Political Analysis Advance Access published September 24, 2013

Political Analysis (2013) pp. 1–22
doi:10.1093/pan/mpt015

Analyzing the Robustness of Semi-Parametric Duration Models for the Study of Repeated Events

Janet M. Box-Steffensmeier
Department of Political Science, Ohio State University, 2140 Derby Hall, 154 N. Oval Mall Columbus, OH 43210

Suzanna Linn
Department of Political Science, Penn State University, 320 Pond Lab, University Park, PA 16802
e-mail: slinn@la.psu.edu (corresponding author)

Corwin D. Smidt
Department of Political Science, Michigan State University, South Kedzie Hall, 368 Farm Lane, S303, East Lansing, MI 48824

Edited by R. Michael Alvarez

Estimators within the Cox family are often used to estimate models for repeated events. Yet, there is much we still do not know about the performance of these estimators. In particular, we do not know how they perform given time dependence, different censoring rates, and a varying number of events and sample sizes. We use Monte Carlo simulations to demonstrate the performance of a variety of popular semi-parametric estimators as these data aspects change and under conditions of event dependence and heterogeneity, both, or neither. We conclude that the conditional frailty model outperforms other standard estimators under a wide array of data-generating processes, and data limitations rarely alter its performance.

1 Introduction

Political scientists often care about the answers to questions involving the effects of variables on the timing and occurrence of outcomes that may recur. Examples of repeated events of interest in political science include individual changes in partisanship (Kuhn 2009) or decisions to turn out to vote (Leighley and Nagler 2013), international acts of terrorism (Dugan, LaFree, and Piquero 2005), the formation of trade agreements (Baccini 2012), instances of pension retrenchment (Fernandez 2010), government elections (Brancati and Snyder 2011) or collapse (Curini 2011), presidential vetoes (Rohde and Simon 1985), policy diffusion (Boehmke and Witmer 2004; Boehmke and Skinner 2012), outbreak of civil wars (Schneider and Wiesehomeier 2008), and mediation success (Greig 2001; Beardsley 2008). These types of processes are not amenable to analysis with generalized linear models because of the prevalence of multiple types of dependencies in the data.

Consider that the simple passage of time may affect the likelihood an event occurs, producing event rates that depend on time. Governments may have natural life cycles, for example, or one's party identification may solidify with age (Plutzer 2002). Additionally, event rates may depend on the existence and number of previous events, a phenomenon referred to as event or occurrence dependence. This type of dependence may arise in the case of terrorist acts if the perceived likelihood of success increases terrorist motivation to instigate more airline hijackings, for example, in the wake of recent hijackings, as suggested by Dugan, LaFree, and Piquero (2005). Event rates may also depend on unaccounted-for differences across cases that make some cases more prone to events than others. The existence of bad leaders can lead to civil wars (Brown 1996) and personality

Authors' note: Thanks to Neal Beck and anonymous reviewers for helpful comments on drafts of the article.

conflicts can lead to cabinet breakdowns (Martin and Vanberg 2003; Boehmke, Morey, and Shannon 2006), for example, but these factors are not readily identified or measured. Different event rates across cases due to either event dependence or unobserved heterogeneity produce *within*-subject correlation.

Researchers have different ways to account for these forms of within-subject correlation when estimating duration models, and these approaches are largely similar and available across semi-parametric, parametric, and discrete time duration models. But we currently have little knowledge of how each approach performs in the presence of model misspecification or data constraints. In response, our study seeks to evaluate the robustness of repeated events models across various data conditions by examining the performance of various Cox models for repeated events.

The Cox semi-parametric duration model has been widely used to test the effects of variables on repeated events data, in particular because it makes no assumptions about the form of time dependence that may exist in the data. Many forms of the Cox model have also been proposed to deal with different forms of within-subject correlation in the context of repeated events processes. Various authors have compared the performance of these models under some specific conditions (Wei, Lin, and Weissfeld 1989; Stukel 1993; Pepe and Cai 1993; Cook, Lawless, and Nadeau 1996; Cook and Lawless 1997; Li and Lagakos 1997; Therneau and Hamilton 1997; Kelly and Lim 2000; Therneau and Grambsch 2000; Box-Steffensmeier and De Boef 2006; Metcalfe and Thompson 2006; Cheung et al. 2010). But no one has examined the joint effect of time dependence and within-subject correlation on the estimates in the family of Cox models. We extend the work of Box-Steffensmeier and De Boef (2006), examining the effect of both types of within-subject correlation—individually and jointly—in the context of three forms of Weibull distributed time-dependent repeated events processes on five models in the Cox family of duration models. Specifically, we consider Andersen and Gill's (1982) model, which is the Cox model with robust standard errors (SEs), the conditional-gap and elapsed-time (Prentice, Williams, and Peterson 1981) models, the frailty model in elapsed time (Oakes 1992), and the conditional frailty model in gap time (Box-Steffensmeier and De Boef 2006). Each of these models has been suggested as promising and/or is heavily used in the literature as a way to model repeated events.

We focus on the role of time dependence because it is a main challenge in modeling repeated events. Such data require making more assumptions about the nature of the data-generating process (DGP), but in many cases modelers make such choices under considerable uncertainty. Researchers often have no strong theoretical basis for specifying whether hazard rate processes are predominately a function of time since last event (gap time) or time since being at risk (elapsed time). Likewise, a researcher's ability to decipher whether processes exhibit significant unobserved heterogeneity or event dependence are often dependent on these assumptions, and these choices may mask misspecification of the baseline hazard process and, consequently, estimates of treatment effects.

For instance, we sampled a distribution of elapsed times drawn from a Weibull hazard distribution with a shape parameter of 1.2, where half the sample is given treatment ($\beta = -1.0$). As shown in Fig. 1, a researcher can easily specify this process as a gap-time process without indication of misspecification. Since the baseline hazard rate changes over time in a nonconstant manner, a gap-time specification of the data suggests the need for a stratified estimate of the baseline hazard (left panel). The misspecified gap-time model estimate provides a substantial underestimate of treatment's true effect ($\hat{\beta} = -0.867$, with a SE of 0.070). However, an examination of a diagnostic Cox-Snell residual plot (right panel) shows they are distributed as unit exponential, providing no indication of the incorrect model specification.

In addition, we are motivated by a second set of concerns that face analysts: data constraints. Theory and data provide us some information about the nature of the DGP and inform our choice of the optimal estimator. But practical issues with the data at hand often raise additional concerns for inference. We may find ourselves with a small number of cases (or cases relative to events), a small number of events, or a high degree of censoring. In some cases, the solution—if possible—is to get more data. Absent that option, what are the consequences of these data constraints for assessing the effects of variables on repeated events processes? The existing literature offers little guidance (Therneau and Grambsch 2000). In the second part of the article, we examine the effects

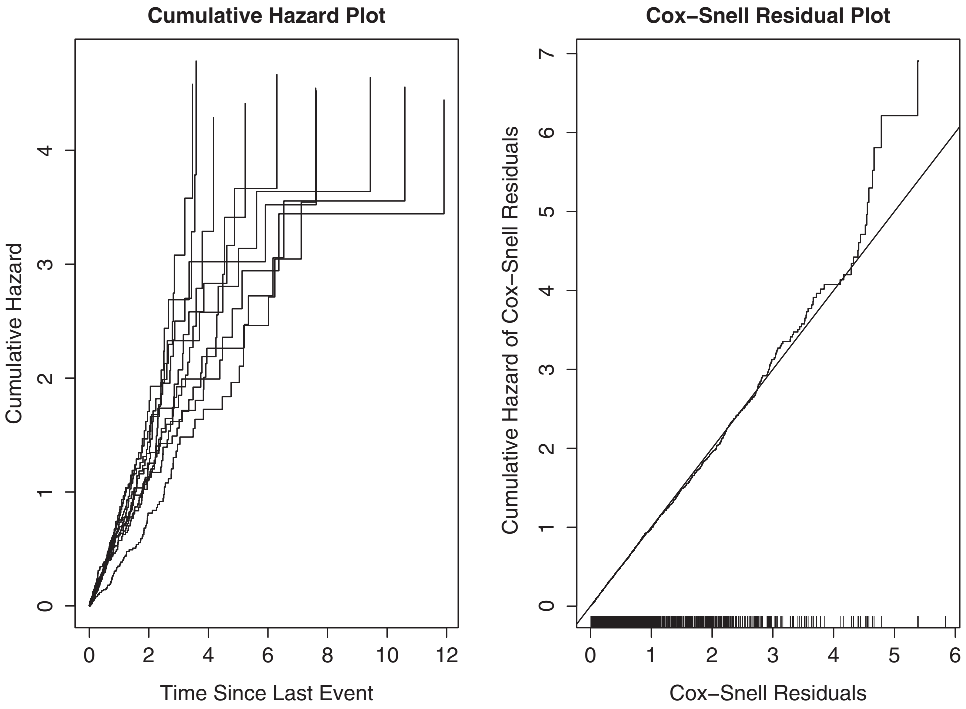

Fig. 1 Diagnostic plots from a misspecified gap-time model. Cumulative hazard and Cox-Snell residual plot from a single simulated data set ($N = 1000$), where the treatment effect, β, is set to -1.0 and the distribution of elapsed times is drawn from a Weibull hazard distribution with a shape parameter of 1.2.

of each of these data constraints on the same five models, again under both types of within-subject correlation—individually and jointly—in the context of exponentially distributed times to event.

Our goal is to assess the robustness of commonly used duration model specifications for repeated events data analysis in situations faced by political scientists. Our primary tool is Monte Carlo analysis. In the following sections, we provide an overview of the Cox semi-parametric model and the extensions we consider here. The simulation setup follows. We turn then to the results of the analysis of the interaction of dependencies in the data due to time dependence and within-subjection correlation. This is followed by analysis of the effects across different time scales and various data constraints. Next we offer some advice to data analysts and some general conclusions about the robustness of the models. We find the conditional frailty model to be the most robust to the widest range of DGPs and data constraints, making it the most attractive option for analysts.

