Published in final edited form as:

Biometrics. 2018 June ; 74(2): 566–574. doi:10.1111/biom.12768.

Regularity of a renewal process estimated from binary data

John D. Rice*, Robert L. Strawderman, Brent A. Johnson

Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY, USA

Summary:

Assessment of the regularity of a sequence of events over time is important for clinical decision-making as well as informing public health policy. Our motivating example involves determining the effect of an intervention on the regularity of HIV self-testing behavior among high-risk individuals when exact self-testing times are not recorded. Assuming that these unobserved testing times follow a renewal process, the goals of this work are to develop suitable methods for estimating its distributional parameters when only the presence or absence of at least one event per subject in each of several observation windows is recorded. We propose two approaches to estimation and inference: a likelihood-based discrete survival model using only time to first event; and a potentially more efficient quasi-likelihood approach based on the forward recurrence time distribution using all available data. Regularity is quantified and estimated by the coefficient of variation (CV) of the interevent time distribution. Focusing on the gamma renewal process, where the shape parameter of the corresponding interevent time distribution has a monotone relationship with its CV, we conduct simulation studies to evaluate the performance of the proposed methods. We then apply them to our motivating example, concluding that the use of text message reminders significantly improves the regularity of self-testing, but not its frequency. A discussion on interesting directions for further research is provided.

Keywords

Estimating equations; gamma distribution; longitudinal binary data; recurrent events; renewal processes

1. Introduction

Assessment of the regularity of a sequence of events over time is an important public health problem. Cancer screening tests such as mammograms or lower endoscopies are recommended to be performed at certain regular intervals (Davis et al., 2010; Centers for Disease Control and Prevention, 2003), and measuring the extent to which patients follow these recommendations is necessary to evaluate their efficacy and inform public health policy. In our motivating example, a study of HIV self-testing behavior among men who

* john_rice@urmc.rochester.edu.

⁷Supplementary Materials

Appendices referenced in Sections 3–6 are available with this paper at the *Biometrics* website on Wiley Online Library, along with R code for fitting the DS, FR, and CO models.

have sex with men (Khosropour et al., 2013), we are interested in determining the effect of an intervention (online-only versus SMS text messaging follow-up) on regularity of self-testing events. While much attention has been paid to increasing the frequency of HIV testing, especially in high-risk populations (Song et al., 2006; Helms et al., 2009), to our knowledge there has been little research into evaluating or otherwise quantifying the regularity of such testing. Indeed, there is no commonly accepted definition of *regular*, and some studies have conflated frequency with regularity (e.g., Mitchell and Horvath, 2015). This is problematic because events could be irregular (e.g., occurring in clusters) while still meeting some criterion for frequency if there are enough occurrences, regardless of their exact times.

Specifically, there are two characteristics of an irregular sequence of HIV tests that present difficulties from the public health perspective. First, long gaps between tests increase the amount of time that an infected individual will go undetected and therefore increase the chances that he or she could spread the infection to others. Second, tests occurring too closely spaced in time will provide little, if any, additional information due to the length of the window period (i.e., the time between the infection event and the point at which infection is detectable in a diagnostic test; see Song et al., 2006, for details).

It is reasonable to assume that each individual sequence of self-testing times arises from some (possibly unobserved) continuous-time point process within a given study period (Foufoula-Georgiou and Lettenmaier, 1986). In this paper, we assume that subjects are experiencing events according to a renewal process whose interarrival distribution may depend on covariates. The regularity of this process can be directly characterized through the behavior of its interevent time distribution. Specifically, assuming that the mean of the interevent time distribution is finite but nonzero, we use its coefficient of variation (CV) to quantify regularity (Dunn et al., 1983; Wheat and Morrison, 1990; Bawa and Ghosh, 1991). As motivation for this perspective, note that in the most extreme case where the mean is finite but the variance is zero (i.e., $CV = \sqrt{\text{variance}}/\text{mean} = 0$), events will occur equally spaced in time. Such rigid periodicity in the event times corresponds to the most intuitively reasonable conception of a “regular” sequence of events. In contrast, as the CV increases, the standard deviation of the interevent times becomes increasingly large relative to the mean interevent time, corresponding to an interevent time density becoming increasingly flat. This results in far greater variation in the observed interevent times, leading to reduced regularity in the sense described above.

Because the times between events in a renewal process have the same distribution, any renewal process may be thought of as having equally spaced event times *on average* (Gakis and Sivazlian, 1993). The distinction being drawn in this paper and in other work characterizing regularity involves studying the degree of variation about that average. Specifically, fixing the mean of the interevent time distribution, a more regular process is one with an event time distribution that has smaller variance. This conception of regularity is also related to results in Winkelmann (1995), who carefully studied the properties of the gamma renewal process. This process, which contains the homogeneous Poisson process as a special case (i.e., with $CV = 1$), exhibits underdispersion (overdispersion) in the event

count distribution when $CV < 1$ ($CV > 1$). A regular process in the sense described above thus corresponds to a particularly extreme form of underdispersion.

In practice, were exact event times observed, regularity could be modeled and assessed using a specified, possibly covariate-dependent parametric renewal process (Dunn et al., 1983; Wheat and Morrison, 1990; Bawa and Ghosh, 1991). However, in our motivating study, exact self-testing times are not observed; instead, we only observe the presence or absence of any self-testing events in each of a sequence of predefined time intervals, i.e., a binary time series for each subject. Compared to existing methods, this coarse observation scheme and our interest in quantifying the regularity (as opposed, e.g., to the mean function) of the process requires new methods for both estimation and inference.

