STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:3518–3533
Published online 12 December 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2434

Repeated events survival models: The conditional frailty model

Janet M. Box-Steffensmeier[1] and Suzanna De Boef[2,*,†]

[1]*Department of Political Science, Ohio State University, Columbus, OH, U.S.A.*
[2]*Department of Political Science, Penn State University, University Park, PA 16802, U.S.A.*

SUMMARY

Repeated events processes are ubiquitous across a great range of important health, medical, and public policy applications, but models for these processes have serious limitations. Alternative estimators often produce different inferences concerning treatment effects due to bias and inefficiency. We recommend a robust strategy for the estimation of effects in medical treatments, social conditions, individual behaviours, and public policy programs in repeated events survival models under three common conditions: heterogeneity across individuals, dependence across the number of events, and both heterogeneity and event dependence. We compare several models for analysing recurrent event data that exhibit both heterogeneity and event dependence. The conditional frailty model best accounts for the various conditions of heterogeneity and event dependence by using a frailty term, stratification, and gap time formulation of the risk set. We examine the performance of recurrent event models that are commonly used in applied work using Monte Carlo simulations, and apply the findings to data on chronic granulomatous disease and cystic fibrosis. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: repeated events survival models; heterogeneity; event dependence; frailty

1. INTRODUCTION

Correlated event times are common in the study of health and related sciences. Correlation may occur when individuals experiencing a single event belong to groups or clusters, such as families or treatment centres, respectively. Alternatively, correlation may be due to recurrent events processes—where the subject experiences the same type of event more than once, such as hospital stays or heart attacks. In the case of recurrent events, correlation can come from two distinct sources:

1. *Heterogeneity across individuals*: In any study, some cases have a higher or lower event rate than other cases due to unknown, unmeasured, or unmeasurable effects. Individuals

*Correspondence to: S. De Boef, Department of Political Science, Penn State University, University Park, PA 16802, U.S.A.
†E-mail: sdeboef@psu.edu

Received 5 August 2004
Accepted 14 September 2005

have varied lifestyles, genetic traits, and experiences, for example, which influence the likelihood that they will succumb to disease but either cannot be measured or are unknown. As a result, some individuals are more prone to disease, experiencing their first, second, third, etc., disease recurrence more quickly than other individuals. This introduces heterogeneity across individuals and produces within-subject correlation in the occurrence and timing of recurrent events within a given subject. At the same time, response rates can be homogeneous within individuals producing within-subject correlation in event times.

2. *Event dependence*: The occurrence of one event may make further events more or less likely. This event (or occurrence) dependence may be produced by a biological weakening (damage effects) or strengthening (resistance effects). Either of these phenomenon implies that the risk for an event is a function of the occurrence of previous spells. This creates within subject correlation as well.

Medical research and clinical experience suggest that both heterogeneity and event dependence are likely to be the rule, rather than the exception, in the study of recurrent events. Our work attempts to separate the effects of event dependence from heterogeneity (see also References [1–3], which we build upon). The Cox [4] proportional hazards model and its extensions have been widely used to model correlated events. In particular, recent work has systematically characterized and compared these models in the context of recurrent events [5, 6]. Our aim is to compare several models for analysing recurrent event data that exhibits both heterogeneity and event dependence. We include a gap time conditional frailty model, which incorporates a random effect—to account for heterogeneity—with stratification and a conditional definition of the risk set—to account for event dependence. We discuss commonly used models for recurrent events and the conditional frailty model in Sections 2 and 3, respectively. We assess the performance of these models using Monte Carlo simulations in Section 4. The simulations focus on how heterogeneity and event dependence both individually and jointly affect estimates and hypothesis tests. In Section 5, we apply the modelling strategy to data sets on patients with chronic granulomatous disease and cystic fibrosis. Section 6 concludes with recommendations for analysts studying repeated events.

2. REPEATED EVENTS PROCESSES: FEATURES AND MODELS

It is well known that any correlation among events—produced individually or jointly by heterogeneity and event dependence—violates the Cox model's assumption that the timing of events is independent. This has two important consequences: the Cox model is both biased and inefficient in typical repeated events contexts [6–9]. Variations of the Cox model, namely *variance-corrected* and *frailty/random effects* models have been proposed for estimation with recurrent events to account for the correlation.

Variance-corrected models were developed to account for unobserved, or at least unaccounted for, heterogeneity by using robust standard errors, and sometimes stratification. Variations within the family of variance-corrected models are based on different definitions of the risk set, i.e. how individuals are defined to be at risk for any given event, k, and whether they allow for event specific baseline hazards. In the Andersen–Gill (AG) [10] model all cases are at risk for each event in all periods and share a common baseline rate function. In

contrast, conditional models stratify the data by event so that the baseline hazard is allowed to vary with each event. Conditional models are estimated in elapsed time or in gap time and cases are designated at risk for event k only after experiencing the k—1st event [11, 12]. Elapsed time estimation produces the hazard of an event since the study began while the gap time formulation gives the hazard since the previous event. For example, in elapsed time an observation with events in months 4, 10 and 14 would have start and stop times of 0–4, 0–10, and 0–14. In gap time, the observation would have start and stop times of 0–4, 0–6, 0–4. The choice of gap *versus* elapsed time depends on the research question at hand. Using elapsed time presumes there are substantive reasons to believe that the 'clock should restart' after each event; such a model is used to determine the effect of covariates on the kth event since the time from the previous event. In contrast, elapsed time models assess the effect of covariates on the kth event since the time from the start of the study. Other variance-corrected alternatives such as the marginal model of Wei *et al.* [13] allow all cases to be at risk for each event so that individuals would be at risk for the first, second, third, etc., event in all periods that the individual is observed. That is, an observation would be at risk for the fourth event before the first event even occurred. Because the previous literature, has largely discredited this model for the study of repeated events, we do not consider it further. See Reference [6] for the hazards and likelihoods for each of the variance-corrected models.