2 The Cox Model and Its Extensions

The Cox model estimates the (instantaneous) ratio of the hazard rate in the *treated* compared with the *control* group. It assumes proportional hazards—that the hazard ratio is constant over time. The Cox proportional hazard model is the most widely used model for the study of survival data because it makes no assumptions about the distribution of time-to-event. The Cox model uses time to order the composition of risk sets and then averages the treatment's hazard ratio across each observed risk set. But it is well known that, given within-subject correlation, the Cox model is both biased and inefficient. Many extensions of the baseline Cox model have been proposed to remedy the problems. We compare the performance of the most widely used and promising extensions. They vary in terms of the time scale over which the hazard is calculated, the risk set or time intervals during which a subject is deemed at risk for a given event k, the choice of common or event-specific baseline hazards, and in terms of how they handle the unobserved heterogeneity in the data. We briefly describe these distinctions here.

1. Time scale. The time scale can be structured in gap time—where time resets with each event occurrence—or in elapsed (also referred to as total) time—where time is measured from the start of treatment.[1]
2. Definition of the risk set. For the unrestricted risk set, all of a subject's risk intervals may contribute to the risk set for any given event, regardless of the number of events experienced by each subject. For the restricted risk set, only those subjects experiencing $(k-1)$ events will be in the risk set (at risk) for the kth event.
3. Choice of a common versus event-specific baseline hazard. Baseline hazard rates may be event specific, in which case the hazard is stratified by event and the model allows the hazard to change with events, or it may be common to all events.
4. Handling of unobserved heterogeneity. Unobserved heterogeneity may be handled with variance-corrected SEs or the inclusion of a frailty term (also referred to as a random effect term).

The Andersen-Gill model is a straightforward extension of the basic Cox model. It is given in elapsed time with an unrestricted risk set where all subjects' risk intervals may contribute to the risk set for any given event, regardless of the number of events experienced by each subject (Andersen and Gill 1982). The model has a common baseline hazard and uses variance-corrected SEs to control for within-subject correlation. Let T_{ik} be true total time of the kth event for the ith subject, C_{ik} be the censoring time of the kth event for the ith subject, and X_{ik} the corresponding observation time, $X_{ik} = \min(T_{ik}, C_{ik})$. z_{ik} is a vector of covariates. Then the Andersen-Gill model has a hazard rate given by

$$\lambda_{ik}(t|z_{ik}) = \lambda_0(t)e^{\beta' z_{ik}}, \tag{1}$$

where cases are at risk whenever $X_{i,k-1} < t < X_{ik}$. So the observed time of the $(k-1)$th event is less than the kth event.

The conditional gap-time model is in gap time with a restricted risk set so that contributions to the kth risk set are restricted to only include the kth event risk intervals of those subjects who have experienced $(k-1)$ events (Prentice, Williams, and Peterson 1981). The model has event-specific baseline hazards and variance-corrected SEs to correct for within-subject correlation. The hazard rate is given by

$$\lambda_{ik}(t|z_{ik}) = \lambda_{0k}(t - t_{k-1})e^{\beta' z_{ik}}, \tag{2}$$

where X_{ik} is replaced by G_{ik}, where $G_{ik} = X_{ik} - X_{i,k-1}$ is the gap time with $X_{i0} = 0$. Cases are at risk when $0 < t - t_{k-1} < G_{ik}$ so contributions to the risk set are restricted to only include the kth event risk intervals for those subjects who have experienced $(k-1)$ events.

The conditional elapsed-time model is in total time with a restricted risk set. The model has event-specific baseline hazards and variance-corrected SEs to correct for within-subject correlation (Cook and Lawless 1997). The hazard is given by

$$\lambda_{ik}(t|z_{ik}) = \lambda_{0k}(t)e^{\beta' z_{ik}}. \tag{3}$$

Contributions to the risk set are restricted to only include the kth event risk intervals for those subjects who have experienced $(k-1)$ events.

The frailty model is in total time with an unrestricted risk set. It has a common baseline hazard and controls for within-subject correlation with a random effect or frailty term (Therneau and Grambsch 2000). The hazard is given by

$$\lambda_{ik}(t|z_{ik}) = \lambda_0(t)e^{\beta' z_{ik} + \omega_i} = \lambda_0(t)e^{\omega_i}e^{\beta' z_{ik}}, \tag{4}$$

[1] Both gap time and elapsed time can be generated from counting process notation.

where e^{ω_i} represents the unknown random effects or frailties associated with each subject i and are assumed to follow a specific distribution, generally Gamma or Gaussian. Cases are at risk when $X_{i,k-1} < t < X_{ik}$. So the observed time of the $(k-1)$th event is less than the kth event.

Finally, we estimate the conditional frailty model, which combines the gap-time formulation with a restricted risk set, and event-specific baseline hazards (Box-Steffensmeier and De Boef 2006). The model controls for within-subject correlation with a random effect. In this way, the conditional frailty model is the most general of the models we consider. The hazard is given by

$$\lambda_{ik}(t|z_{ik}) = \lambda_{0k}(t - t_{k-1})e^{\beta' z_{ik}+\omega_i}. \tag{5}$$

X_{ik} is again replaced by G_{ik}. So contributions to the risk set are restricted to only include the kth event risk intervals for those subjects who have experienced $(k-1)$ events.

3 Simulations: An Overview of the DGPs

Our goal is to assess the robustness of a family of Cox models under a variety of conditions given small samples and common data constraints. For this task, we rely primarily on simulations to empirically gauge model performance as we change the form of time dependence (distribution of time-to-event), the time scale (gap time or elapsed time), the degree of censoring, the number of events, and the sample size. For each case, we generate data that (1) contain event (occurrence) dependence such that the occurrence of an event makes its recurrence more or less likely, raising and lowering the baseline hazard, respectively, (2) contain unobserved heterogeneity such that some cases have higher or lower event rates, (3) contain both event dependence and heterogeneity, and finally (4) contain neither event dependence nor heterogeneity. We gauge model performance on three dimensions: the bias in the estimated treatment effects as well as in the estimated variance of the random effect, bias in the SEs, and rate at which the estimated confidence intervals include the true parameter. Our assessments of censoring focus on the mean estimates of β. Sample size experiments focus on the mean squared error of the estimates.

The basic DGP for the experiments draw duration times (G_{ik}) from a Weibull distribution and varying combinations of heterogeneity and event dependence, time scale, number of events, and censoring. The baseline hazard was initially set to 1.0 and departures are noted below. The constant treatment effect was set to $\beta = -1.0$, as treatments are typically designed to reduce event rates.[2] Maximum follow-up time was set so that there was no censoring, with the maximum number of events set to 10, unless otherwise noted. All models were estimated in R using the survival package.[3]

In our first set of experiments, we isolate the effect of the time dependence/distribution of time-to-event, essentially fixing censoring rates at zero by observing all cases until they reached the maximum number of events, $k = 10$.[4] For all our time-to-event (gap-time) data, we draw the time of an individual i's kth event—t_{ik}—from a Weibull distribution with hazard rate λ_{ik}, where

$$\lambda_{ik}(t|z_i) = p(t)^{p-1}\lambda_{0k}e^{\omega_i}e^{\beta' z_i}. \tag{6}$$

λ_{0k} is the baseline hazard rate and may vary by k. Event dependence enters in the form of the baseline hazard, λ_{0k}. If the baseline is constant across events, there is no event dependence, $\lambda_{0k} = \lambda_0$. When event dependence occurs, the baseline hazard is allowed to vary as some function of k. We follow Box-Steffensmeier and De Boef (2006) and specify $\lambda_{0k} = k\lambda_0$. This produces inter-event times that decrease as the number of events experienced grows, producing different and correlated event-specific baseline hazards. Time dependence is introduced through the Weibull shape parameter p. For values of $p < 1.0$, the hazards decrease with age, such as when chances of regime turnover fall the longer a regime lasts. For $p > 1$, failure is increasing as with age, such as when rates of policy change increase as existing policies become antiquated. When $p = 1$, the

[2] See Box-Steffensmeier and De Boef (2006) for the effect of a treatment designed to raise event rates.

[3] Replication materials for this article are available from the *Political Analysis* dataverse at http://hdl.handle. net/1902.1/21914.

[4] In some cases with unobserved heterogeneity, some observations were censored when cases drew failure times so large that the model could not be estimated.

Weibull reduces to the exponential distribution and the failure rate is constant over time and there is no duration dependence. We considered $p = 0.8$, $p = 1.2$, and $p = 1.0$. Heterogeneity is introduced in the random effect, e^{ω_i}. It is set equal to 1 for the case of no unobserved heterogeneity—when sequential events are independent—or is drawn from a multiplicative Gamma distribution with a mean and variance equal to 1.0 when heterogeneity is introduced in the data.[5] The larger variance represents greater heterogeneity and produces higher correlations in event times. In addition, z_i is a dichotomous and time-invariant covariate, indicating, for example, whether a subject has received treatment or not.[6]

4 Varying the Form of Time Dependence

Research analyzing the behavior of Cox-type estimators has typically assumed the data are distributed exponentially, which implies that time dependence itself is irrelevant. But Bender, Augustin, and Blettner (2005) demonstrate the importance of the distribution of time-to-event (see also Harezlak and Tu 2006; Metcalfe and Thompson 2006), finding "complex effects" where the distribution of time-to-event "influences results in various directions and concerning different variables" (Bender, Augustin, and Blettner 2005, 1721). We focus on data generated using the Weibull distribution because it can produce either decreasing or increasing monotonic hazard rates and nests the exponential model. We assess the effect of each form of time dependence in combination with within-subject correlation on the behavior of the estimates of the covariate effects and the estimates of the random effects. By fully factoring our three time dependence conditions ($p = (1.0{,}0.8{,}1.2)$), our two event dependence conditions ($\lambda_{0k} = (\lambda_0, k\lambda_0)$), and our two unobserved heterogeneity conditions ($e^{\omega_i} = (1, \sim \text{Gamma}(1/\theta,\theta)$), we compare model performance across twelve conditions, summarized in Tables 1–4.

4.1 *Time Dependence, No Unobserved Heterogeneity, and No Event Dependence*

The first case we consider is that of Box-Steffensmeier and De Boef (2006). Our results replicate theirs: In the absence of time dependence ($p = 1.0$) and within-subject correlation ($\lambda_{0k} = \lambda_0$, $e^{\omega_i} = 1.0$), the Cox model assumptions are met and all the estimated models perform well. See Table 1, Panel A.