We view the observed data on each subject as a coarsening of an unobserved renewal process in continuous time, allowing us to estimate the parameters governing this continuous-time model from binary data. By coarsening, we mean two distinct operations: discretization and clipping (Foufoula-Georgiou and Lettenmaier, 1986). The former is due to the grouping of data within disjoint time intervals, and the latter to dichotomizing the count of events within intervals into the categories *zero* or *at least one*. Various existing approaches to estimation cannot be used in the current setting since we observe neither exact event times nor exact counts of events in the specified intervals.

The goal of this paper is to develop suitable methods for estimating the parameters associated with a renewal process when only binary summaries of the process are observed. To facilitate measurement of regularity, we propose two approaches to estimation and inference. First, we propose a likelihood-based method which only uses data on the first interval containing at least one testing event. Second, we propose a quasi-likelihood approach based on the residual life, or forward recurrence time, distribution that uses all of the available data on each subject. Both methods make possible the inclusion of baseline covariates, allowing for heterogeneity in both the frequency and regularity of event times. In the coming sections, we introduce the methodology, conduct simulation studies to evaluate performance, and return to our motivating example. We close the paper with a discussion on future directions.

2. Gamma renewal processes: regularity and the CV

Consider a general ordinary renewal process $N(t) = \max\{n: \sum_{r=1}^n X_r \leq t\}$, where the positive interarrival times $X_1, X_2, \dots$, are independent and identically distributed (iid). Our definition of regularity is based on the CV of the interevent time distribution; this does not require that the interevent times have any particular parametric distribution beyond the basic requirement of having a finite second moment, and the estimation and inference methods to be developed below are general and do not rely on specific distributional assumptions. However, we will focus on the gamma distribution because of its monotone hazard, nesting of the exponential distribution as a special case, prevalence in prior literature on the regularity of events (e.g., Dunn et al., 1983; Wheat and Morrison, 1990; Bawa and Ghosh, 1991), and the ability to naturally parameterize this distribution in terms of its CV.

A gamma renewal process is obtained when the interevent time $X_r > 0$ has a gamma distribution with density $f(x; \alpha, \lambda) = \lambda^\alpha x^{\alpha-1} e^{-\lambda x} / \Gamma(\alpha)$ where $\alpha > 0$ and $\lambda > 0$ respectively denote the shape and rate parameters. Under this parameterization of the gamma distribution, the mean and variance of X_r are respectively $\mu = \alpha/\lambda$ and $\sigma^2 = \alpha/\lambda^2$. This implies that the CV of X_r is $\sigma/\mu = 1/\sqrt{\alpha}$, which clearly decreases as α increases. Dunn et al. (1983), Wheat and Morrison (1990) and Bawa and Ghosh (1991) each use this feature in attempting to quantify regularity. However, this is inherently problematic because (i) the mean and variance of the interevent times both diverge as $\alpha \rightarrow \infty$ and (ii) the moment generating function (mgf) $\mathbb{E}(e^{tX_r}) = (1 - t/\lambda)^{-\alpha}$ converges to zero for $t < \lambda$. This behavior obscures the desired interpretation of the series of events becoming more regular in the sense described earlier and has not been addressed in the aforementioned work on regularity.

The gamma distribution is easily reparameterized in terms of μ and α (e.g., McCullagh and Nelder, 1989, p. 287). Let $f(x; \alpha, \alpha/\mu)$ be the density of a gamma random variable with shape parameter α and rate parameter α/μ . Under this “constant mean” parameterization, the mean interevent time is μ , the variance is μ^2/α and the CV is again $1/\sqrt{\alpha}$. Thus, for a fixed mean μ , the variance of X_r now goes to zero as $\alpha \rightarrow \infty$. This same behavior is directly reflected in the behavior of the mgf $(1 - \mu t/\alpha)^{-\alpha}$, which converges to $e^{\mu t}$ as $\alpha \rightarrow \infty$ (i.e., the distribution of X_r becomes degenerate at μ).

In the constant mean parameterization, the ability to quantify regularity in terms of the shape parameter α remains unchanged, with underdispersion in the gamma renewal process conforming to an intuitively reasonable characterization of “regularity” (i.e., that events occur at times $r\mu$ for $r = 1, 2, \dots$ as $\alpha \rightarrow \infty$). In studying the behavior of a single gamma renewal process, regularity in the sense of approaching the indicated deterministic process clearly increases with α ; in comparing two such processes with respective shape parameters α_1 and α_2 , the process with the larger shape parameter may be considered to be more regular.

3. Estimation with binary time series data

Previous work in the area of recurrent events has focused on analysis of data for which exact event times are available, or when exact event counts are recorded in a set of intervals. Cook and Lawless (2007) provide an extensive treatment of recurrent events, including reviews of the relevant literature. The literature on panel count data, where only counts of events between specified observation times are available (i.e., discretization without clipping), has considered nonparametric (Guédon and Coccozza-Thivent, 2003; Zhang and Jamshidian, 2003), semiparametric (Wellner and Zhang, 2007; Yao et al., 2016), and parametric (Lawless and Zhan, 1998) approaches for estimation and inference.