Variance-corrected models present one way to deal with the efficiency problems produced by heterogeneity across individuals. A subset of these models attempts to incorporate event dependence by allowing the baseline hazards to vary with the number of events an individual experiences. However, applied work using alternative variance-corrected estimators often leads to different inferences about estimated effects [3, 9–17] leading to different policy recommendations. These differences are explainable because variance-corrected models do not incorporate the heterogeneity into the estimates themselves and therefore remain biased. Research using simulations to examine the estimates of treatment effects for a subset of the variance-corrected models assuming normally distributed and uniformly distributed random effects, respectively, has found that heterogeneity induces negative biases in the estimates of the treatment effect so that treatment effects are underestimated [5, 6]. The finding is similar to Aalen's findings about the Cox model itself when applied to a repeated events context [7]. Therneau and Grambsch suggested that the magnitude of the bias *may* be tolerable in *some* circumstances, but such a conclusion is unsatisfying. Additional work comparing variance-corrected models in the bivariate case with event dependence as well as heterogeneity demonstrates significant bias in estimated effects, poor size and power of hypothesis tests, and large model mean squared errors with this data generating process [18].

In contrast to the variance-corrected models, frailty or random effects models incorporate heterogeneity into the estimator by making assumptions about the frailty distribution and incorporating it into the model estimates and thus present a more promising alternative than variance-corrected models for dealing with heterogeneity [5, 19]. The underlying logic of frailty models is that some subjects (or groups or clusters) are intrinsically more or less prone to experiencing the event of interest than are others, and that the distribution of these effects can be at least approximated. Frailty models treat repeated events as a special case of more general unit-level heterogeneity. In this case the random effect is across individuals and constant over time, rather than across groups or clusters; there is only a single individual for each value of the random effect, so that it is *shared* over time by a single individual, rather than shared across families or groups. The proportional hazards frailty model for subject i is

written as

$$\lambda_i(t) = \lambda_0(t)\mathrm{e}^{X_i\beta+\omega_i} \tag{1}$$

where X_i is the ith row of the covariate matrix X. X and β correspond to p fixed effects in the model, ω is a vector containing the unknown random effects or frailties [20]. The event times are assumed to be independent conditional on the chosen parametric distribution, so inference may be made in standard fashion. Scholars are actively investigating how to choose a distribution and look at the ramifications of mis-specifying the distribution and determining the problematic conditions [21, 22], while others focus on how to relax the parametric assumption altogether [23–25]. Because the hazards are necessarily positive, the distribution is usually chosen from the class of positive distributions; in applied work, the most widely used are the gamma, Gaussian, and t distributions, with the gamma being by far the most frequent due to the flexibility of that distribution [22, 26].

Like the AG model, the baseline hazard rate for the standard frailty model does not vary by event number, k. However, heterogeneity is directly incorporated via a random effect. Simulations examining the estimates of treatment effects in frailty models with gamma distributed random effects find that frailty models produce unbiased estimates of covariate effects when the variance of the random effect is known [5]. Importantly, these simulations do not consider data generating processes that also contain event dependence. And simulations designed to examine event dependence and heterogeneity are limited to two events and do not include frailty models [18].

Frailty models are better than variance-corrected models for dealing with heterogeneity, but ignore the biasing effects induced by event dependence. Lawless [27] sheds some light on the connection between the stratified analysis, which deal with event dependence, and the frailty model. Specifically, with the semi-parametric Gamma–Poison mixture, with conditional rate $\lambda(t|u = u\lambda(t)$ where $u \sim G(u;\phi)$ with $E(u) = 1$, $\mathrm{var}(u) = \phi$, and $N()$ denotes the number of events and $H()$ denotes the 'history' of the process, the marginal intensity function is given by

$$\lim_{\Delta t\to 0} \frac{P(N(t+\Delta t^-) - N(t^-)|H(t))}{\Delta t} = \lambda(t)\cdot\frac{1+\phi N(t^-)}{1+\phi\Lambda(t)} \tag{2}$$

where $\Lambda(t) = \int_0^t \lambda(u)\,\mathrm{d}u$, which can be seen to 'jump' at event occurrence.

Neither frailty nor variance-corrected models present a general modelling strategy for repeated events processes that are characterized by *both* event dependence and heterogeneity (as our simulations show). Nor do they present a reliable solution when analysts are unsure which features of the data underlie the correlation. Both variance-corrected and frailty models recognize the problems associated with violations of the independence assumption of the Cox model. But as currently proposed, variance-corrected and frailty models are inconsistent with the typical data generating process for repeated events, which feature both event dependence *and* heterogeneity. Nor does the literature investigate how well existing models account for the two different types of correlation. This means that our confidence in inferences and policy prescriptions is low. We cannot reliably estimate the effects of policies or conditions, for example, if unobserved or unmeasured characteristics of individuals or their circumstances affect the risk for multiple heart attacks or if heart attacks themselves increase the risk of future attacks. We seek a modelling solution that is independent of the unknown features of

the data generating process, i.e. that performs well whether event dependence and/or heterogeneity are features of the process of interest. In the next section, we consider the conditional frailty model as one way to meet this goal.

3. THE CONDITIONAL FRAILTY MODEL

The conditional frailty model combines a random effect to incorporate unobserved heterogeneity with event-based stratification (varying baseline hazards) to incorporate event dependence. The model is formulated in gap time so that parameter estimates can be interpreted as a risk estimate for the kth event since the previous event (for right censored failures only).

The hazard or risk of a particular event k occurring for a specific individual i, (λ_{ik}), for the conditional frailty model follows from [5, 6]:

$$\lambda_{ik}(t) = \lambda_{0k}(t - t_{k-1})\mathrm{e}^{X_{ik}\beta+\omega_i} \tag{3}$$

where k denotes event number; λ_{0k} is the baseline hazard rate and varies by event number; $(t - t_{k-1})$ incorporates a gap time data structure so that the hazard gives the risk for event k since the previous event occurred; X is a vector of independent variables, which may be time varying; and β gives the effect parameters. The remaining portion of the hazard incorporates the random effect. Here each subject i has a random effect that is *shared*, i.e. constant, over time (across events) and ω is a vector containing the unknown random effects or frailties [20].