Adding time dependence by allowing a decreasing ($p = 0.8$) or increasing ($p = 1.2$) baseline hazard in the DGP given in equation (6) changes the implications of estimator choice. For nonconstant hazard rate processes with this DGP, only the gap-time models are correctly specified because as each event occurs, the risk starts over again/the clock restarts, as in the generated data. The gap-time models perform well when there is no heterogeneity and no event dependence because the only factor impacting conditional event rates is the treatment. The elapsed-time models, however, fail to compensate for the nonconstant hazard rates, here specifically how the hazard resets each time an event occurs. In the case of the exponentially distributed data, resetting never matters since the hazard is constant. Given increasing or decreasing hazard rates under the Weibull, however, once an event occurs the hazard rate returns to the level it was at the start.[7] So the hazard rate is much higher right after the event occurs, and then it decays over time under decreasing hazards. And under increasing baseline hazard rates, the hazard rate is much lower immediately after the event occurs, increasing over time. In general, elapsed-time models cannot accurately capture this relationship. The conditional elapsed-time model, however, allows for different baseline hazards by event and so picks up the "resetting" effect.

For decreasing baseline hazard rates, this misspecification increases estimates of the treatment's effectiveness. Treatment not only lowers the rate of an event, but also lowers the rate at which the

[5] Specifically, $e^{\omega_i} \sim \text{Gamma}(1/\theta,\theta)$, where $\theta = 1$. We use the multiplicative Gamma distribution here because of its flexibility and to build upon previous results.

[6] Half of the samples received treatment.

[7] Depending on the quantity of interest, some might choose to attribute a treatment's effect on the resetting process to a part of a treatment's total effect (Cheung et al. 2010).

Table 1 The role of time dependence: No event dependence, no heterogeneity

	$\hat{\beta}$	*SD*	*SE*	*Coverage rate*	$\hat{\theta}$	*Rejection rate*
Panel A: No time dependence						
Exponential						
Conditional frailty, gap	−1.002	0.070	0.072	0.963	0.003	0.001
Frailty, elapsed	−1.008	0.078	0.079	0.950	0.003	0.004
Andersen-Gill	−1.002	0.077	0.077	0.943		
Conditional, gap	−1.000	0.070	0.070	0.955		
Conditional, elapsed	−1.010	0.106	0.098	0.934		
Panel B: Decreasing hazards						
Weibull ($p = 0.8$)						
Conditional frailty, gap	−1.003	0.072	0.073	0.958	0.003	0.004
Frailty, elapsed	−1.237	0.105	0.093	0.294	0.052	0.56
Andersen-Gill	−1.162	0.094	0.093	0.600		
Conditional, gap	−1.000	0.071	0.070	0.946		
Conditional, elapsed	−1.015	0.101	0.094	0.935		
Panel C: Increasing hazards						
Weibull ($p = 1.2$)						
Conditional frailty, gap	−1.002	0.070	0.072	0.964	0.003	0.001
Frailty, elapsed	−0.874	0.065	0.076	0.600	0.000	0.000
Andersen-Gill	−0.874	0.065	0.065	0.479		
Conditional, gap	−1.000	0.070	0.070	0.955		
Conditional, elapsed	−1.017	0.110	0.103	0.937		

Notes. Simulation results are for $M = 1000$ replications. The DGP is given in equation (6) with $\beta = -1.0$, $\lambda_0 = 1.0$, and $N = 100$ cases. The true value of $\theta = 0$ and $\lambda_{0k} = \lambda_0$. Reported SEs for the Andersen-Gill, conditional elapsed-time, and conditional gap-time models are robust SEs:

$$\mathbf{V} = \mathbf{I}^{-1}\mathbf{B}\mathbf{I}^{-1},$$

where $\mathbf{I}^{-1}$ is the usual variance estimate of a Cox model (the inverse of the information matrix $\mathbf{I}$) and $\mathbf{B}$ is a correction factor based on the correlation within cases. SEs for the frailty and conditional frailty models follow Gray (1992): $\mathbf{V} = \mathbf{H}^{-1}\mathbf{I}\mathbf{H}^{-1}$. $\mathbf{H}$ is the second derivative matrix for the penalized likelihood.

hazard rate spikes up again. Elapsed-time models misattribute the difference in the baseline hazard as purely a result of treatment condition. This bias operates in a reverse direction for distributions with increasing baseline risk. Since the control group is more likely to experience an event, this will then reset its baseline risk to a lower level. Consequently, elapsed-time models misattribute this process as one where treatment has less of an effect. The only exception to this pattern is the result from the conditional elapsed-time model, which again picks up the resetting of the hazard rate via the varying baseline hazard rate, accommodating model misspecification. Nevertheless, the gap-time models perform the best when the baseline hazard is nonconstant; see Table 1, Panels B and C.

The coverage rates for our estimates of β are all near the nominal 95% rate. They remain so for the conditional frailty model under both increasing and decreasing baseline hazard rates. Similarly, coverage rates for both conditional models are near 95%. The Andersen-Gill and elapsed-time frailty model underestimate the coverage rate by as much as approximately 50%. Both the frailty model and the conditional frailty model correctly estimate random effects that are effectively 0.

4.2 *Time Dependence, Unobserved Heterogeneity, No Event Dependence*

Next we consider the case of unobserved heterogeneity with no event dependence. For a model without frailty terms, failure to account for unobserved heterogeneity produces estimates biased toward 0 in a form that is dependent on the magnitude of the frailty variance (Henderson and

Table 2 The role of time dependence: Unobserved heterogeneity, no event dependence

	$\hat{\beta}$	*SD*	*SE*	*Coverage rate*	$\hat{\theta}$	*Rejection rate*
Panel A: No time dependence						
Exponential						
Conditional frailty, gap	−0.984	0.228	0.206	0.918	0.950	1.000
Frailty, elapsed	−0.998	0.232	0.206	0.922	0.941	1.000
Andersen-Gill	−0.512	0.148	0.141	0.087		
Conditional, gap	−0.480	0.131	0.123	0.023		
Conditional, elapsed	−0.251	0.087	0.076	0.000		
Panel B: Decreasing hazards						
Weibull ($p = 0.8$)						
Conditional frailty, gap	−0.978	0.228	0.208	0.918	0.941	1.000
Frailty, elapsed	−1.177	0.274	0.233	0.820	1.349	1.000
Andersen-Gill	−0.589	0.164	0.158	0.270		
Conditional, gap	−0.472	0.124	0.117	0.014		
Conditional, elapsed	−0.278	0.100	0.084	0.000		
Panel C: Increasing hazards						
Weibull ($p = 1.2$)						
Conditional frailty, gap	−0.994	0.230	0.204	0.910	0.968	1.000
Frailty, elapsed	−0.673	0.159	0.156	0.428	0.488	1.000
Andersen-Gill	−0.402	0.121	0.115	0.004		
Conditional, gap	−0.500	0.149	0.139	0.063		
Conditional, elapsed	−0.235	0.077	0.070	0.000		

Notes. Simulation results are for $M = 1000$ replications. The DGP is given in equation (6) with $\beta = -1.0$, $\lambda_0 = 1.0$, and $N = 100$ cases. The random effect is generated from a multiplicative Gamma(1,1) such that the true value of $\theta = 1$. $\lambda_{0k} = \lambda_0$. Reported SEs for the Andersen-Gill, conditional elapsed-time, and conditional gap-time models are robust SEs:

$$\mathbf{V} = \mathbf{I}^{-1}\mathbf{B}\mathbf{I}^{-1},$$

where $\mathbf{I}^{-1}$ is the usual variance estimate of a Cox model (the inverse of the information matrix $\mathbf{I}$) and $\mathbf{B}$ is a correction factor based on the correlation within cases. SEs for the frailty and conditional frailty models follow Gray (1992): $\mathbf{V} = \mathbf{H}^{-1}\mathbf{I}\mathbf{H}^{-1}$. $\mathbf{H}$ is the second derivative matrix for the penalized likelihood.

Oman 1999). Within a model with negative treatment effect ($\beta = -1.0$) and when $\lambda_{0k} = 1.0$ at any time t, the observed population log of the hazard ratios equals

$$\begin{aligned}\log\left(\frac{h(t)_T}{h(t)_C}\right) &= \log\left[\frac{pt^{p-1}\exp(-1)[1-\theta\log(S(t))]^{-1}}{pt^{p-1}[1-\theta\log(S(t))]^{-1}}\right]\\ &= \log\left[\frac{pt^{p-1}\exp(-1)[1+\theta\exp(-1)t^p]^{-1}}{pt^{p-1}[1+\theta t^p]^{-1}}\right]\\ &= -1+\log\left[\frac{1+\theta t^p}{1+\theta\exp(-1)t^p}\right].\end{aligned}$$

At $t = 0$, the log of the population hazard ratios is − 1.0, but as t approaches infinity, the second component, which is always positive, gets larger at a rate based on θ, until it reaches its limit at 1.0. This failure to account for heterogeneity distorts the differences in hazards caused by the treatment. Eventually, both groups are observed to experience relatively equal hazard rates because at that point only the extremely strong reside in each group.

These properties are supported by our simulations. When the DGP is characterized by heterogeneity, only models designed to capture that unobserved heterogeneity—the frailty and conditional frailty models—perform well. These models eliminate the imbalance across treatment and control groups not due to the treatment. As predicted, all other models underestimate the treatment

Table 3 The role of time dependence: Event dependence, no heterogeneity

	$\hat{\beta}$	*SD*	*SE*	*Coverage rate*	$\hat{\theta}$	*Rejection rate*
Panel A: No time dependence						
Exponential						
Conditional frailty, gap	−1.002	0.070	0.072	0.964	0.003	0.001
Frailty, elapsed	−2.450	0.262	0.156	0.000	0.003	0.001
Andersen-Gill	−1.525	0.171	0.162	0.104		
Conditional, gap	−1.000	0.070	0.070	0.955		
Conditional, elapsed	−1.008	0.104	0.094	0.924		
Panel B: Decreasing hazards						
Weibull ($p = 0.8$)						
Conditional frailty, gap	−1.002	0.070	0.072	0.964	0.003	0.001
Frailty, elapsed	−3.378	0.411	0.199	0.000	1.801	1.000
Andersen-Gill	−1.619	0.208	0.193	0.129		
Conditional, gap	−1.000	0.070	0.070	0.955		
Conditional, elapsed	−1.122	0.106	0.099	0.772		
Panel C: Increasing hazards						
Weibull ($p = 1.2$)						
Conditional frailty, gap	−1.002	0.070	0.072	0.964	0.003	0.001
Frailty, elapsed	−1.807	0.173	0.130	0.000	0.385	1.000
Andersen-Gill	−1.370	0.138	0.133	0.199		
Conditional, gap	−1.000	0.070	0.070	0.955		
Conditional, elapsed	−0.944	0.104	0.095	0.859		

Notes. Simulation results are for $M = 1000$ replications. The DGP is given in equation (6) with $\beta = -1.0$, $\lambda_0 = 1.0$, and $N = 100$ cases. The true value of $\theta = 0$ and $\lambda_{0k} = k\lambda_0$. Reported SEs for the Andersen-Gill, conditional elapsed-time, and conditional gap-time models are robust SEs:

$$\mathbf{V} = \mathbf{I}^{-1}\mathbf{B}\mathbf{I}^{-1},$$

where $\mathbf{I}^{-1}$ is the usual variance estimate of a Cox model (the inverse of the information matrix $\mathbf{I}$) and $\mathbf{B}$ is a correction factor based on the correlation within cases. SEs for the frailty and conditional frailty models follow Gray (1992): $\mathbf{V} = \mathbf{H}^{-1}\mathbf{I}\mathbf{H}^{-1}$. $\mathbf{H}$ is the second derivative matrix for the penalized likelihood.

effects. See Table 2, Panel A. The frailty model's elapsed time also falters once there is time dependence and nonconstant hazards (Panels B and C). We find the conditional frailty model the most accurate; most other models experience bias toward zero, although for the elapsed-time model in Panel B this bias does not outweigh its original overestimate of treatment's effectiveness illustrated in Table 1.