In the setting of interest, subjects $i = 1, \dots, m$ are followed up to time τ and it is assumed that one can only observe whether or not a subject has experienced at least one event in each of k intervals: $0 \equiv t_0 < t_1 < \dots < t_k = \tau$; that is, the observed response data on subject i is

$$Y_{ij} = \mathbb{1}(\Delta N_{ij} > 0), \quad j = 1, \dots, k, \quad (1)$$

where $N_{ij} = N_i(t_j) - N_i(t_{j-1})$ is the (unobserved) event count in the interval $(t_{j-1}, t_j]$. Existing methods for analyzing panel count data require the availability of the N_{ij} s and are not applicable to data in the form of (1).

Figure 1 shows a small sample of event histories and associated means of the Y_{ij} variables for three values of the shape parameter α when the underlying renewal process is gamma. The leftmost panels depict what we would classify as an irregular process ($\alpha = 0.1$): visible here are the sporadic long gaps between events, alternating with sequences of multiple events in quick succession. The center panels depict events occurring with uniformity ($\alpha = 1$); that is, a homogeneous Poisson process. In this case, there is no duration dependence: the probability of an event occurring in a given interval is the same regardless of when any previous events occurred. The rightmost panels depict events occurring with regularity ($\alpha = 20$), where the spacing between events is much more consistent across and within subjects than for $\alpha = 0.1$ or $\alpha = 1$.

We assume in this paper that we only have access to coarsened data, in the form of equation (1), derived from a sample of ordinary gamma renewal processes $\{N_i(t), 0 \leq t \leq \tau, i = 1, \dots, m\}$. For an ordinary renewal process, all subjects begin the observation period at the time of a renewal. To allow for the possibility that regularity and mean interevent time could vary between subjects, we parameterize α_i and μ_i such that the effects of covariates are represented by finite-dimensional parameters γ and β , respectively. Specifically, suppose that for each subject we have two covariate vectors, $\mathbf{x}_i$ of dimension p , acting on the mean; and $\mathbf{z}_i$ of dimension q , acting on the shape parameter. These vectors may or may not contain the same set of covariates: that is, elements of $\mathbf{x}_i$ may or may not be shared by $\mathbf{z}_i$. Under this model, the interevent times for the i th subject are distributed as gamma with mean μ_i and variance μ_i^2 / α_i , where $\alpha_i = e^{\mathbf{z}_i' \gamma}$ and $\mu_i = e^{\mathbf{x}_i' \beta}$ and use of the log link preserves positivity. The following subsections provide two possible methods for estimating $\theta = (\gamma', \beta')'$.

3.1 Likelihood-based estimation in a discrete-time survival model

For any set of ordinary renewal processes $\{N_i(t), t \geq 0\}, i = 1, \dots, m$ we have

$$P\{N_i(t) = 0\} = 1 - F_i(t), \quad (2)$$

where $F_i(\cdot)$ denotes the cdf of the interarrival times for the i th process (Ross, 1996, p. 99), allowed here to depend on subject-specific covariates. Let $T_i = \min\{j: Y_{ij} = 1\}$ if $\sum_{j=1}^k Y_{ij} > 0$ and $k + 1$ otherwise, where $Y_{ij} = \mathbb{1}(\Delta N_{ij} > 0)$. As written, T_i represents the time to the first interval with at least one event for the i th subject, with $T_i = k + 1$ indicating that no events were observed to occur during follow-up. Using (2), the distribution of T_i is then given by $P(T_i = j) = F_i(t_j) - F_i(t_{j-1}), 1 \leq j \leq k+1$, where it is assumed that $t_{k+1} = \infty$ and $F_i(\infty) = 1, i = 1, \dots, m$.

Guttorp (1986) makes use of (2) to obtain method-of-moments estimators for the parameters of a Weibull interarrival distribution. In the current context, the above construction can instead be used to generate a likelihood for the “time-to-first-event-interval” data. Specifically, assume that each $N_i(\cdot)$ is a gamma renewal process with an interevent time

distribution that follows the “constant mean” parameterization of Section 2. Then, $F_i(t) = F(t; \alpha_i, \alpha_i/\mu_i)$, $i = 1, \dots, m$ and the observed data log-likelihood is simply

$$\log L_m(\theta) = \sum_{i=1}^m \sum_{j=1}^{k+1} \mathbb{1}(T_i = j) \log \{ F(t_j; \alpha_i, \alpha_i/\mu_i) - F(t_{j-1}; \alpha_i, \alpha_i/\mu_i) \}. \quad (3)$$

Here, $\theta = (\gamma', \beta')'$ is the set of parameters describing (α_i, μ_i) for $i = 1, \dots, m$. As equation (3) defines a proper likelihood for a discrete-time survival model, estimation and inference are conducted in the usual manner (Kalbfleisch and Prentice, 2002, Section 2.4).

3.2 Quasi-likelihood estimation using the residual life distribution

The likelihood-based analysis of the previous section yields efficient estimators if one is only able to follow subjects until the interval of their first event. With data of the form (1) on m subjects, the log-likelihood (3) discards information on $Y_{i,T_i+1}, \dots, Y_{ik}$, $i = 1, \dots, m$ thereby making inefficient use of the observed data. We now introduce a method that utilizes all of the data assumed to be available. In what follows, we define $\bar{H}(\cdot) = 1 - H(\cdot)$ for any function $H(\cdot)$. In addition, for any ordinary renewal process $N(\cdot)$, we let $R(t)$ denote the residual life, or forward recurrence time, at time t . Importantly, it then follows that $P\{R(t) > x\} = \bar{F}(t+x) + \int_0^t \bar{F}(t+x-y) dM(y)$, where $F(\cdot)$ is the interarrival distribution and $M(t) = \mathbb{E}N(t)$ (Ross, 1996, p. 132).