The partial likelihood for this model, conditional on the frailties, follows directly from, among others, [6] and is given by

$$L(\beta) = \prod_{i=1}^{n}\prod_{k=1}^{K}\left(\frac{\mathrm{e}^{X_{ik}\beta+\omega_i}}{\sum_{i=1}^{n}\sum_{k=1}^{K}Y_{ik}\mathrm{e}^{X_{ik}\beta+\omega_i}}\right)^{\delta_{ik}} \tag{4}$$

where k refers to the event number, δ is a censoring variable equal to 1 if observed or 0 if censored, and Y is an at risk indicator, which is equal to 1 when the individual is at risk for the current event k and 0 otherwise.

The gamma frailty model can be laid out in terms of a penalized partial log likelihood solution [5, 232–233]. The penalty is imposed so that the individual effects do not account for too much of the fit. The gamma frailty model is equivalent to a penalized Cox model with penalty function:

$$p(\omega) = (1/\theta)\sum[\omega_i - \exp(\omega_i)] \tag{5}$$

The ω's are distributed as the logs of iid gamma random variables and the tuning parameter θ is their variance. The correlation of subjects within groups—here a subject over time—(Kendall's tau) is $\theta/(2 + \theta)$. The unconditional likelihood is then given by equation (4) multiplied by the relevant penalty function [28, 29]. If the frailties were known, we could, of course, write out the full likelihood. Given that they are unknown, we have two options. We can treat them as missing data and use an EM algorithm with the full likelihood, integrating over an assumed distribution for the ω. In this case, the EM algorithm provides an estimate of the ω. Or we can maximize the penalized version of the log of the usual Cox partial likelihood. Both give the same result for any fixed value of θ (see References [5, 28]). There are advantages for using the penalized version, such as computational speed.

The conditional frailty model allows for the possibility that both heterogeneity and event dependence make important contributions to the hazard rate or an individual's risk for a particular event (re)occurrence. The conditional frailty gap time model has not been explicitly compared to other common recurrent event models and seems particularly promising for addressing both heterogeneity and event dependence because it allows for both an event specific baseline hazard rate and a random effect.

4. SIMULATIONS

We use simulations to compare the conditional frailty model and several variance-corrected and frailty models with a known data generating process that exhibits heterogeneity, event dependence, both, and neither. We focus our comparison on the three most popular and promising variance-corrected models: the Andersen–Gill, conditional gap time, and conditional elapsed time models; and the basic frailty model estimated with a gamma random effect. 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 standard errors, and rate at which the estimated standard errors includes the true parameter. Our simulations suggest that the conditional frailty model can estimate the effects of both sources of correlation simultaneously and retrieve the parameters of the true data generating process better in all four cases. Further, in the simulations we have investigated, the conditional frailty model performs similarly to, or better than, the variance-corrected and frailty alternatives. In the case of both heterogeneity and event dependence, only the conditional frailty model performs well. So, in cases where there is a possibility of both, and often we cannot rule either out, the conditional frailty model is recommended.

4.1. Testing the performance of the conditional frailty model

We begin by testing the performance of the conditional frailty model using simulations that extend those used to evaluate variance-corrected and frailty models in published work. In particular, Therneau and Grambsch [5] and Kelly and Lim [6] have compared some subset of the models under a set of conditions in which a constant treatment is given to a random sample of subjects, unobserved heterogeneity is present, the range of events experienced is from 0 to some small number, and there is neither event dependence nor time dependence [5, 6]. Our initial simulations proceed from this data generating process (DGP), but generalize the DGP further to allow for the addition of event dependent baseline hazard rates. Specifically, we generated data by drawing the time to an individual i's kth event—t_{ik}—from an exponential distribution with rate λ_{ik}:

$$\lambda_{ik}(t) = \lambda_{0k}(t)e^{X_i\beta+u_i} \tag{6}$$

where λ_{0k} is the baseline hazard rate and may vary by k; u_i is a random effect that allows for the introduction of heterogeneity; X_i is a dichotomous and time invariant covariate, indicating for example, whether a subject has received treatment or not; and β is the effect parameter [5, 30]. By drawing from an exponential distribution, we assume that there is no duration dependence, i.e. that time itself has no effect on event rates. This maximizes comparability

with published work. Satten [31] considers the model for interval-censored recurrent event data using a parametric approach.

Heterogeneity enters the DGP in the form of the random effect, u_i. It is set to 0 for the case of no unobserved heterogeneity or is drawn from a normal distribution with $\sigma_u = 0.6$, 1, and 2 across the simulations. The larger variance represents greater heterogeneity and produces higher correlations in event times. Event dependence enters in the form of the baseline hazard, λ_{0k}. If the baseline is constant across events, there is no event dependence, $\lambda_{0k} = \lambda_0$ and we have the case considered in previously published work. Event dependence occurs whenever the baseline hazard is allowed to vary as some function of k. We begin with a simple form of event dependence, setting $\lambda_{0k} = k\lambda_0$. This produces inter-event times that decrease as the number of events experienced grows. One can think of this as drawing from a distribution with baseline hazards that change after each event, producing different and correlated event-specific baseline hazards. In other simulations we consider $\lambda_{0k} = (1/k^2)\lambda_0$, so that event dependence is somewhat weaker.

We initially set a constant treatment effect of β_1 to -1, the baseline hazard to 1.0 and the maximum follow-up time to 4. The maximum number of events, here 6, is determined by the follow-up time and the baseline hazard rate. The set of parameter values we have chosen to create the simulated data produces a distribution of events that follows those of Reference [5] and matches those of many clinical studies, re: small numbers of subjects having large numbers of events, most have few events, etc., over a given typical span of observation (although here with artificial metric). The treatment effect chosen is negative, as treatments efforts typically reduce event rates, however we note the effect of a positive 'treatment,' different baseline hazards, and higher numbers of events in the paper and discuss the findings in the results section.

4.2. *Results*

We generate 1000 data sets following equation (6) and estimate the conditional frailty model and each of the four models identified above: three variance-corrected models (Anderson–Gill, conditional elapsed time, conditional gap time); and a frailty model with a gamma random effect. All the data was simulated and models estimated in R/S+ using the basic survival package in R. We investigate any biases and evaluate inferences to assess model performance. Specifically, we focus our attention on the model estimates of β, the standard errors, and estimates of the random effect for the frailty models. Results are presented in Table I. We consider a broader array of models and DGPs than the previous literature in our comparisons. For example, Kelly and Lim [6] only consider variance-corrected models. We present results here for the case where $\sigma_u = 1$ and event dependence is positive and linear in the baseline hazard. Results for $\sigma_u = 0.6$ reveal the same pattern in results, but with more attenuated differences. Results for $\sigma_u = 2$ produce more dramatic differences for the conditions with heterogeneity and provide stronger evidence for the conditional frailty model. Alternative forms of event dependence support our conclusion that the conditional frailty model is preferred.