Across the distributions, the coverage rates for $\hat{\beta}$ are all somewhat smaller than the nominal 95% level; in some cases only the conditional frailty model has coverage rates that approach 95%. In the case of the conditional models under an increasing baseline hazard, the rates for the Andersen-Gill and both conditional gap- and elapsed-time models have rejection rates approaching zero. In virtually all cases, the SEs are too small. Finally, although in the exponential both the conditional frailty and frailty models' estimate of θ approach 1.0, under decreasing baseline hazards the estimate of θ is 35% too big and under increasing baseline hazards the estimate is approximately 50% too low. The frailty model in elapsed time cannot parse the treatment effect and the random effect. In all cases, the null that $\hat{\theta}$ equals zero is rejected.

4.3 *Time Dependence, Event Dependence with No Heterogeneity*

Given a negative treatment effect, cases in the control group will experience more events than in the treatment group. When the data exhibit positive event dependence, the effect is exacerbated, raising

Table 4 The role of time dependence: Unobserved heterogeneity and event dependence

	$\hat{\beta}$	*SD*	*SE*	*Coverage rate*	$\hat{\theta}$	*Rejection rate*
Panel A: No time dependence						
Exponential						
Conditional frailty, gap	−0.991	0.230	0.204	0.908	0.963	1.000
Frailty, elapsed	−2.129	0.585	0.621	0.588	0.956	1.000
Andersen-Gill	−0.564	0.187	0.176	0.287		
Conditional, gap	−0.493	0.149	0.141	0.069		
Conditional, elapsed	−0.286	0.113	0.088	0.000		
Panel B: Decreasing hazards						
Weibull ($p = 0.8$)						
Conditional frailty, gap	−0.985	0.229	0.205	0.915	0.956	1.000
Frailty, elapsed	−2.912	0.810	0.457	0.000	4.795	1.000
Andersen-Gill	−0.611	0.200	0.190	0.450		
Conditional, gap	−0.483	0.143	0.135	0.043		
Conditional, elapsed	−0.334	0.137	0.101	0.000		
Panel C: Increasing hazards						
Weibull ($p = 1.2$)						
Conditional frailty, gap	−0.995	0.230	0.204	0.911	0.969	1.000
Frailty, elapsed	−1.666	0.411	0.418	0.655	2.255	1.000
Andersen-Gill	−0.526	0.174	0.163	0.181		
Conditional, gap	−0.502	0.157	0.146	0.086		
Conditional, elapsed	−0.255	0.098	0.080	0.000		

Notes. Simulation results are for $M = 1000$ replications. The DGP is given in equation (6) with $\beta = -1.0$, $\lambda_0 = 1.0$, and $N = 100$ cases. The random effect is generated from a multiplicative Gamma(1,1) such that the true value of $\theta = 1$. $\lambda_{0k} = k\lambda_0$. Reported SEs for the Andersen-Gill, conditional elapsed-time, and conditional gap-time models are robust SEs:

$$\mathbf{V} = \mathbf{I}^{-1}\mathbf{B}\mathbf{I}^{-1},$$

where $\mathbf{I}^{-1}$ is the usual variance estimate of a Cox model (the inverse of the information matrix $\mathbf{I}$) and $\mathbf{B}$ is a correction factor based on the correlation within cases. SEs for the frailty and conditional frailty models follow Gray (1992): $\mathbf{V} = \mathbf{H}^{-1}\mathbf{I}\mathbf{H}^{-1}$. $\mathbf{H}$ is the second derivative matrix for the penalized likelihood.

the event rate in the control group relative to the treatment group. Treatment thus looks more effective than it is, unless the models are stratified. A positive treatment effect would have a similar result because the treatment effect produces more cases having more events and event dependence produces yet more cases having more events. The effect of the treatment once again looks bigger than it is, as the effects of the event dependence are picked up by the estimated treatment effect. Thus, for the exponential, the frailty models that do not stratify overestimate treatment effects, whereas models that stratify perform well; see Table 3, Panel A. Introducing time dependence either exaggerates ($p = 0.8$) or mitigates ($p = 1.2$) elapsed model biases in the presence of event dependence, where the misspecified conditional elapsed-time model appears to be somewhat more robust to misspecification in this case; see Table 3, Panels B and C.

Under event dependence, we also find that SEs on the estimate of β tend to be too small, with the exception of the conditional frailty model, where they are slightly larger than the standard deviation (SD) of the estimates. Coverage rates for the frailty model are zero in all cases. This makes sense given how far off the estimates of β actually are. The Andersen-Gill model has similarly small rejection rates. The estimates of θ and the rejection rates on the null that $\theta = 0$ are not surprising. The conditional frailty model picks up event dependence with the event-specific baseline hazards so that it correctly estimates θ to be 0 and essentially never rejects the null. The elapsed-time frailty model's estimate of θ, although centered on 0 for the exponential case, misattributes event dependence's interaction with the nonconstant hazard rate as a product of unobserved heterogeneity. For both Weibull distributions, it estimates the frailty variance as larger than 0 and with rejection rates that approach 1.0.

4.4 *Time Dependence, Unobserved Heterogeneity, and Event Dependence*

The case where both event dependence and heterogeneity are present with time dependence is more complicated. Here, imbalance in conditional event rates in the control and treatment arms due to heterogeneity interact with those caused by event dependence. Under a negatively signed treatment effect, fewer cases experience a small number of events and more cases experience a large number of events in the control group. This magnifies the imbalance in conditional event rates due to heterogeneity so that models that do not control for both heterogeneity and event dependence will, in the first case, underestimate the treatment effect, and in the second case overestimate it. If treatment effects are positive, we see the same result because the probability of having an event conditional on the subject having no event or a small number of events is higher in the control group than in the treatment group. This makes treatment look less effective than it is, biasing estimates toward zero.

Importantly, we fail to find the combination of specification errors canceling out any misspecification bias. For instance, the Andersen-Gill estimates still exhibit substantial positive bias from unobserved heterogeneity, despite the negative biasing effects of event dependence (Table 4) or time dependence with decreasing hazards (Panel B). The effect of the combination of conditions is to increase the average SE estimate, but this only slightly improves the coverage rate.

Importantly, whereas heterogeneity and event dependence affect the conditional probability of experiencing an event in the treatment and control groups, the conditional frailty model controls for both heterogeneity and event dependence so that any imbalance in conditional event rates must be due to the treatment. We see the superior performance of the conditional frailty model across all the scenarios; see Table 4, Panels A–C.

Finally, when there is both event dependence and heterogeneity, we see that the SEs tend to be too small, in some cases substantially so. Coverage rates quickly fall off from 90% for the conditional frailty model to a 0% rejection rate for the frailty model under decreasing hazards, and the conditional elapsed-time model under both increasing and decreasing baseline hazards. Like the case of the DGP with event dependence only, $\hat{\theta}$ is too big, by 480%, and in the case of increasing hazards it is 226% too big. Using *R*'s penalized likelihood estimator, all estimates of θ had a 100% rejection rate, rejecting the null that $\theta = 0$ in every case.

4.5 *The Interaction of Time and Within-Subjection Correlation*

Examining the effects of time dependence under these four conditions provides us with a number of consistent findings. First, regardless of the existence of time dependence, omitting a random effect when there is unobserved heterogeneity results in an underestimation of the treatment effect, as was suggested in previous work (Box-Steffensmeier and De Boef 2006). Similarly, failure to estimate an event-specific baseline hazard when there is positive event dependence resulted in estimates of the treatment effect that were too big.

Moreover, both these biases are exacerbated by nonconstant hazard rates when incorrectly specifying a baseline hazard process. For instance, in Table 1 (Panel B) we found that decreasing hazard rates make treatment effects appear more effective in elapsed model estimates since treatment delays the experience of an event and, thus, keeps hazard rates at their lower levels for greater duration. But this difference in hazard rates is even greater if there is positive event dependence (Table 3, Panel B), as each successive event raises the hazard rate. Since the disparity in the rate at which the hazard rates change is associated with treatment, elapsed models attribute it to be a function of treatment. Likewise, since increasing hazard rates and unobserved heterogeneity both make treatment appear less effective, the combination of conditions make elapsed models underestimate treatment effects to a greater degree (Table 2, Panel C).

Importantly, in all cases, the conditional frailty model exhibits little bias and consistently exhibits the highest coverage rate. The flexibility of the model also appears to come at little cost, as the model performs equal to its more restrictive counterparts when there is no event dependence or unobserved heterogeneity (or both).

5 Robustness to Misspecifying Time Scale

The conditional frailty model exhibits superior performance across all conditions when the hazard rate is associated with time since last event. In contrast, the elapsed-time models performed especially poorly when hazard rates increased or decreased over time. But do these findings hold when the time scale is switched? In many practical applications, researchers face uncertainty over whether hazard rates are predominately a function of time since last event (gap time) or time since being at risk (elapsed time). For instance, the risk of civil war or international conflict may predominately be a function of how long a state, dyad, or regime has existed instead of time since last dispute (e.g., Maoz 1996). The relative performance of the conditional frailty model may change if the time scale process is misspecified.