The development of the quasi-likelihood estimator relies directly on this last identity. To see this, observe that $Y_{ij} = 1$ (i.e., $N_{ij} > 0$) implies that there is at least one renewal in $(t_{j-1}, t_j]$; hence, we may write $\mathbb{E}Y_{ij} = P(\Delta N_{ij} > 0) = P\{R_i(t_{j-1}) \leq t_j - t_{j-1}\}$ where $R_i(t)$ is the forward recurrence time at time t for subject i . It now follows that

$G_i(t_j) \equiv \mathbb{E}Y_{ij} = F_i(t_j) - \int_0^{t_j-1} \bar{F}_i(t_j-y) dM_i(y)$, where $F_i(\cdot)$ is the same distribution function appearing in (2) because all subjects begin follow-up with a renewal at time 0 (i.e., all interevent times, including the time to the first event, have the same distribution).

In general, $G_i(t_j) = P\{R_i(t_{j-1}) \leq t_j - t_{j-1}\}$ depends on both t_j and t_{j-1} ; however, in the case where $t_j - t_{j-1} = t_1$ for $1 \leq j \leq k$ (i.e., observation times are equally spaced over the study period),

$$G_i(t_j) = F_i(t_{j-1} + t_1) - \int_0^{t_j-1} \bar{F}_i(t_{j-1} + t_1 - y) dM_i(y) \quad (4)$$

depends only on t_{j-1} . Making the same distributional assumptions as in Section 3.1, it follows that $G(t; \alpha_i, \alpha_i/\mu_i) = G_i(t)$ for $i = 1, \dots, m$ and one can now construct the quasi-likelihood (McCullagh and Nelder, 1989, Section 9.3)

$$Q_m(\theta) = \sum_{i=1}^m \sum_{j=1}^k \{ Y_{ij} \log G(t_j; \alpha_i, \alpha_i/\mu_i) + (1 - Y_{ij}) \log \bar{G}(t_j; \alpha_i, \alpha_i/\mu_i) \}. \quad (5)$$

The parameter θ being estimated under (5) is the same as that under (3). The objective function (5) arises from a model for an outcome $Y_{ij} \in \{0, 1\}$ for which we specify only its

first moment $\mathbb{E}Y_{ij} = G(t_j; \alpha_i, \alpha_i / \mu_i)$ and the mean-variance relationship $\text{Var } Y_{ij} = \mathbb{E}Y_{ij}(1 - \mathbb{E}Y_{ij})$. Maximization of (5) with respect to $\boldsymbol{\theta}$ is equivalent to solving

$$\mathbf{U}_m(\boldsymbol{\theta}) = \sum_{i=1}^m \sum_{j=1}^k \frac{Y_{ij} - G(t_j; \alpha_i, \alpha_i / \mu_i)}{G(t_j; \alpha_i, \alpha_i / \mu_i) \bar{G}(t_j; \alpha_i, \alpha_i / \mu_i)} \frac{\partial G(t_j; \alpha_i, \alpha_i / \mu_i)}{\partial \boldsymbol{\theta}} = \mathbf{0}. \quad (6)$$

Inference for the resulting parameter estimates obtained under this estimating equation may be carried out using the sandwich covariance estimator

$$\mathbf{J}_m(\hat{\boldsymbol{\theta}})^{-1} \sum_{i=1}^m \left\{ \mathbf{S}_i(\hat{\boldsymbol{\theta}})' (\mathbf{Y}_i - \bar{\mathbf{Y}}_m) (\mathbf{Y}_i - \bar{\mathbf{Y}}_m)' \mathbf{S}_i(\hat{\boldsymbol{\theta}}) \right\} \mathbf{J}_m(\hat{\boldsymbol{\theta}})^{-1}, \quad (7)$$

where $\mathbf{J}_m(\boldsymbol{\theta}) = -2 Q_m' \boldsymbol{\theta} \boldsymbol{\theta}'$, $\bar{\mathbf{Y}}_m = m^{-1} \sum_{i=1}^m \mathbf{Y}_i$ and $\mathbf{S}_i(\boldsymbol{\theta})$ is a $(p+q) \times k$ matrix with j th column $\{G(t_j; \alpha_i, \alpha_i / \mu_i) \bar{G}(t_j; \alpha_i, \alpha_i / \mu_i)\}^{-1} \partial G(t_j; \alpha_i, \alpha_i / \mu_i) / \partial \boldsymbol{\theta}$. (See Appendix B of the Supplementary Materials for a derivation of the asymptotic distribution of these estimators.)

With only minor changes in notation, both (3) and (5) are easily adapted for use with other interevent time distributions. Because (5) uses all of the observed data, it is reasonable to expect that estimating $\boldsymbol{\theta}$ via (5) will result in greater efficiency in comparison to estimating $\boldsymbol{\theta}$ via (3). However, (5) is not a true likelihood; hence, this anticipated efficiency improvement is not guaranteed. In comparison to (3), we further note that the calculation of (5) is more challenging because $G(\cdot)$ in (4) typically does not have a closed form. In our simulations and data analyses, we calculate $G(\cdot)$ using a discrete approximation due to Arnold and Groeneveld (1981). This approximation relies on (i) a formula for the residual life derived from discrete-time renewal theory; and, (ii) the construction of a discrete random variable U_L such that $P(U_L/L \geq t) \rightarrow R(t)$ as $L \rightarrow \infty$, where $R(\cdot)$ is the interevent time distribution; see Appendix A of the Supplementary Materials for some additional details of this approximation and its accuracy. We found that $L = 200$ gave adequate accuracy for our purposes.