4.2.1. Heterogeneity. Heterogeneity in the data means that some people have higher, and others lower, event rates for some unknown or unmeasured reason. This has two implications. First, it means that event prone individuals will dominate the sample of individuals at risk for

Table I. Simulation results for $N = 100$, $\beta = -1.0$, $\theta = 1.0$, and baseline hazard $= 1.0$.

	$\hat{\beta}$	Std dev	Standard error	Coverage rate	$\hat{\theta}$	Rejection rate θ
Panel A: No event dependence, heterogeneity						
Conditional frailty, gap	−0.9856	0.2762	0.2347	0.8990	0.9418	0.9900
Frailty, elapsed	−0.9708	0.2542	0.2259	0.9030	0.8746	1.0000
Andersen–Gill	−0.7491	0.1920	0.1875	0.7410		
Conditional, gap	−0.5520	0.1516	0.1471	0.1550		
Conditional, elapsed	−0.4821	0.1410	0.1353	0.0520		
Panel B: Event dependence, no heterogeneity						
Conditional frailty, gap	−1.0119	0.1007	0.1006	0.9510	0.0056	0.0100
Frailty, elapsed	−1.6867	0.1859	0.1572	0.0080	0.3690	1.0000
Andersen–Gill	−1.2986	0.1458	0.1404	0.4340		
Conditional, gap	−1.0047	0.0983	0.0969	0.9480		
Conditional, elapsed	−1.0120	0.1185	0.1105	0.9370		
Panel C: Event dependence and heterogeneity						
Conditional frailty, gap	−0.9860	0.2502	0.2145	0.9040	0.8309	1.0000
Frailty, elapsed	−1.4081	0.4004	0.2940	0.6500	1.8283	1.0000
Andersen–Gill	−0.7083	0.1747	0.1708	0.5830		
Conditional, gap	−0.5238	0.1351	0.1287	0.0520		
Conditional, elapsed	−0.3948	0.1148	0.1057	0.0010		
Panel D: No event dependence, no heterogeneity						
Conditional frailty, gap	−1.0214	0.1636	0.1540	0.9480	0.0142	0.0150
Frailty, elapsed	−1.0053	0.1424	0.1399	0.9440	0.0167	0.0390
Andersen–Gill	−1.0018	0.1418	0.1362	0.9400		
Conditional, gap	−1.0086	0.1562	0.1486	0.9400		
Conditional, elapsed	−1.0103	0.1590	0.1505	0.9534		

Notes: Reported standard errors for the Andersen–Gill, conditional elapsed time and conditional gap time models are all robust standard errors:

$$\mathbf{V} = \mathbf{I}^{-1}\mathbf{B}\mathbf{I}^{-1} \tag{7}$$

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 [32]. Standard errors for the frailty and conditional frailty models follow Gray [33] $\mathbf{V} = \mathbf{H}^{-1}\mathbf{I}\mathbf{H}^{-1}$. $\mathbf{H}$ is the second derivative matrix for the penalized likelihood.

large numbers of events. When treatment effectively reduces event rates for all individuals, the conditional event rates in the treatment group will become higher as the number of events increases due to the dominance of these people in higher strata. This makes treatment look less effective than it really is, introducing negative bias. In contrast, if treatment increases event rates, i.e. $\beta = 1$, the situation is reversed: the conditional probability of an event in the treatment is smaller than the control arm (there are more cases with smaller numbers of events in the treatment group than in the control group). This means that positive treatment looks less effective, too. Second, in addition to the bias introduced by imbalance in the treatment arms, we expect models that do not directly incorporate heterogeneity in the estimation stage—all the variance-corrected models—will perform particularly poorly. The frailty and conditional frailty models correct for the heterogeneity in the estimation stage and should thus produce better estimates.

Heterogeneity has been the focus of attention in simulations [5, 6]. Table I, panel A, shows the expected downward bias in each of the models, especially in the conditional gap and elapsed models. The statistically significant estimate of the frailty term is more accurate, 0.94, for the conditional frailty model than for the frailty model. We draw four general conclusions from these simulations. First, correcting for standard errors in the presence of heterogeneity is not enough—bias results from the imbalance in the data if the analyst does not control for heterogeneity. Second, controlling for heterogeneity by estimating a random effect enhances model performance substantially, specifically by reducing the bias in the estimated β. Thus both the standard frailty and conditional frailty models give better estimates of the treatment effect than the alternative models. Third, the robust standard errors are smaller than the standard deviation of the estimated $\hat{\beta}$ for all models. Of course, the frailty and conditional frailty model standard deviations of $\hat{\beta}$ are larger than in the remaining models because more parameters are estimated. The standard errors do improve as N increases. Fourth, stratifying is problematic unless a random effect is also estimated—conditional models exhibit large biases as the stratification attempts to pick up the heterogeneity. Because the conditional model controls for the heterogeneity, adding the stratification to the frailty model does not induce bias.

4.2.2. Event dependence. The next two conditions (event dependence as well as event dependence and heterogeneity) constitute extensions to the DGPs examined in the current body of experimental work. The independence assumption of the Cox model is violated as a result of event dependence; the true data generating process has baseline hazard rates that vary by event number. Therefore, we expect models that stratify by event number will necessarily perform better than models that do not. More specifically, by omitting strata, the AG and frailty models must average effects across event numbers. In general, these averages will be biased and increase the more separation there is among the hazard rates across the number of events.