We ran a similar set of experiments to compare the performance of the conditional frailty model to elapsed-time models but under a different time scale, making the hazard rate a function of time since origin.[8] A hazard rate that is flat over elapsed time is no different than our previous simulations with a hazard rate that is flat over gap time, so we only tabulate results when hazard rates are a function of a Weibull process that either increases ($p = 1.2$) or decreases ($p = 0.8$) over time. This results in four conditions for each hazard rate, where we again varied the presence or absence of event dependence and the presence or absence of unobserved heterogeneity.[9]

Table 5 summarizes our estimates across the four experimental conditions with an increasing hazard rate over elapsed time. In Panel A, the three elapsed-time models are all correctly specified and have coverage rates near 95%. The misspecified conditional frailty model underestimates the effect of treatment, but its estimates are not far off, with a coverage rate around 54%. The advantages of the model's flexibility are apparent when the DGP changes. If there is either event dependence or heterogeneity (Panels B and C), the correctly specified elapsed-time models produce unbiased estimates, but the other elapsed-time models perform much worse. In contrast, the small bias in the misspecified conditional frailty model shows little change. In both panels, the conditional frailty model exhibits the second best coverage rate. When there is both unobserved heterogeneity and event dependence, all four models are misspecified since none of the elapsed-time models account for both unobserved heterogeneity and event dependence. In this situation, the misspecified conditional frailty model performs the best, with a relatively small bias and a coverage rate of 85.1%.

The conditional frailty model's performance also compares favorably if the hazard rate instead decreases across time (Table 6). With a decreasing hazard rate over time, the conditional frailty model estimates treatment to have a slightly stronger preventative effect. But the model's flexibility makes its estimates robust, such that the relatively small bias is consistent across all four conditions. Again, when either unobserved heterogeneity or event dependence are present, the conditional frailty model's coverage rates are second to only the correctly specified model (Panels B and C). If the hazard rate process has both event dependence and heterogeneity, then the gap-time conditional frailty model outperforms all the elapsed-time models, with a much smaller bias and the highest coverage rate.

These results demonstrate the flexibility and robustness of the conditional frailty model to time-scale specification. Compared with elapsed model estimates of gap-time processes (Panels B and C in Tables 1–4), the conditional frailty estimates of a misspecified time scale perform either

[8] The simulation setup only differs in that simulations were done in successive order per subject, where each draw's value on the Weibull cumulative distribution function sets the range of the next uniform draw within the Weibull's inverse cumulative distribution function.

[9] We exclude the results for the conditional gap-time model since it rarely outperformed the conditional frailty model in the previous simulations.

Table 5 Elapsed-time Weibull, increasing hazards ($p = 1.2$, $\beta = -1.0$, $\theta = 1$)

	$\hat{\beta}$	*SD*	*SE*	*Coverage rate*	$\hat{\theta}$	*Rejection rate*
Panel A						
No event dep., no het.						
Conditional frailty, gap	−0.872	0.061	0.070	0.542	0.000	0.000
Frailty, elapsed	−1.008	0.078	0.079	0.950	0.004	0.004
Andersen-Gill	−1.002	0.077	0.077	0.943		
Conditional, elapsed	−1.010	0.106	0.098	0.934		
Panel B						
No event dep., het.						
Conditional frailty, gap	−0.841	0.195	0.185	0.866	0.758	1.000
Frailty, elapsed	−1.015	0.236	0.210	0.920	1.003	1.000
Andersen-Gill	−0.451	0.144	0.134	0.040		
Conditional, elapsed	−0.231	0.084	0.072	0.000		
Panel C						
Event dep., no het.						
Conditional frailty, gap	−0.859	0.062	0.069	0.450	0.000	0.000
Frailty, elapsed	−2.450	0.262	0.156	0.000	0.850	1.000
Andersen-Gill	−1.525	0.171	0.162	0.104		
Conditional, elapsed	−1.008	0.104	0.094	0.924		
Panel D						
Event dep., het.						
Conditional frailty, gap	−0.833	0.194	0.185	0.851	0.750	1.000
Frailty, elapsed	−2.238	0.669	0.472	0.325	3.200	1.000
Andersen-Gill	−0.540	0.189	0.173	0.242		
Conditional, elapsed	−0.282	0.112	0.087	0.000		

Notes. Simulation results are for $M = 1000$ replications. The DGP is given in equation (6) with $\beta = -1.0$, $\lambda_0 = 1.0$, and $N = 100$ cases. The random effect is generated from a multiplicative Gamma(1,1) such that the true value of $\theta = 1$. $\lambda_{0k} = k\lambda_0$. Reported SEs for the Andersen-Gill, conditional elapsed-time, and conditional gap-time models are robust SEs:

$$\mathbf{V} = \mathbf{I}^{-1}\mathbf{B}\mathbf{I}^{-1},$$

where $\mathbf{I}^{-1}$ is the usual variance estimate of a Cox model (the inverse of the information matrix $\mathbf{I}$) and $\mathbf{B}$ is a correction factor based on the correlation within cases. SEs for the frailty and conditional frailty models follow Gray (1992): $\mathbf{V} = \mathbf{H}^{-1}\mathbf{I}\mathbf{H}^{-1}$. $\mathbf{H}$ is the second derivative matrix for the penalized likelihood.

equivalently or better. Moreover, when event dependence and heterogeneity are present, the more flexible conditional frailty model performs the best.

In combination, for repeated events data we find that an incorrect assumption of a duration's time scale (elapsed or gap time) is not as problematic as a failure to account for forms of within-subject correlation. For instance, when only unobserved heterogeneity was present the frailty and conditional frailty models were the best performers across both settings (Panels B and C of Table 2 and Panel B of Tables 5 and 6). The different time-scale specifications do not change the relative balance of observations and event rates in the treatment and control group, only the composition or ordering of the risk sets among these observations. This produces some error when the Cox model divides out the baseline hazard, but it does not change the relative contribution of treatment and control observations to the partial likelihood. In contrast, much stronger biases occur when estimates fail to account for event dependence or unobserved heterogeneity. When either is present, the model estimates fail to account for within-subject correlation and the differences they create in the treatment and control group's event rate. Thus, a flexible approach to modeling within-subject

Table 6 Elapsed-time Weibull, decreasing hazard ($p = 0.8$, $\beta = -1.0$, $\theta = 1$)

	$\hat{\beta}$	*SD*	*SE*	*Coverage rate*	$\hat{\theta}$	*Rejection rate*
Panel A						
No event dep., no het						
Conditional frailty, gap	−1.201	0.086	0.082	0.329	0.036	0.399
Frailty, elapsed	−1.008	0.078	0.079	0.950	0.004	0.004
Andersen-Gill	−1.002	0.077	0.077	0.943		
Conditional, elapsed	−1.010	0.106	0.098	0.934		
Panel B						
No event dep., het.						
Conditional frailty, gap	−1.191	0.278	0.230	0.792	1.279	1.000
Frailty, elapsed	−1.003	0.234	0.206	0.921	1.002	1.000
Andersen-Gill	−0.493	0.145	0.139	0.072		
Conditional, elapsed	−0.243	0.086	0.074	0.000		
Panel C						
Event dep., no het.						
Conditional frailty, gap	−1.227	0.089	0.086	0.254	0.048	0.595
Frailty, elapsed	−2.450	0.262	0.156	0.000	0.850	1.000
Andersen-Gill	−1.525	0.171	0.162	0.104		
Conditional, elapsed	−1.008	0.104	0.094	0.924		
Panel D						
Event dep., het.						
Conditional frailty, gap	−1.219	0.286	0.233	0.764	1.337	1.000
Frailty, elapsed	−2.168	0.650	0.455	0.333	3.282	1.000
Andersen-Gill	−0.555	0.187	0.175	0.267		
Conditional, elapsed	−0.284	0.112	0.087	0.000		

Notes. Simulation results are for $M = 1000$ replications. The DGP is given in equation (6) with $\beta = -1.0$, $\lambda_0 = 1.0$, and $N = 100$ cases. The random effect is generated from a multiplicative Gamma(1,1) such that the true value of $\theta = 1$. $\lambda_{0k} = k\lambda_0$. Reported SEs for the Andersen-Gill, conditional elapsed-time, and conditional gap-time models are robust SEs:

$$\mathbf{V} = \mathbf{I}^{-1}\mathbf{B}\mathbf{I}^{-1},$$

where $\mathbf{I}^{-1}$ is the usual variance estimate of a Cox model (the inverse of the information matrix $\mathbf{I}$) and $\mathbf{B}$ is a correction factor based on the correlation within cases. SEs for the frailty and conditional frailty models follow Gray (1992): $\mathbf{V} = \mathbf{H}^{-1}\mathbf{I}\mathbf{H}^{-1}$. $\mathbf{H}$ is the second derivative matrix for the penalized likelihood.

correlation consistently outperforms the more restrictive, and the conditional frailty model remains preferable because it offers such flexibility in the face of uncertainty about the nature of the DGP.

6 Data Constraints

We turn next to the problem of data constraints. These are both different from and similar to the problems created by dependencies in the DGP. The nature of the problems is philosophically different; dependencies in the data are typically features of the underlying process. They make the process what it is. But data constraints are often limitations in sample size from a lack of resources, foresight of those coming before us to collect data, or the slow march of time since a historic event. Given enough time or resources the problem would, presumably, disappear. Yet they are similar to the problems above as well. Data constraints can create imbalance in the control and treatment arms that make treatment look more or less effective than it is, unless an estimator corrects for this imbalance in some way. Or, more insidious, data constraints may falsely create the illusion of balance in the control and treatment arms of the data, that without using the best model can lead to biased and inefficient estimates of the treatment effect.

7 Sample Size: Number of Cases and Number of Events

Flexible model specifications, like the conditional frailty model, may perform better in large N simulation studies, but repeated events data rarely meet these standards. More commonly, researchers often encounter a small number of cases or a small number of events per case. However, the research literature has yet to establish the number or distribution of each needed to produce stable frailty estimates (Therneau and Grambsch 2000, 256). Consequently, our next set of experiments look at the role of sample size for the mean squared error of the estimated treatment effects to examine the potential efficiency gains in parsimony and when, if ever, they change the choice of estimator.

Sample sizes vary tremendously across disciplines and subfields. Suciu (2002) defines a small sample as forty-five or fewer observations. In contrast, data sets of thousands of observations are also possible. In the repeated events case, sample size is equal to the product of the number of events and the number of individuals plus the product of any censored observations and the number of individuals who are censored. This means that sample size is a function both of number of events and of number of cases. We separately assess the effects of both cases and events.

