model when the significance level is 0.05 and $\beta_A = \log 0.75$ when the censoring process is taken to be independent of the event process (i.e. $g = 1.0$). For this sample size calculation we first obtain the limiting value $\beta^\dagger$ when $\beta = \log 0.75$ under the Markov models with independent censoring and then obtain the corresponding asymptotic variance. Figure S.3 shows the power of no treatment effect test under Andersen-Gill and Cox models when the event process is Markov process and the censoring process possibly depends on the cumulative number of events (i.e. the censoring intensity is state-dependent). The top two plots are based on the sample size derived from the Andersen-Gill model when the data generation is based on a Markov process with independent censoring while the bottom two plots are based on the sample size obtained under a Cox model for the time to the first event under the Markov process with independent censoring. From this plot we see that within the framework of the Cox model the power is unaffected by the relation between the censoring and event processes; this is natural since state-dependent censoring is not manifest until the first event is observed and this is the failure time in this analysis. Although Cox analyses give consistent estimate in this setting, the power is smaller for Cox analysis compared with the analysis based on the Andersen-Gill model.

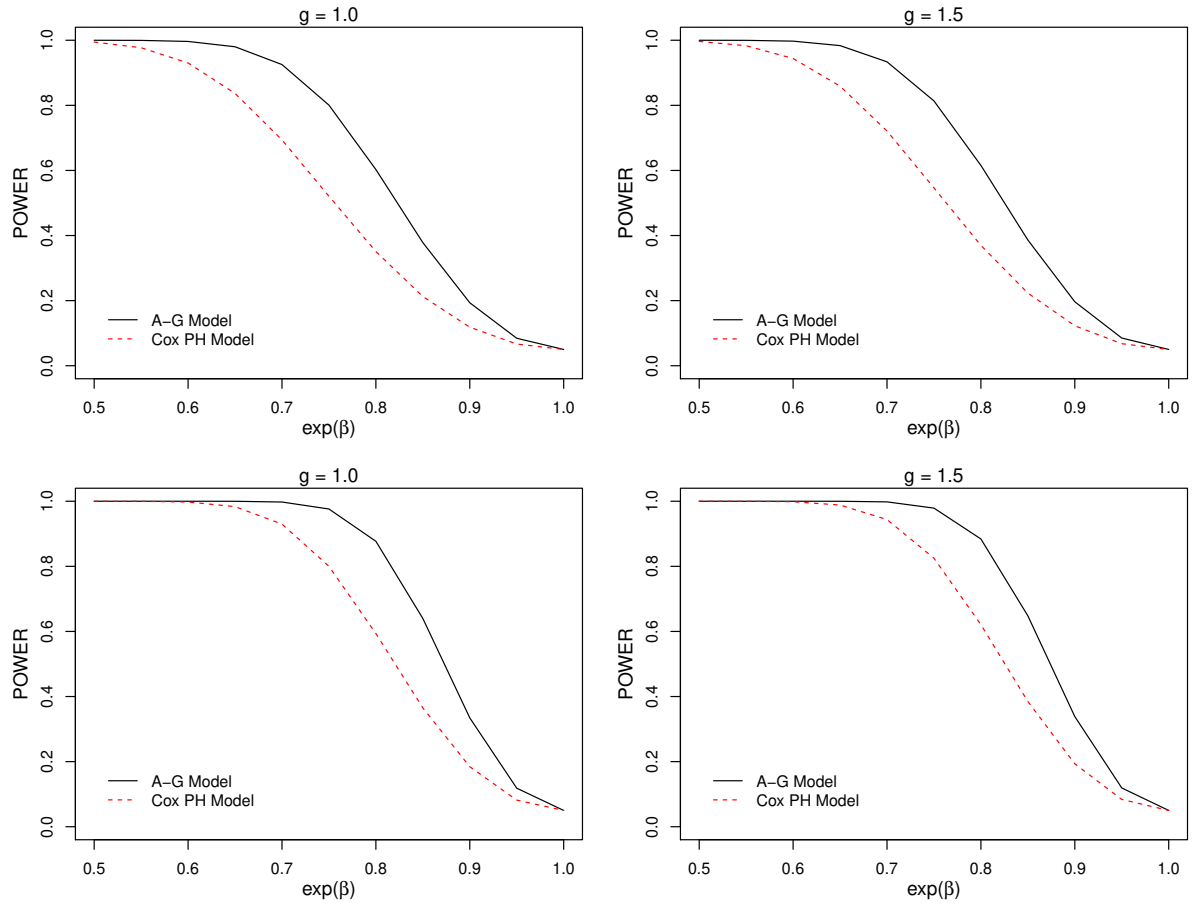

Figure S.3: Power of test under Andersen-Gill model and Cox model as a function of $\exp(\beta)$ when the event process is a Markov process ($\alpha = \log 1.5$) and the intensity function for censoring process depends on the cumulative number of events (top two plots are based on the sample size obtained from Andersen-Gill model for Markov process with independent censoring and bottom two plots are based on the sample size derived from Cox model for Markov process with independent censoring)