Table I, panel B, shows the expected, large upward bias, which ranges from about 30 per cent for the AG model to over 68 per cent for the frailty model. This bias is explained by imbalance in the treatment and control groups as well, but importantly the patterns produced by event dependence are distinct from those produced by heterogeneity. Specifically, if the treatment reduces event rates, event dependence means that the control group will have more cases having larger numbers of events relative to the treatment group. The end result is that the treatment designed to reduce event rates looks more effective than it is. The same is true for a treatment or covariate with a positive effect. More cases have more events in the treatment group, making the effect look even bigger. Because more events occur in later strata, the effects look bigger unless we correctly incorporate the shift in the baseline hazard. Because the conditional frailty models stratify, this bias does not occur.

The true variance of the random effect is zero in this data generating process. Yet for the frailty models that do not stratify, the random effect necessarily picks up the correlation in events. In contrast, the conditional frailty model (frailty with stratification) accurately reveals the random effect to be zero.

In short, our preliminary results caution against the current recommendation by Therneau and Grambsch [5] for the AG model. When the data exhibit event dependence, stratification is essential. The addition of the random effect in the conditional frailty model, does not appear

to come at any cost. That is, the conditional frailty model correctly estimates the variance of the random effect to be zero while producing a relatively unbiased estimate of treatment effects.

4.2.3. Event dependence and heterogeneity. The condition of joint event dependence and heterogeneity is the most important to consider for two reasons. First, it seems likely that both sources of correlation may simultaneously describe many of the processes that we care about. Second, even if only one source of correlation exists, we typically cannot know which source drives the correlation *a priori* and therefore do not know which model to use [20, 30, 34, 35].

We know that heterogeneity biases effects toward zero due to the imbalance in the number of events that occur in the treatment and control groups. If in addition to heterogeneity, we also have positive event dependence, this problem will be exacerbated. Specifically, even fewer cases experience small numbers of events and more cases experience large numbers of events under event dependence, magnifying imbalance. In particular, if the true treatment is negative, then the control group will have more cases with more events and fewer people with zero (and small numbers of) events. This means that the conditional probability of having an event in the control group becomes less than in the treatment group after a small number of events and thus treatment looks less effective than it is—effects are biased toward zero. The same bias toward zero occurs with positive treatment as well. In this case, the imbalance is reversed and the conditional probability of having another event, given that the subject has had one or two is higher in the control group than in the treatment group, because so few have no or one event. In other words, the probability of treated cases having subsequent events conditional on having had k events is smaller than in the control arm; the control arm has a higher probability of more events than the treatment. This means that the treatment effect will be biased toward zero.

The simulation results suggest that two things occur when we have both event dependence and heterogeneity, and they work to adversely affect the estimates from the conditional models (see Table I, panel C). First, as noted above, event dependence exacerbates the problems with the conditional models that are due to heterogeneity. So while the conditional models clearly outperform the marginal models for event dependence alone, they do not do well with heterogeneity. Second, as expected, in models that do not stratify by event number, the varying baseline hazards are averaged. In general, this averaging produces biased estimates.

The conditional frailty model performs the best for the condition of event dependence and heterogeneity, showing that the effects of both can be simultaneously and distinctly well estimated. The conditional frailty model stratifies by event, controlling for event dependence, and allows for the estimation of a random effect, controlling for the heterogeneity. In doing so, the conditional frailty models have very little bias in either β or θ.

We present the densities of the estimated treatment effects for each of the five models in Figure 1. While the estimates using the conditional frailty model are less tightly clustered around -1.0 then we would like, the mass is centered around the true value. Further, when compared to the alternatives estimators, the majority of the estimates are closer to the true value.

4.2.4. No event dependence or unobserved heterogeneity. We end our simulations by generating data that is consistent with the assumptions of the Cox model; there is no unobserved heterogeneity and no event dependence. In this case, all the estimated models should be

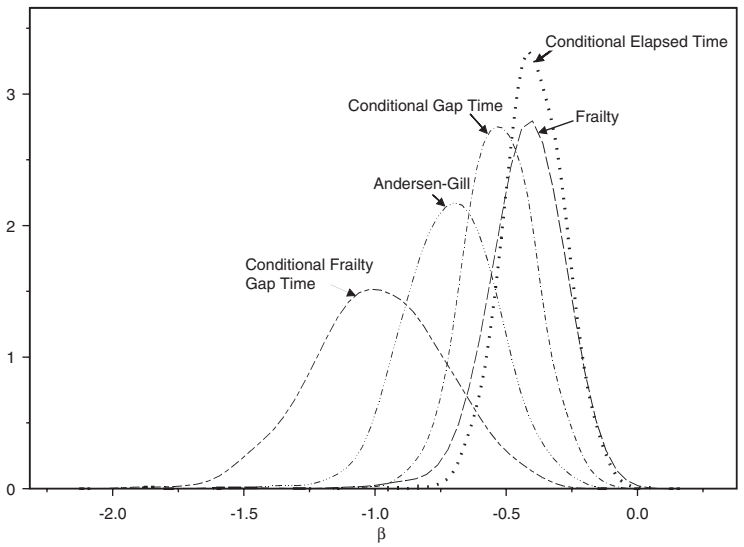

Figure 1. Densities of the estimated β (heterogeneity and event dependence, $N = 100$). The densities of the estimated effects, $\hat{\beta}$, are calculated for conditions under which the $\lambda_0 = 1.0$, $\theta = 1.0$, $\beta = -1$, $\lambda_{0k} = k\lambda_0$, the maximum follow up time is 4. The maximum number of events, which is determined by the baseline hazard and the follow-up time, is 6.

unbiased and the standard errors should be close to the standard deviation. Estimators that allow for varying baseline hazards or that estimate a random effect should, however, be less efficient than estimators that do not, given that they require the estimation of an additional parameter(s). The conditional frailty models, by allowing for both event dependence and heterogeneity should be the least efficient. Table I, panel D, shows that our expectations for these simulations are largely born out in the data. Most importantly, the only cost associated with using the conditional frailty model when there is no correlation in the data is a slight loss of efficiency.