4. Simulation studies

To evaluate the performance of the proposed methods in finite samples, we conducted simulation studies comparing the estimators respectively derived under (3) and (5). We abbreviate estimates derived from the forward recurrence quasi-likelihood (5) as FR, while those derived from the discrete survival likelihood (3) are abbreviated as DS. We also estimated parameters from the underlying continuous data, which is assumed to be unavailable for analysis, using a point process likelihood based on the continuously observed event times in order to understand the impact of using coarsened data on bias and variance (see Appendix C of the Supplementary Materials for details); these estimates are abbreviated as CO.

Two scenarios were considered in which the underlying point process is a gamma renewal process having shape and scale parameters of the form

$$\alpha_i = e^{\gamma_0 + \gamma_1 Z_i}, \mu_i = e^{\beta_0 + \beta_1 Z_i}, \quad (8)$$

where Z_i denotes a scalar covariate; here, the desired parameter $\theta = (\gamma_0, \gamma_1, \beta_0, \beta_1)'$. First, to gain some understanding of the ability of the methods to distinguish between clearly defined high-, low-, and Poisson-regularity groups, we generated data assuming Z_i is a binary covariate. Second, to assess the performance of the methods in a more general regression setting, we generated data assuming that Z_i is a continuous covariate. Simulation settings and results for each scenario are discussed in this section; more details appear in Appendix C of the Supplementary Materials.

Implementation of all methods was by numerical maximization of the objective functions in R using `optim()`. Standard error calculations were based on numerical differentiation of objective and score functions using the `hessian=TRUE` option in `optim()` and the `numDeriv` package (Gilbert and Varadhan, 2015).

4.1 Binary covariate

The purpose of the simulations in the first scenario was to examine the effect of different levels of regularity in two groups on the ability of the proposed methods to detect these differences. The sample size was $m = 500$. We held the length of follow-up $\tau = 12$ and the number of observation intervals $k = 6$ constant for all simulations. Data were generated according to model (8) in which Z_i is a Bernoulli random variable with mean 0.5, intended to mimic the situation with a control and a treatment group with equal assignment probabilities. This implies that only two values are possible for α_i and μ_i . For each group (treatment and control), three values of the shape parameter a_i were considered: $e^{-1.5}$, e^0 , and $e^{1.5}$; the correspondence between α_i for each treatment group and γ_0, γ_1 is detailed in Appendix C. The regression parameters β_0, β_1 were determined by holding $\mathbb{E}N(\tau)$ fixed at either 3 or 9 and solving numerically for β_0, β_1 in the renewal equations for the gamma renewal process under the constant mean parameterization with the log link (see Appendix C for details). Therefore, the true values of β_0, β_1 vary depending on α_i and $\mathbb{E}N(\tau)$. Note, however, that $\mathbb{E}N(\tau)$ does not vary between treatment groups.

Detailed results for this scenario appear in Appendix C of the Supplementary Materials. It is clear from these results that the forward recurrence method is able to distinguish reliably between two groups on the basis of regularity when only binary summaries of the renewal process are available, regardless of the levels of regularity involved and the rate of events. The same cannot be said for the discrete survival method, which suffers substantial bias when the event rate is high.

4.2 Continuous covariate

For the second scenario, data were generated according to model (8), where $\gamma_0 = -1$, $\gamma_1 = 1.5$, $\beta_0 = \beta_1 = 0.5$ and the covariate was $Z_i \sim \mathcal{N}(0.5, 0.5^2)$. This choice matches the first two moments of a Gaussian distribution with those of a Bernoulli with mean 0.5 (as in our binary covariate setting). These simulations are intended to show the ability of the methods to go beyond simply fitting separate models to two or more groups.

We varied the follow-up time τ and the number of intervals k to assess the impact of these parameters, which relate to study design, on the behavior of the estimators; the sample size was $m = 500$. Because of our focus in this paper on regularity, we present results only for the estimates of γ_0, γ_1 here in Tables 1 and 2; further results, including tables for the β_0, β_1 estimates, appear in Appendix C of the Supplementary Materials.

As expected, the CO method offers the best performance because all event times are exactly observed. Its performance is unaffected by the value of k as the number of observation intervals is only relevant for the coarsened data; the slight variation seen in the tables is a consequence of the independent sequences of data sets generated for each combination of τ and k . For the FR and DS methods, there are also some interesting general efficiency trends that can be observed. Most importantly, the results show that FR has either the same or substantially better efficiency in all cases. For both methods, the variance of $\hat{\gamma}_0$ decreases with increasing k for a fixed τ , whereas the variance of $\hat{\beta}_0$ shows little change except for larger values of τ where some decreases in variance are apparent; in contrast, the variance of $\hat{\beta}_0$ decreases with increasing τ for a fixed k and the variance of $\hat{\gamma}_0$ shows little change. In addition, for both methods, the variance of $\hat{\gamma}_1$ decreases with increasing k for a fixed τ and also with increasing τ for a fixed k ; the variance of $\hat{\beta}_1$ decreases slightly with increasing k for a fixed τ and decreases more sharply when τ increases for a fixed k . For the most part, these results are expected, as the overall amount of available information tends to grow with both k and τ . Finally, in comparison with CO, the FR method is able to narrow the gap in MSE considerably for both $\hat{\beta}_0$ and $\hat{\beta}_1$ in settings involving larger k and smaller τ . However, as might be expected, neither DS nor FR are able to estimate the shape-related parameters with precision comparable to CO.