Box-Steffensmeier and De Boef (2006) show that when an estimator is able to correctly capture the DGP, more events provide more information with which to estimate covariate effects and thus lead to a reduction in bias. When the estimator does not capture the features of the DGP, more events yield more biased estimates. However, this work did not address the role of an increasing sample size that occurs with increasing the numbers of events. We hold censoring and sample size constant to focus on the role played by the number of events experienced and the number of cases. This allows us to answer the question of how sample size matters, whether it is the number of events experienced across the control and treatment arms of a variable or the number of cases.

We first vary the number of cases from ten to 330, whereas the number of events is fixed at three per case. Next we isolate the effect of the number of events, fixing the number of cases at a hundred while allowing the number of events to grow from two to ten. In either scenario, the sample size ranges up to approximately a thousand. In all cases, the data are generated from exponentially distributed times-to-events as in equation (6). Once again, we compare results across conditions in which there is neither heterogeneity nor event dependence, heterogeneity but no event dependence, event dependence without heterogeneity, and event dependence with heterogeneity. These results are summarized in two four-panel graphs in Figs. 2 and 3, which plot our estimates of each model's mean squared error.

When we increase cases while holding events per case constant, the results are straightforward. For each modeling approach, as the ratio of cases to events within cases increases, model estimates consistently improve and grow nearer to the true treatment effect. We consistently see the greatest improvements for each estimator when the number of cases moves from ten to thirty to fifty. For data with a small number of cases relative to instances of repeated events, it is clear that models that fail to accommodate the true DGP are often far from accurate. After about fifty cases, however, these error rates are reduced, although they never reach 0.

Another finding from these results is that efficiency concerns in small samples rarely motivate the more parsimonious estimator. In instances with no event dependence and no heterogeneity all models do well, since all model specifications include the DGP. The elapsed models do slightly worse than the two gap-time specifications and, among these two groups, estimation of an unnecessary frailty parameter increases the mean squared error at most by 4%. Likewise, for event dependence with no heterogeneity, the mean squared error is reduced by about 4% if one can correctly choose the more parsimonious conditional gap model. In cases with unobserved heterogeneity and event dependence, the misspecified conditional gap model might be preferred on efficiency grounds over the conditional frailty if the number of cases is as low as ten. But in all other instances, the unbiased conditional frailty model also possesses the smallest mean squared error.

Figure 3 shows that when cases are held constant, but repeated events increase per case, the differences in estimators grow. Those models that do not encompass the true DGP frequently produce worse estimates as the number of observed events per case increase. Increasing the number of events magnifies the effects of imbalance in the treatment and control arms, so it

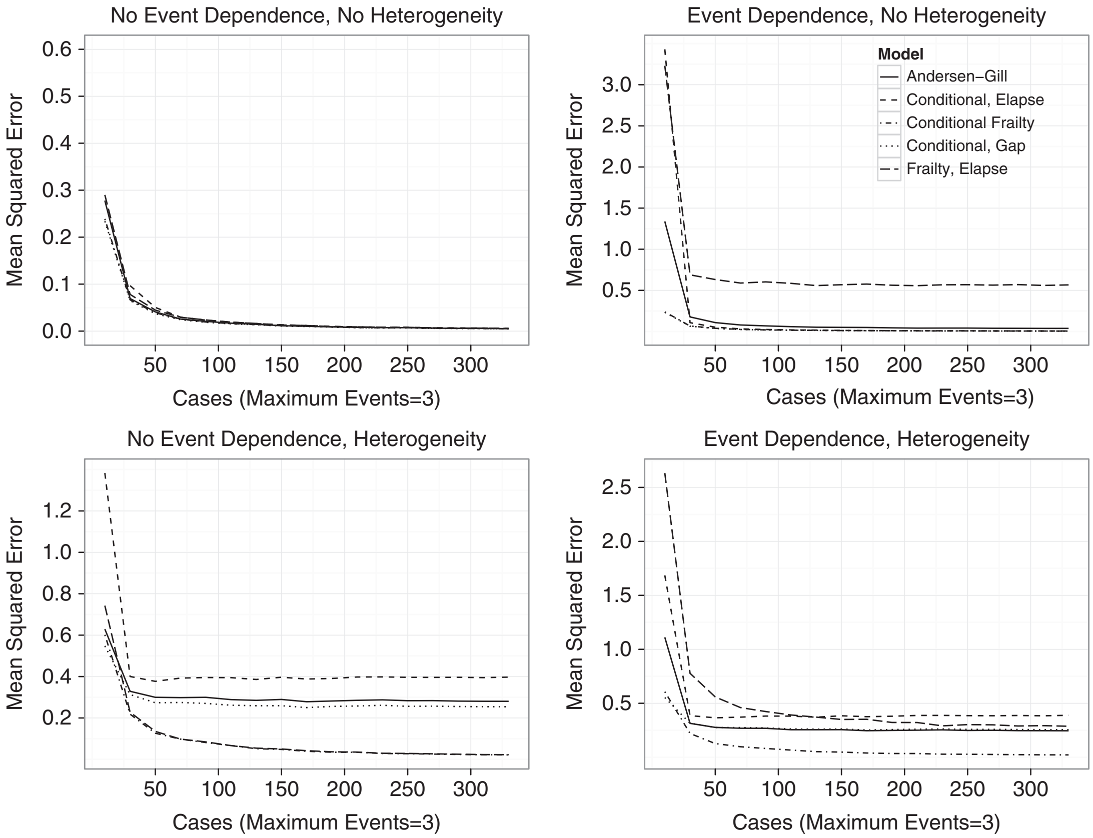

Fig. 2 The effects of sample size: estimates of mean squared error as the number of cases increases. Simulation results are for $M = 1000$ replications with exponentially distributed times-to-event. The DGP is given in equation (6) with $p = 1.0$, $\beta = -1.0$, and $\lambda_0 = 1.0$. The random effect is generated from a multiplicative Gamma(1,1) such that the true value of $\theta = 1$ under heterogeneity. Under event dependence, $\lambda_{0k} = k\lambda_0$, otherwise $\lambda_{0k} = \lambda_0$. N ranges from ten to 330 and $k = 3$ for sample sizes ranging from 30 to 990.

becomes easier to see the effects of imbalance on the estimators that do not encompass the DGP. More information, in the form of more events, means that the estimator performs as poorly or more poorly since it enhances the effects of misspecification.

Taken together, we find that concerns about misspecification are perhaps greatest not when sample sizes are small, but when the number of repeated events relative to the number of sampled cases is high. An increase in the number of cases relative to events lessens some concerns of misspecification, as it allows some misspecified models to approach the true treatment effect. But this reduction in bias decreases marginally and to a limit, such that these estimates consistently remain biased and less efficient. More generally, we find that conditional and frailty estimators perform well even with few observations, such that the conditional frailty model is the optimal estimator regardless of sample size or events per case considerations.

8 Censoring

Our last set of simulations focus on the role of censoring. Censoring occurs when information is incomplete for an individual. In the repeated events case, it is only the last observation for an individual that is counted as censored. In political science, we have censoring when participants drop out of the study or the study ends before an event occurs. Censoring is common to studies of repeated events and rates vary across real-world applications. Yet, the repeated events literature is surprisingly silent on the effects of varying degrees of censoring.

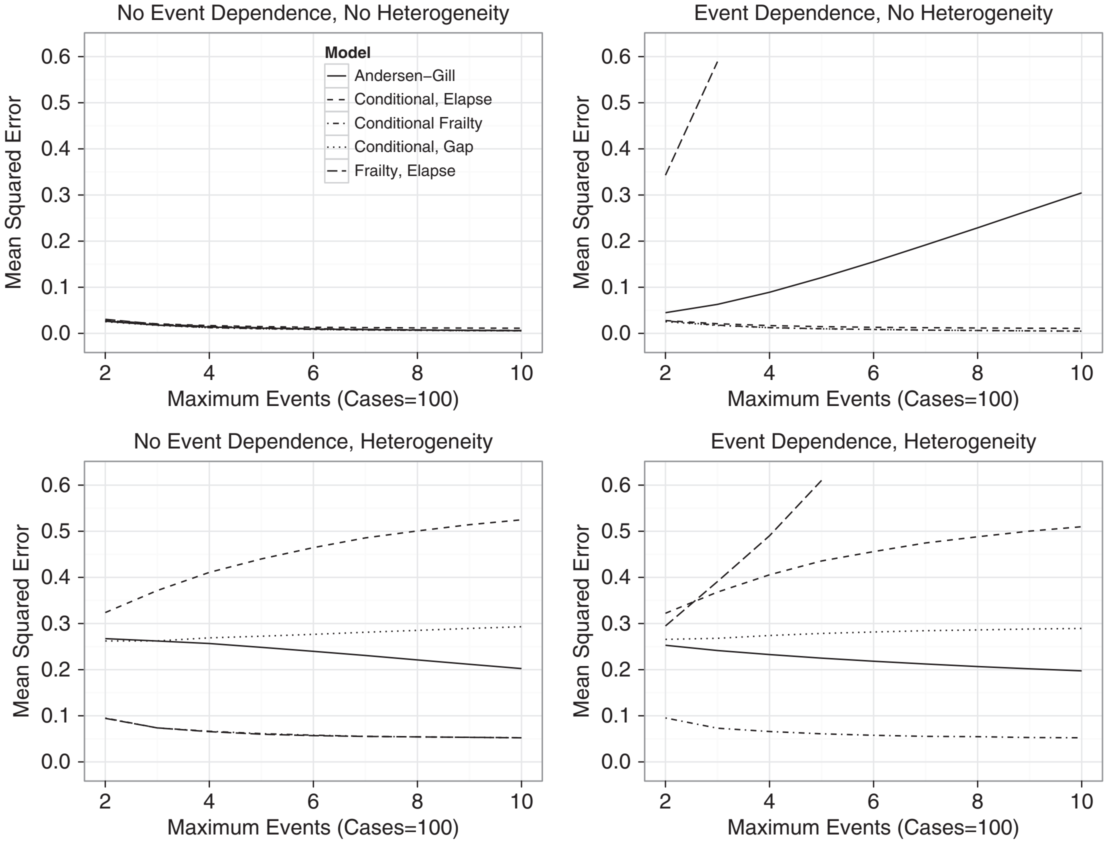

Fig. 3 The effects of sample size: estimates of mean squared error as the number of events increases. Simulation results are for $M = 1000$ replications with exponentially distributed times-to-event. The DGP is given in equation (6) with $p = 1.0$, $\beta = -1.0$, and $\lambda_0 = 1.0$. The random effect is generated from a multiplicative Gamma(1,1) such that the true value of $\theta = 1$ under heterogeneity. Under event dependence, $\lambda_{0k} = k\lambda_0$, otherwise $\lambda_{0k} = \lambda_0$. $N = 100$ and k ranges from two through ten for sample sizes ranging from two hundred to a thousand.