4.2.5. Extensions. In addition to the simulations reported here, we examined the performance of these models setting the baseline hazard to 0.10. This produces times between events that are quite long in the case of no dependence and no heterogeneity so that very few events occur within the sample period. When we add heterogeneity and/or event dependence, the event times and average number of events change, with more events occurring more quickly. These simulations reveal three interesting findings. First, the bias in $\hat{\beta}$ across the estimators is larger than can be explained by the Monte Carlo uncertainty. Second, the estimators reject the null hypothesis on $\hat{\beta}$ at frequencies well below the nominal 95 per cent rate, ranging from 77.9 to 79.5 per cent. Third, the estimated variance of the random effect, θ, is often very biased and the true null is rejected almost 95 per cent of the time for the frailty gamma model. The conditional frailty gamma model estimate of θ has a much larger bias, but the null that $\theta = 0$ is rejected at rates near the nominal 5 per cent level. The former two problems are explained by the small number of cases experiencing multiple events and therefore producing a *rare events bias* and imprecise estimates. In simulations using the larger

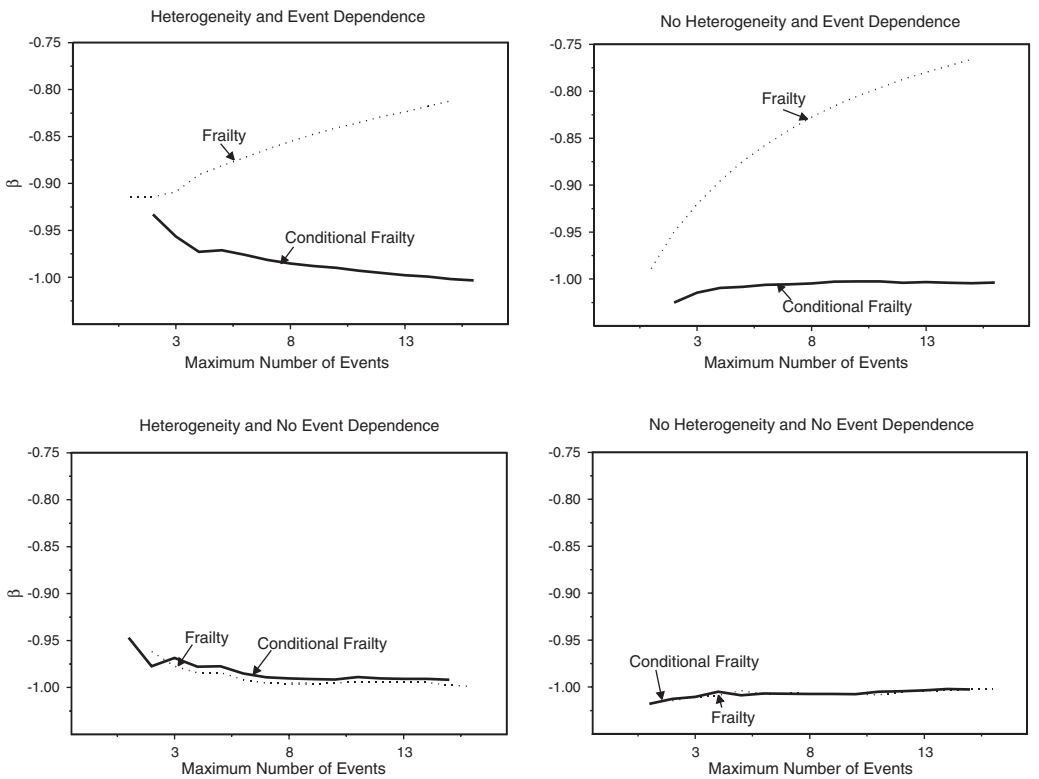

Figure 2. Impact of number of events on estimated β. For all conditions, the baseline hazard $\lambda_0 = 1.0$ and $\beta = -1$. Under conditions of heterogeneity, a random effect was added in which $\theta = 1.0$. Under event dependence, λ_{0k} was set to $k\lambda_0$.

baseline hazard rate, a smaller proportion of the sample experiences no events and these findings disappear.

4.2.6. Gap time versus elapsed time. We have compared the traditional frailty model, which is specified in elapsed time, and the conditional frailty model, which we chose to specify in gap time [6, 18, 36]. It is important to remember that the interpretation of the estimated parameters from a gap time and elapsed time model differ (see References [37, 38]). A natural question arises then about whether the superior performance of the conditional frailty model is due to the formulation of the risk set in gap time rather than stratification also playing a major role under certain conditions. We performed additional simulations to compare a gap time frailty model with the conditional frailty gap time model. The only difference in these simulations is the presence of stratification in the conditional frailty model. The two models perform similarly for two conditions: (a) heterogeneity only and (b) no heterogeneity and no event dependence. However, the existence of event dependence demonstrates the superior performance of the conditional frailty gap time model over a broad range of conditions. Specifically, Figure 2 presents the mean estimated β for each of these two models across a range of maximum numbers of events and a fixed follow-up time. We varied the maximum

number of events from 2 to 12 with a constant follow-up period of 50 time points. As the number of events increase, the conditional frailty model performance increases relative to the gap time frailty model.

5. APPLICATIONS

We present two applications of the conditional frailty model using data from References [5, 39]. The first application examines the effect of interferon gamma on patients with chronic granulomatous disease (cgd), a heterogeneous group of uncommon inherited disorders that manifest in recurrent pyogenic infections. The data are from a double-blind placebo-controlled trial in 128 patients with cgd. In addition to interferon gamma, age at start of the study, pattern of inheritance, and use of corticosteriods are included in the models.

Table II presents estimates for the treatment for the AG, conditional elapsed and conditional gap time, frailty, and conditional frailty gap time models. The largest estimated effect is found in the AG model, followed by the frailty model. The remaining models all estimate a similar and smaller effect. This pattern suggests that infection risk is event dependent, specifically that infection occurrence makes reoccurrence more likely. There is no evidence of heterogeneity in the conditional frailty model (the variance of the random effect is not statistically significant, $p = 0.930$). In contrast, $\hat{\theta}$ is significant at $\alpha < 0.10$, with $p = 0.065$ in the frailty model, but this is not surprising as the estimated random effect will pick up the effects of event dependence as shown in our simulation patterns. The cumulative baseline hazards for the conditional gap time and conditional frailty gap time models present further evidence that only event dependence underlies infection rates. The figures of the cumulative baseline hazards (not shown) are virtually identical, as we would expect when there is no heterogeneity, since both models nest the true data generating process. The separation of the baseline hazards by event number also confirms the existence of event dependence.