In general, the bias of the FR estimators is slightly greater than that of the DS estimators: plots of the average value of the quasi-score functions (6) for the simulations (not shown) reveal that the discrete approximation to the forward recurrence time distribution introduces some modest bias into the (theoretically unbiased) estimating equations that decreases as the number of partitions L increases. However, this bias is more than compensated for by the reduced variability of the FR method, resulting in consistently better MSE performance for the forward recurrence method relative to the discrete survival method.

The relative efficiency of FR versus DS is governed by two factors. The FR method is a quasi-likelihood method that assumes a working independence correlation structure; it uses all of the observed Y_{ij} data but is not based on a full likelihood. The DS method uses a full likelihood, but is only informed by data from intervals up to and including that containing the first event to estimate model parameters. The improved efficiency of the FR method relative to DS can evidently be attributed to increased data utilization, despite the use of an otherwise inefficient estimator.

5. Analysis of *Checking In!* study

5.1 Description of the data

We apply the proposed methods to data from the *Checking In!* study (Khosropour et al., 2013). This study randomized eligible men who have sex with men (MSM) to either receive text message (SMS) or online follow-up, and collected data at two-month intervals for a period of twelve months. Variables on which data was collected included baseline demographic variables age, education level, and race. The coarseness of the HIV testing data follows directly from the longitudinal study design and the primary retention endpoint of the *Checking In!* study. In our notation, Y_{ij} is the variable indicating whether the i th subject has self-tested for HIV at least once during the j th interval, $j = 1, \dots, 6$. Neither the date of the HIV test nor number of tests per two-month interval was collected.

Since self-testing behavior was self-reported, missing outcome data was assumed to indicate that the subject had not self-tested during that interval. This allows for more conservative conclusions with respect to the rates of self-testing in the study population: the rates we find are a lower bound on the true rates of testing had we been able to observe the missing intervals. The validity of the renewal process model requires that the observation period for each subject begin at a self-testing event, which is fulfilled for this data set as only those subjects who tested negative upon return of an initial HIV self-test kit were followed prospectively (Khosropour et al., 2013).

A sample of size $m = 645$ subjects was available for the analysis, divided approximately equally between the Online and SMS treatment groups. The scientific question of interest in our analysis is whether or not the intervention (SMS versus online follow-up) has any effect on regularity of self-testing for HIV. To evaluate this, we fit regression models by maximizing (3) and (5) where we assume that the time between self testing events follows a gamma distribution with $\alpha_i = e^{\gamma_0 + \gamma_1 \text{trt}_i}$, $\mu_i = e^{\beta_0 + \beta_1 \text{trt}_i}$, where trt_i is equal to 1 if the i th subject is randomized to the SMS treatment group and 0 otherwise. Models adjusting both shape and mean for age, education, and race were also fit, but none of these covariates were found to be significant (see Appendix D of the Supplementary Materials). Therefore, the final analysis presented here includes only the treatment variable.

5.2 Results of gamma renewal analysis

The estimates from both the discrete survival and the forward recurrence methods are presented in Tables 3 and 4; the former displays the estimates of $\gamma_0, \gamma_1, \beta_0, \beta_1$, while the latter gives the corresponding estimated shape and mean parameters by treatment group. The Online group is self-testing in a significantly irregular manner according to both methods ($p < 0.0001$). However, there is no evidence that the self-testing process for the SMS group is either more or less regular than Poisson, when judged by the estimates obtained using the discrete survival method ($p = 0.4149$). The forward recurrence method, by contrast, allows us to conclude that the SMS group is testing significantly more regularly than a Poisson process ($p = 0.0489$). Both the discrete survival and forward recurrence models lead to the conclusion that there is significant evidence of a difference in regularity between the two groups, as $p < 0.0001$ for the Wald test of $\gamma_1 = 0$ in each case.

Under the DS method, there is no evidence of a difference in mean time between self-testing events for the two groups ($p = 0.1136$ for the test of $\beta_1 = 0$). The mean time between self-tests for both groups is considerable, at 10.1 months for the online group and 17.4 months for the SMS group. We should interpret these results with caution, however, as these values are either close to or beyond the time horizon for the entire study (12 months). The estimates of the mean interarrival time obtained under the FR method deviate substantially from those of the discrete survival model. The mean time between self-tests is estimated to be lower for both groups, and the test for a difference in group means is significant ($p = 0.0048$).

Differences in parameter estimates obtained between the two approaches can arise in many ways, including as a result of model misspecification. Goodness of fit for the estimated models was therefore examined by comparing plots of the forward recurrence distribution given in equation (4) corresponding to the estimated shape and mean for each group and both the DS and FR methods with the empirical values of $m^{-1} \sum_{i=1}^m Y_{ij}$, shown in Figure 2. There appears to be good agreement for the Online group between the fitted gamma distributions and the observed proportions of subjects in each treatment group self-testing at least once, regardless of which method is used. The fit of both models to the SMS group data is notably worse, although the forward recurrence method is much closer to the observed data visually: the forward recurrence predictions are able to reproduce the slight upward trend in the observed data, while the discrete survival predictions are close to flat (but slightly downward sloping on closer examination).