Censoring represents a combination of potential effects that may improve the relative performance of parsimonious models over frailty or conditional model specifications. That is, censoring may have an impact by reducing sample size and by creating an imbalance in the number of events observed for each case. To compare estimator performance, we again work with data generated from exponentially distributed times-to-event ($p = 1.0$).

We present our results for a random censoring mechanism that is not correlated with treatment effects or the hazard rate. We begin by generating data so that each case experienced the maximum number of events. Next, we randomly selected a case to be censored. In the third step, we randomly selected the event number for which the case was to be censored. Finally, we randomly selected the time at which the case was censored for that event number. This was repeated until we had censored enough cases to reach the desired censoring rate, which ranged from 0% to 50%.[10] Censoring the data in this manner has the effect of producing an identical range of censoring rates *regardless of the presence/absence of heterogeneity and event dependence*, only the event times will vary across conditions.

At times, we supplement our discussion with additional findings from a second set of experiments that censored the data by varying the follow-up time for our fixed baseline hazard rate,

[10] In all cases, the baseline hazard was fixed at 1.0 and we began with a follow-up time of 1 and ended with a follow-up time of 100.

selecting time intervals that gave us a wide range of variation in censoring rates. This form of censoring is correlated with treatment effects and other factors associated with the hazard rate, and may be common to certain political science data sets. The shorter the follow-up time, the more cases were censored before experiencing an event, especially those cases with lower hazard rates.[11]

We present a summary of our results in Fig. 4. When there is no censoring for the setting with no unobserved heterogeneity and no event dependence, the mean estimates of the treatment effect match those in Table 1, Panel A (by definition). The only factor affecting conditional event rates is the treatment. As the amount of censoring increases, the effect is to increase the size of the estimated treatment effect, but only mildly so (on the order of 5%–10%).

In this setting, the degree of bias is slightly larger for the three conditional models and most likely a function of biased likelihood estimation in small samples. With 50% random censoring, there is a 50% chance of being censored before each event for each observation. Meaning, on average, we have two hundred observations where fifty out of a hundred cases get censored before event 1, twenty-five out of the remaining fifty cases get censored before event 2, and we average less than one observation remaining before event 7. Since the conditional models generate their estimates by comparing within different groups defined by the number of previous events, the estimates of the different baseline hazard are more prone to small sample bias as observations per stratum dwindle (Sun and Yang 2000; Therneau and Grambsch 2000, 68). Therefore, researchers should seek a sufficient number of observations per stratum to ensure accurate estimation of conditional models.

When we have event dependence and no heterogeneity, the conditional elapsed- and gap-time models and conditional frailty model, each of which captures the true DGP by stratifying, perform well with no censoring. Again, similar to what we observe in the previous setting, there is a slight increase in the size of the estimated treatment effect, on the order of 10%, when censoring rates increase to 0.5 and observations per stratum are limited. Thus, regardless of whether event dependence exists, stratified estimates slightly overestimate treatment effects when there are fewer observations to compare within stratum. But this degree of bias is relatively small and is likely easily addressed if researchers combine the few cases with many events into one stratum.

As already noted, the Andersen-Gill and frailty models overestimate the treatment effect size, in this case because they do not control for event dependence (see Table 3, Panel A). The degree of overestimation ranges from 35% to 50% for the Andersen-Gill model and 50% to 150% for the frailty model in elapsed time. Again, cases in the control group will experience k events more quickly than in the treatment group, even more so than for the conditional models because the event dependence is left uncontrolled and assumed to be a function of treatment. Censoring can limit this bias, however, because it can lessen the number of events observed per case. Again, with 50% censoring, half of the observations are censored before the first event occurs, such that the average number of observations (both censored and uncensored) per case is two. With fewer events experienced, the differences in the rate these k events are experienced in the control and treatment groups diminish. Since fewer events per case reduce bias in the size of the estimated treatment effect (Fig. 3), it pulls the mean $\hat{\beta}$ back toward -1.0.

In the case of heterogeneity only (bottom left graph), the frailty model and the conditional frailty model both encompass the true DGP so that bias is small under no censoring (Table 2, Panel A). However, censoring and the reduction in number of observations have different consequences for these two estimators. In this setting, increased censoring and reduced events per case produce an underestimate of θ and a slight underestimate of the treatment effect. For a frailty model in elapsed time, the mean estimate is biased downward by about 10% with censoring rates of approximately 0.3. This bias is slightly less for the conditional frailty model because, as discussed above, it also possesses a small positive bias from stratification.

[11] In all cases, the baseline hazard was fixed at 1.0 and we began with a follow-up time of 1 and ended with a follow-up time of 50. This produced censoring rates from near 0% to 60%, although censoring rates varied depending on whether heterogeneity and/or event dependence were present.

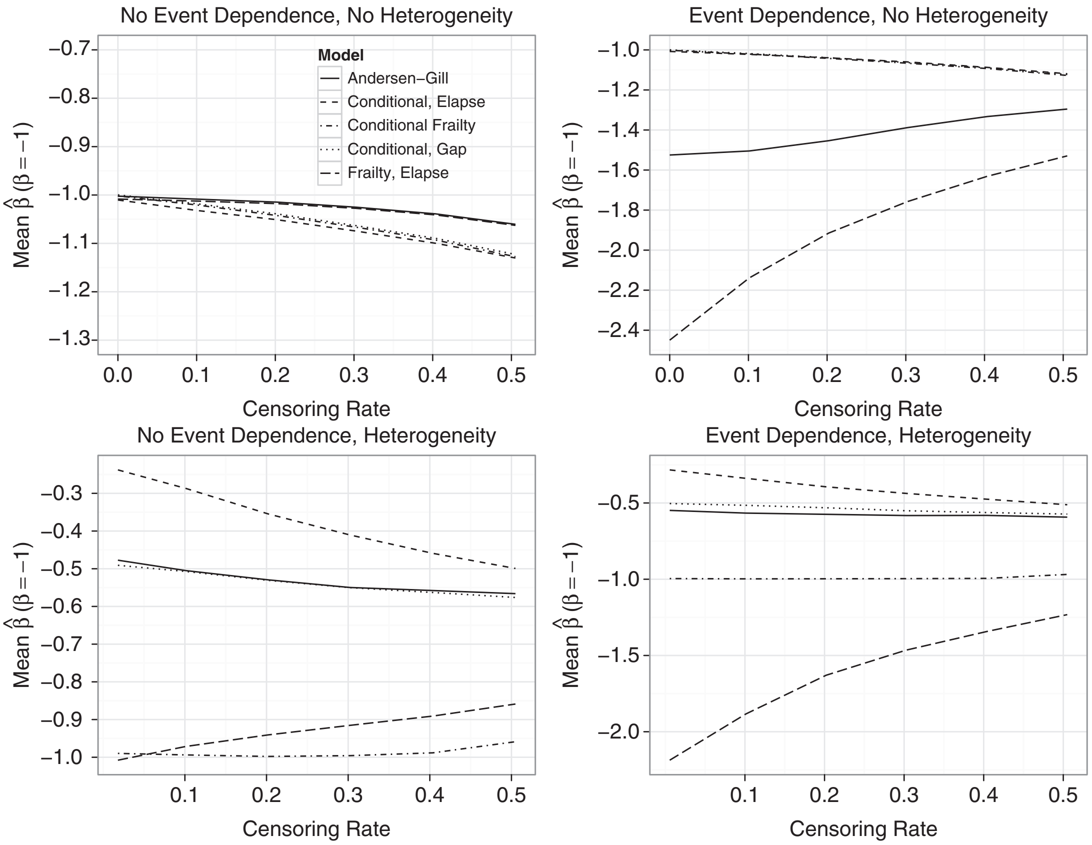

Fig. 4 The effects of random censoring: mean estimates of β as censoring rate increases. Simulation results are for $M = 1000$ replications with exponentially distributed times-to-event. The DGP is given in equation (6) with $p = 1.0$, $\beta = -1.0$, $\lambda_0 = 1.0$, and $N = 100$. The random effect is generated from a multiplicative Gamma(1,1) such that the true value of $\theta = 1$ under heterogeneity. Under event dependence, $\lambda_{0k} = k\lambda_0$, otherwise $\lambda_{0k} = \lambda_0$. Details on the censoring procedure can be found in the text.

In contrast, the models that lack a frailty term all underestimate the treatment effect because heterogeneity means that the frail will be disproportionately in the higher strata for the treatment group. And again, with fewer cases experiencing all the events, censoring produces fewer cases in these higher strata such that the disproportionality is reduced. However, even in cases of extreme censoring of 50%, the mean estimates of the treatment effect are about half the true size. Based on our additional set of simulations, we found it is possible that censoring can make this bias even smaller for these three estimators if censoring is correlated with time. In short, the frail in the treatment group are not only more likely to be in higher strata, but also are more likely to be censored, which reduces the imbalance. However, even in these settings the imbalance is never lessened to a point that their bias is less than 20%.

We believe these results that indicate data limitations should rarely constrain researchers from including frailty terms, even in cases of extreme censoring or severe imbalance in repeated events. For instance, compare the Andersen-Gill and the elapsed frailty estimates across the two graphs in the right column of Fig. 4. Across all censoring conditions, the elapsed frailty model performs essentially equal to the Andersen-Gill model when no heterogeneity is present.[12] However, even with only 50% of cases providing multiple observations, the frailty model is a substantially better

[12] Even with 50% censoring, the Andersen-Gill provides only 0.1% reduction in the mean squared error versus the elapsed model.

estimator than the Andersen-Gill when unobserved heterogeneity is present. Thus, the benefits of accounting for unobserved heterogeneity far outweigh costs from data limitations.