The cystic fibrosis example looks at the effects of rhDNase—pulmozyme (DNase I), a cloned and highly purified recombinant DNase I designed to mimic that produced by the human body, deoxyribonuclease—on the incidence of pulmonary exacerbations. The double-blind, placebo-controlled study was conducted in 1992. We estimate all five models for 956 patients with cystic fibrosis. Here we find a different pattern in the estimated treatment effects (see Table III). The conditional frailty model finds the largest effect, followed closely by the frailty model, the two conditional models estimate the smallest effect. We note that the variance of the random effect is statistically significant for both the conditional frailty and frailty model,

Table II. Effects of inteferon gamma in patients with chronic granulotamous disease.

Model	Estimate	Robust standard error	$\hat{\theta}$	p-value
Conditional frailty gap	−0.8997	0.2818	0.0000	0.9300
Frailty, elapsed	−1.0286	0.2643	0.6109	0.0650
Andersen–Gill	−1.0998	0.3093	NA	NA
Conditional gap	−0.8997	0.2978	NA	NA
Conditional elapsed	−0.9047	0.3018	NA	NA

Table III. Effects of DNase I (rhDNase I, pulmozyme) in patients with cystic fibrosis.

Model	Estimate	Robust standard error	$\hat{\theta}$	p-value
Conditional frailty gap	−0.3727	0.1128	2.0713	0.0000
Frailty, elapsed	−0.3331	0.1083	1.1953	0.0000
Andersen–Gill	−0.2951	0.1312	NA	NA
Conditional elapsed	−0.2153	0.1125	NA	NA
Conditional gap	−0.2161	0.1083	NA	NA

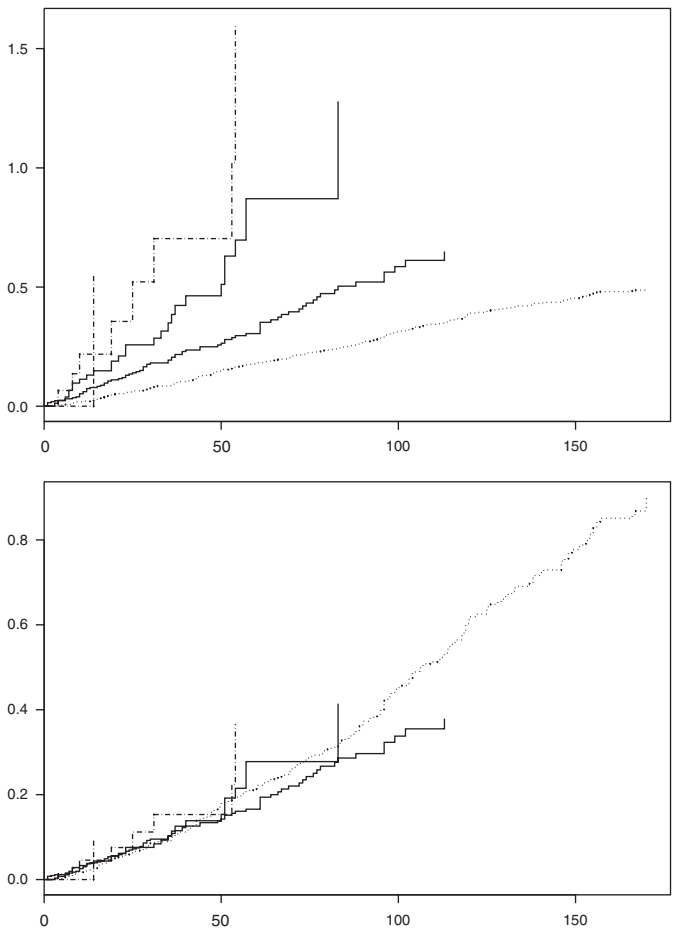

Figure 3. Estimated cumulative hazard, DNase rx and fev. The top panel presents cumulative hazards by event as estimated from the conditional gap time model. The bottom panel presents the same cumulative hazards when the model estimated is the conditional frailty model also in gap time model.

suggesting that heterogeneity is present. The size of the variance of the random effect is estimated at 2.07 and 1.19, respectively. The difference in the estimated effect of treatment in these two models suggest that either heterogeneity and event dependence or only heterogeneity

underlie the true data generating process. Graphs of the cumulative hazards clarify which condition best reflects the data. A graph of the cumulative hazard function by event number from the conditional gap time model (top panel Figure 3) suggests that the baseline hazard varies by event number, however, heterogeneity will masquerade as event dependence when heterogeneity itself is not modelled. Comparing with a graph of the cumulative hazards by event number from the conditional frailty model (bottom panel Figure 3) suggests that there is little or no event dependence in the data—the hazards do not vary significantly by event number once heterogeneity is incorporated into the model. For medical reasons, one might expect event dependence, but the subject enrolment process means that we have no information on the number of pulmonary exacerbations prior to the study. Thus, we are cautious about the generality of the conclusion of no event dependence in other studies of this disease because of the lack of information prior to enrolment.

These results suggest that using the conditional frailty model not only better allows the analyst to capture the effects of both heterogeneity and event dependence, but also helps us to diagnosis the source of correlation in the data. By comparing the magnitude of the estimated effects across the models and then by graphing estimated cumulative baseline hazard rates and comparing patterns across the models, we can triangulate the information to identify the source of correlation.

6. DISCUSSION AND CONCLUSIONS

Our work sheds light on the ability of estimators to capture *both* event dependence and unobserved heterogeneity. Variance-corrected models directly model event dependence through stratification, while attempting to account for heterogeneity with ex post fixes to the standard errors. However, our simulations show that these models fall short given heterogeneity. In contrast, frailty models directly estimate the effects of heterogeneity while ignoring event dependence and thus similarly fall short. Yet repeated events data are likely to exhibit both event dependence and heterogeneity. Certainly it is unlikely we can rule either out *a priori*. Under these conditions, a modelling strategy that is robust to both heterogeneity and event dependence is desirable. We recommend the conditional frailty model.