These conclusions are substantiated by the root mean-square errors given in the figure, with the forward recurrence method's predictions giving approximately half the RMSE of the discrete survival method in both the Online and SMS groups. This is consistent with the results of our simulation studies, where the MSE of the parameter estimates in models of the same dimension and similar sample size showed substantial improvements with the forward recurrence method as compared with the discrete survival method (see Appendix C in the Supplementary Materials for additional simulation results).

In order to put some of these results in context, we also provide a comparison with the estimates of the inter-test time distribution found by Song et al. (2006) in a study population of MSM visiting a publicly funded clinic in Seattle for HIV testing between 1996 and 2000. This appears in Appendix D of the Supplementary Materials.

6. Discussion

In this paper we have proposed two approaches to estimating the rate and regularity of a renewal process from coarsely measured data. We have shown through simulation studies the substantial gains that may be made with our proposed quasi-likelihood-based method over a likelihood-based method that uses only information available up to the first interval in which a testing event occurs. We applied the method to data on HIV self-testing behavior among men who have sex with men, concluding that the use of SMS reminders significantly improves regularity of self-testing while decreasing its frequency.

We are inherently limited in our analysis of binary data collected at fixed intervals when it is thought to be generated by an underlying stochastic process in continuous time. Our results show these limitations may be mitigated by making certain further assumptions about that underlying process. However, note that estimability problems may arise in extreme circumstances for both the DS and FR methods. This may best be understood with reference to the ratio of μ_i to t_1 (assuming all observation intervals are equally spaced): as $\mu_i/t_1 \rightarrow 0$, the observed Y_{ij} will equal 1 with probability approaching 1; as $\mu_i/t_1 \rightarrow \infty$, the reverse happens, with the observed Y_{ij} equaling 0 with probability approaching 1. In both cases, homogeneous response profiles become increasingly likely, which will in turn lead to difficulties in estimating the parameters associated with the renewal process if a sufficient proportion of subjects' μ_i values tend to one of these two limiting cases. However, the model parameters remain identifiable, in the sense that unique values of μ_i and α_i correspond to unique distributions of $(Y_{i1}, \dots, Y_{ik})$, except in the most extreme cases (i.e., $\mu_i = 0$ or $\mu_i = \infty$).

We have focused on the gamma renewal process because of its familiarity in the literature on regularity of events, but in any given application, it may not provide the best possible fit to the data. Further work is needed to determine how much we may weaken distributional assumptions while still being able to answer questions about regularity. Extensions to allow for varying risk levels over time (e.g., using the modulated gamma renewal process of Berman, 1981) are also possible in principle, at least for the DS method, but would be difficult to implement in our coarse-data setting due to greatly increased model complexity. We discuss such possibilities in more detail in Appendix E of the Supplementary Materials.

Although we hypothesize that increasing regularity of testing will lead to a decrease in HIV incidence, there is no data available that explicitly supports this hypothesis. Because the incidence of HIV is quite low in the *Checking In!* study population, the possibility of stopping a subject's testing process due to a positive result was not addressed in our analysis. This represents an important direction for future work: how to determine the effect of regularity of testing on the occurrence of a low-probability event, i.e., HIV infection. Song et al. (2006) present an analysis in which they seek to understand the effect of testing behavior on HIV infection, concluding that while there is no evidence that choosing to test is associated with recent HIV infection, there is a significantly higher rate of infection among subjects who tested frequently. This latter fact could imply that subjects who perceive themselves to be at elevated risk of infection test more frequently because of that perception. An analysis of the regularity of testing could illuminate whether those subjects testing more frequently were also testing more regularly, and determine which of the two factors, frequency or regularity, was more highly associated with risk of HIV infection.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank Patrick Sullivan for providing access to the *Checking In!* data set. Additionally, we acknowledge the partial support from the University of Rochester CTSA award UL1TR000042 from the National Center for

Advancing Translational Sciences of the NIH and from the University of Rochester Center for AIDS Research grant P30AI078498 (NIH/NIAID) and the University of Rochester School of Medicine and Dentistry.