Finally, under both event dependence and heterogeneity, we see the various estimators performing as they did when they did not encompass the true DGP under the occurrence of event dependence or heterogeneity. The three estimators that do not account for heterogeneity underestimate the treatment effect substantially for all censoring rates because unobserved heterogeneity is the main force of bias in their treatment effect estimates. The frailty model that does not account for event dependence overestimates the treatment effect, with the greatest infraction occurring with no censoring. In this case, the failure to include event dependence is the main contributor of bias, although this again reduces by censoring rates. These findings only slightly change within our additional set of simulations that examine time-correlated censoring. The Andersen-Gill model switches from averaging an underestimate to an overestimate when censoring rates are above 20% in this setting because the biasing effect of a failure to model event dependence surpasses the biasing effects of unobserved heterogeneity.

Once again, however, the conditional frailty model performs very well over the full censoring range we have examined. There is a very small drop in the estimate of the treatment effect size as censoring increases beyond about 30%. But even when half the cases are censored, the mean estimate of the treatment effect is -0.954. Again, the slight downward bias of frailty estimates with high censoring compensates for the slight upward bias associated with stratifying estimates when there are few cases within each strata. Consequently, even for data with high censoring rates, where there are very few observations per case and an imbalance of events per case, the conditional frailty specification has little costs associated with it. Stratified estimates of event dependence provide a marginal overestimation of treatment when cases per stratum are few, but these can often be easily accounted for by the researcher. Moreover, although frailty estimates can underestimate unobserved heterogeneity and, thus, treatment effect with high rates of censoring, their estimates still outperform more parsimonious specifications.

9 Conclusions

Analysts wishing to understand the effects of covariates on repeated events processes face many challenges. There is often considerable uncertainty about the nature of the DGP. Specifically, researchers have little theoretical basis for selecting the form of time dependence underlying the distribution of event recurrence. Adding to the problem, analysts often cannot measure or do not know all the variables that influence event recurrence and have limited knowledge about whether the occurrence of an event changes the risk associated with future events.

We have shown that these features of the DGP are not always innocuous. In particular, choosing an estimator that does not properly account for within-subject correlation typically produces large biases in the estimated treatment effect because the misspecification, whatever its form, is attributed to the independent variable. Heterogeneity and event dependence induce bias if left unmodeled when there is no time dependence and the problem is exacerbated when there is time dependence in the data, specifically when the underlying hazard increases or decreases monotonically over time. These findings are robust to time scales and hold even when the analyst has the luxury of a large sample size and no censoring. The message from our analysis is clear: The only consistent defense against the problems associated with uncertainty about the DGP and within-subject correlation is to estimate the most flexible model, like the conditional frailty model.

Further, in many cases not only does uncertainty about the DGP plague the analysis, but data collected are also affected by small samples due either to a small number of cases or events and also by censoring. Our results pinpoint the source of the sample size problem: When the number of repeated events relative to the number of cases is high, the mean squared error associated with the models grows, unless the model encompasses the DGP. Again, using the conditional frailty model serves as the best defense against the problem. Censoring, even moderate levels, often produces large biases in the estimated treatment effects which, when combined with the problems of unobserved heterogeneity and event dependence, are difficult to predict. This is true even without the complication of time dependence. Under these conditions, the bias incurred using the conditional frailty model is minimal and always at least as small as the alternatives.

It is important to note that our findings apply only for the Weibull and exponential distributions of times-to-event. More nuanced forms of event dependence associated with alternative distributions for times-to-event may result in baseline hazards that increase, then decrease. Nonmonotonic forms of time dependence are more complicated and may not be well captured by the models we estimate here. This remains for future research.

Our comparisons are also specific to the Cox family of models, but our findings and conclusions likely apply to different repeated events estimators as well. In discrete time settings, researchers may estimate a logit with clustered SEs to account for repeated events. But accounting for event dependence or unobserved heterogeneity is less common in published work, as few studies specify a random effects logit and fewer still include event-specific time dummy variables or event-specific time splines. Our findings suggest that the failure to consider these sources of within-subject correlation will substantially bias results, even if repeated events are infrequent.

The takeaway message from our analysis, though, is clear. For continuous event-history data, the conditional frailty model has the flexibility to capture multiple aspects of DGPs and observed features of the data that other models commonly used in applied research and analyzed here cannot. It has proven to be versatile and robust, outperforming or performing equally as well as more parsimonious models, such as the elapsed-time frailty and nonfrailty conditional gap-time model. Importantly, the conditional frailty model did so under substandard conditions, such as a misspecified time scale, low sample size, or high censoring. This strongly suggests that we use the conditional frailty model when estimating models for repeated events in general. Thus, we are more confident in the utility of the conditional frailty model as "recommended" across the commonly applied settings we analyze here.

Funding

This work was supported by the National Science Foundation [SES-0648469 and SES-0647936 to J.M.B-S. and S.L.].

References

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

Baccini, L. 2012. Democratization and trade policy: An empirical analysis of developing countries. *European Journal of International Relations* 18(3):455–79.

Beardsley, K. 2008. Agreement without peace? International mediation and time inconsistency problems. *American Journal of Political Science* 52(4):723–40.

Bender, R., T. Augustin, and M. Blettner. 2005. Generating survival times to simulate Cox proportional hazards models. *Statistics in medicine* 24(11):1713–23.

Boehmke, F., D. Morey, and M. Shannon. 2006. Selection bias and continuous-time duration models: Consequences and a proposed solution. *American Journal of Political Science* 50(1):192–207.

Boehmke, F. J., and P. Skinner. 2012. State policy innovativeness revisited. *State Politics and Policy Quarterly* 12:304–30.

Boehmke, F. J., and R. Witmer. 2004. Disentangling diffusion: The effects of social learning and economic competition on state policy innovation and expansion. *Political Research Quarterly* 57(1):39–51.

Box-Steffensmeier, J. M., and S. De Boef. 2006. Repeated events survival models: The conditional frailty model. *Statistics in Medicine* 25(20):3518–33.

Brancati, D., and J. Snyder. 2011. Rushing to the polls: The causes of premature postconflict elections. *Journal of Conflict Resolution* 55(3):469–92.

Brown, M. 1996. *The international dimensions of internal conflict*. Cambridge, MA: MIT Press.

Cheung, Y., Y. Xu, S. Tan, and P. Milligan. 2010. Estimation of intervention effects using first or multiple episodes in clinical trials: The Andersen-Gill model re-examined. *Statistics in Medicine* 29(3):328–36.

Cook, R. J., J. Lawless, and C. Nadeau. 1996. Robust tests for treatment comparisons based on recurrent event responses. *Biometrics* 52(2):557–71.

Cook, R. J., and J. F. Lawless. 1997. An overview of statistical methods for multiple-failure time data in clinical trials: Discussion. *Statistics in Medicine* 16(8):841–43.

Curini, L. 2011. Government survival the Italian way: The core and the advantages of policy immobilism during the first republic. *European Journal of Political Research* 50(1):110–42.

Dugan, L., G. LaFree, and A. Piquero. 2005. Testing a rational choice model of airline hijackings. *Intelligence and Security Informatics* 3495:513–29.

Fernandez, J. 2010. Economic crises, high public pension spending, and blame-avoidance strategies: Pension policy retrenchments in 14 social-insurance countries, 1981–2005. Technical report, MPIfG Discussion Paper.

Gray, G. R. 1992. Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *Journal of the American Statistical Association* 87(420):942–51.

Greig, J. M. 2001. Moments of opportunity: Recognizing conditions of ripeness for international mediation between enduring rivals. *Journal of Conflict Resolution* 45(6):691–718.

Harezlak, J., and W. Tu. 2006. Estimation of survival functions in interval and right censored data using STD behavioral diaries. *Statistics in Medicine* 25(23):4053–64.

Henderson, R., and P. Oman. 1999. Effect of frailty on marginal regression estimates in survival analysis. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 61(2):367–79.

Kelly, P. J., and L. L.-Y. Lim. 2000. Survival analysis for recurrent event data: An application to childhood infectious disease. *Statistics in Medicine* 19:13–33.

Kuhn, U. 2009. Stability and change in party preference. *Swiss Political Science Review* 15(3):463–94.

Leighley, J. E., and J. Nagler. 2013. *Who votes now? Demographics, issues, inequality, and turnout in the United States*. Princeton, NJ: Princeton University Press.

Li, Q., and S. Lagakos. 1997. Use of the Wei-Lin-Weissfeld method for the analysis of a recurring and a terminating event. *Statistics in Medicine* 16(8):925–40.

Maoz, Z. 1996. *Domestic sources of global change*. Ann Arbor, MI: University of Michigan Press.

Martin, L., and G. Vanberg. 2003. Policing the bargain: Coalition government and parliamentary scrutiny. *American Journal of Political Science* 48(1):13–27.

Metcalfe, C., and S. Thompson. 2006. The importance of varying the event generation process in simulation studies of statistical methods for recurrent events. *Statistics in Medicine* 25(1):165–79.

Oakes, D. 1992. Frailty models for multiple event times. In *Survival analysis, state of the art*, eds. John P. Klein and P. K. Goel. The Netherlands: Kluwer Academic Publishers.

Pepe, M., and J. Cai. 1993. Some graphical displays and marginal regression analysis for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association* 88(423):811–20.

Plutzer, E. 2002. Becoming a habitual voter: Inertia, resources, and growth in young adulthood. *American Political Science Review* 96(1):41–56.

Prentice, R., B. Williams, and A. Peterson. 1981. On the regression analysis of multivariate failure time data. *Biometrika* 68(2):373–9.

Rohde, D. W., and D. M. Simon. 1985. Presidential vetoes and congressional response: A study of institutional conflict. *American Journal of Political Science* 29:397–427.

Schneider, G., and N. Wiesehomeier. 2008. Rules that matter: Political institutions and the diversity: Conflict nexus. *Journal of Peace Research* 45(2):183–203.

Stukel, T. 1993. Comparison of methods for the analysis of longitudinal interval count data. *Statistics in Medicine* 12(14):1339–51.

Suciu, G. P. 2002. Nonparametric survival comparison methods. Typescript.

Sun, J., and I. Yang. 2000. Nonparametric tests for stratum effects in the Cox model. *Lifetime Data Analysis* 6(4):321–30.

Therneau, T. M., and P. M. Grambsch. 2000. *Modeling survival data: Extending the Cox model*. Statistics for Biology and Health. New York: Springer.

Therneau, T. M., and S. A. Hamilton. 1997. rhDNase as an example of recurrent event analysis. *Statistics in Medicine* 16(18):2029–47.

Wei, L. J., D. Y. Lin, and L. Weissfeld. 1989. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* 84(408):1065–73.