REFERENCES

1. Aalen OO, Husebye E. Statistical analysis of repeated events forming renewal processes. *Statistics in Medicine* 1991; **10**(8):1227–1240.
2. Lawless JF, Fong DYT. State duration models in clinical and observational studies. *Statistics in Medicine* 1999; **18**(17–18):2365–2376.
3. Cook RJ, Ng ETM, Mukherjee J, Vaughan D. Two-state mixed renewal processes for chronic disease. *Statistics in Medicine* 1999; **18**(2):175–188.
4. Cox DR. Regression models and life tables. *Journal of the Royal Statistical Society* 1972; **B34**(2):86–94.
5. Therneau TM, Grambsch PM. *Modeling Survival Data*: *Extending the Cox Model*. Springer: New York, 2000.
6. Kelly PJ, Lim LL. Survival analysis for recurrent event data: an application to childhood infectious disease. *Statistics in Medicine* 2000; **19**(1):13–33.
7. Aalen OO. Heterogeneity in survival analysis. *Statistics in Medicine* 1988; **7**(11):1121–1137.
8. Lawless JF, Nadeau C. Some simple robust methods for the analysis of recurrent events. *Technometrics* 1995; **37**:158–168.
9. Nelson WB. *Recurrent Events Data Analysis for Product Repairs*, *Disease Recurrences*, *and Other Applications*. ASA-SIAM Series on Statistics and Applied Probability, Philadelphia, PA, 2003.
10. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. *The Annals of Statistics* 1982; **10**(4):1100–1120.

11. Prentice RL, Williams BJ, Peterson AV. On the regression analysis of multivariate failure time data. *Biometrika* 1981; **68**(2):373–379.
12. Gail MH, Santner TJ, Brown CC. An analysis of comparative carcinogenesis experiments based on multiple times to tumor. *Biometrics* 1980; **36**(2):255–266.
13. Wei LJ, Lin DY, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* 1989; **84**(408):1065–1073.
14. Clayton D. Some approaches to the analysis of recurrent event data. *Statistical Methods in Medical Research* 1994; **3**(3):244–262.
15. Lin DY. Cox regression analysis of multivariate failure time data. *Statistics in Medicine* 1994; **13**(21): 2233–2247.
16. Gao S, Zhow XH. An empirical comparison of two semi-parametric approaches for the estimation of covariate effects from multivariate failure time data. *Statistics in Medicine* 1997; **16**(18):2049–2062.
17. Klein JP, Moeschberger ML. *Survival Analysis Techniques for Censored and Truncated Data*. Springer: Telos, 1997.
18. Bowman ME. An evaluation of statistical models for the analysis of recurrent events data: with application to needlestick injuries among a cohort of female veterinarians, Ohio State University, 1996.
19. Oakes DA. Frailty models for multiple event times. In *Survival Analysis*, *State of the Art*, Klein JP, Goel PK (eds). Kluwer Academic Publishers: Dordrecht, 1992.
20. Hougaard P. *Analysis of Multivariate Survival Data*. Springer: New York, 2000.
21. Manatunga A. Use of non-parametric frailty in multivariate survival data. *Presented at the International Conference on Reliability and Survival Analysis*, Columbia, SC, 2003.
22. Kosorok M. Robust inference for proportional hazards univariate frailty regression models. *Presented at the International Conference on Reliability and Survival Analysis*, Columbia, SC, 2003.
23. Ohman-Strickland P. Fitting the frailty distribution using empirical Bayes. *Presented at the International Conference on Reliability and Survival Analysis*, Columbia, SC, 2003.
24. Andersen PK, Klein JP, Zhang M. Testing for centre effects in multi-centre survival studies: a Monte Carlo comparison of fixed and random effects tests. *Statistics in Medicine* 1999; **18**(12):1489–1500.
25. Li Y, Lin X. Functional inference in frailty measurement error models for clustered survival data using the SIMEX approach. *Journal of the American Statistical Association* 2003; **98**(461):191–204.
26. Wei LJ, Glidden DV. An overview of statistical methods for multiple failure time data in clinical trials. *Statistics in Medicine* 1997; **16**(8):883–939.
27. Lawless JF. The analysis of recurrent events for multiple subjects. *Applied Statistics* 1995; **44**(4):487–498.
28. Therneau TM, Grambsch PM, Pankratz VS. *Penalized Survival Models and Frailty*. Mayo Foundation: Rochester, MN, 2000.
29. Fan J, Lin F. Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics* 2002; **30**(1):74–99.
30. Lawless JF. *Statistical Models and Methods for Lifetime Data* (2nd edn). Wiley: New York, 2003.
31. Satten G. Estimating the extent of tracking in interval-censored chain-of-events data. *Biometrics* 1999; **55**(4):1228–1231.
32. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 1980; **48**(4):817–838.
33. Gray GR. Flexible methods for analysing survival data using splines, with application to breast cancer prognosis. *Journal of the American Statistical Association* 1992; **87**(420):942–951.
34. Cook RJ, Lawless JF. Analysis of repeated events. *Statistical Methods in Medical Research* 2002; **11**(2): 141–166.
35. Fong DYT, Lam KF, Lawless JF, Lee YW. Dynamic random effects models for times between repeated events. *Lifetime Data Analysis* 2001; **7**(4):345–362.
36. Box-Steffensmeier JM, Zorn CJW. Duration models for repeated events. *Journal of Politics* 2002; **64**(4): 1069–1094.
37. Hosmer D, Lemeshow S. *Applied Survival Analysis*: *Regression Modelling of Time to Event Data*. Wiley: New York, 1999.
38. Hannan PJ, Shu XO, Weisdorf D, Goldman A. Analysis of failure times for multiple infections following bone marrow transplantation: an application of the multiple failure time proportional Hazards model. *Statistics in Medicine* 1998; **17**:2371–2380.
39. Fleming TR, Harrington DP. *Counting Processes and Survival Analysis*. Wiley: New York, 1991.
40. Therneau TM. Extending the Cox model. *Proceedings of the First Seattle Symposium in Biostatistics*. Springer: New York, 1997.
41. Mahe C, Chevret S. Analysis of recurrent failure time data: should the baseline hazard be stratified? *Statistics in Medicine* 2001; **20**(24):3807–3815.
42. Therneau TM, Hamilton SA. rhDNase as an example of recurrent event analysis. *Statistics in Medicine* 1997; **16**(18):2029–2047.