REFERENCES

- Arnold BC and Groeneveld RA (1981). On excess life in certain renewal processes. *Journal of Applied Probability* 18, 378–389.
- Bawa K and Ghosh A (1991). The covariates of regularity in purchase timing. *Marketing Letters* 2, 147–157.
- Berman M (1981). Inhomogeneous and modulated gamma processes. *Biometrika* 68, 143–152.
- Centers for Disease Control and Prevention (2003). Colorectal cancer test use among persons aged $\geq$ 50 years—United States, 2001. *Morbidity and Mortality Weekly Report* 52, 193–196. [PubMed: 12653456]
- Cook RJ and Lawless JF (2007). *The Statistical Analysis of Recurrent Events*. Springer.
- Davis WW, Parsons VL, Xie D, Schenker N, Town M, Raghunathan TE, and Feuer EJ (2010). State-based estimates of mammography screening rates based on information from two health surveys. *Public Health Reports* 125, 567–578. [PubMed: 20597457]
- Dunn R, Reader S, and Wrigley N (1983). An investigation of the assumptions of the NBD model as applied to purchasing at individual stores. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 32, 249–259.
- Foufoula-Georgiou E and Lettenmaier DP (1986). Continuous-time versus discrete-time point process models for rainfall occurrence series. *Water Resources Research* 22, 531–542.
- Gakis K and Sivazlian B (1993). Distribution of order statistics of waiting times in an ordinary renewal process and the covariance of the renewal increments. *Stochastic Analysis and Applications* 11, 441–458.
- Gilbert P and Varadhan R (2015). numDeriv: Accurate Numerical Derivatives. R package version 2014.2-1.
- Guédon Y and Coccozza-Thivent C (2003). Nonparametric estimation of renewal processes from count data. *The Canadian Journal of Statistics* 31, 191–223.
- Guttorp P (1986). On binary time series obtained from continuous time point processes describing rainfall. *Water Resources Research* 22, 897–904.
- Helms DJ, Weinstock HS, Mahle KC, Bernstein KT, Furness BW, Kent CK, Rietmeijer CA, Shahkolahi AM, Hughes JP, and Golden MR (2009). HIV testing frequency among men who have sex with men attending sexually transmitted disease clinics: implications for HIV prevention and surveillance. *Journal of Acquired Immune Deficiency Syndromes* 50, 320–326. [PubMed: 19194309]
- Kalbfleisch JD and Prentice RL (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, second edition.
- Khosropour CM, Johnson BA, Ricca AV, and Sullivan PS (2013). Enhancing retention of an Internet-based cohort study of men who have sex with men (MSM) via text messaging: Randomized controlled trial. *Journal of Medical Internet Research* 15, e194. [PubMed: 23981905]
- Lawless JF and Zhan M (1998). Analysis of interval-grouped recurrent-event data using piecewise constant rate functions. *The Canadian Journal of Statistics* 26, 549–565.
- McCullagh P and Nelder JA (1989). *Generalized Linear Models*. Chapman & Hall, second edition.
- Mitchell JW and Horvath KJ (2015). Factors associated with regular HIV testing among a sample of US MSM with HIV-negative main partners. *Journal of Acquired Immune Deficiency Syndromes* 64, 417–423.
- Ross SM (1996). *Stochastic Processes*. John Wiley & Sons, second edition.
- Song R, Karon JM, White E, and Goldbaum G (2006). Estimating the distribution of a renewal process from times at which events from an independent process are detected. *Biometrics* 62, 838–846. [PubMed: 16984327]
- Wellner JA and Zhang Y (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *The Annals of Statistics* 35, 2106–2142.

- Wheat RD and Morrison DG (1990). Estimating purchase regularity with two interpurchase times. *Journal of Marketing Research* 27, 87–93.
- Winkelmann R (1995). Duration dependence and dispersion in count-data models. *Journal of Business & Economic Statistics* 13, 467–474.
- Yao B, Wang L, and He X (2016). Semiparametric regression analysis of panel count data allowing for within-subject correlation. *Computational Statistics and Data Analysis* 97, 47–59.
- Zhang Y and Jamshidian M (2003). The gamma-frailty Poisson model for the nonparametric estimation of panel count data. *Biometrics* 59, 1099–1106. [PubMed: 14969490]

Figure 1.

Sample event histories: randomly generated gamma renewal processes for $m = 1000$ subjects. The mean time between events in these simulated data sets was set to be $\mu = 3$ months. The top row displays a random subset of 10 of the processes in continuous time, while the bottom row displays a summary of the data we would be able to observe in our setting. Specifically, each panel of the bottom row is a barplot of the observed mean of $Y_{ij}, j = 1, \dots, k = 6$.

Figure 2.

Fitted and observed values for the HIV self-testing data. The filled circles represent the empirical proportions of subjects with at least one self-testing event in that interval for each treatment group; the lines are 95% confidence limits for these proportions. The accuracy of the predicted proportions for the discrete survival (DS) and forward recurrence method (FR) as measured by root mean-square error is given in the legend in parentheses. Note the different scale of the y-axis between the panels.

Table 1

Simulation results for estimating the gamma shape intercept parameter $\gamma_0 = -1$: Continuous covariate $Z \sim \mathcal{N}(0.5, 0.5^2)$, $m = 500$. Reported here are the bias, average standard error (ASE), empirical standard deviation (ESD), empirical coverage probabilities for 95% confidence intervals (ECP), and mean-square error (MSE) of the estimates of γ_0 . The estimation methods are the forward recurrence quasi-likelihood (FR) and discrete survival likelihood (DS), both of which make use of binary data which we assume is available; the continuously observed point process likelihood (CO) method represents the ideal case (i.e., the efficient estimator, if one had access to the exact event times).

τ	k	Method	Bias	ASE	ESD	ECP	MSE
2	4	FR	-0.011	0.144	0.147	0.948	0.022
		CO	0.003	0.046	0.045	0.956	0.002
		DS	-0.001	0.163	0.166	0.948	0.028
	6	FR	-0.008	0.131	0.138	0.937	0.019
		CO	0.007	0.046	0.046	0.946	0.002
		DS	-0.003	0.143	0.148	0.941	0.022
	12	FR	-0.014	0.118	0.122	0.934	0.015
		CO	0.003	0.046	0.046	0.957	0.002
		DS	-0.008	0.123	0.125	0.942	0.016
4	4	FR	-0.007	0.145	0.150	0.951	0.023
		CO	0.003	0.036	0.036	0.956	0.001
		DS	-0.007	0.171	0.177	0.940	0.031
	6	FR	-0.008	0.130	0.132	0.949	0.017
		CO	0.002	0.036	0.036	0.953	0.001
		DS	-0.001	0.145	0.150	0.939	0.023
	12	FR	-0.011	0.114	0.117	0.939	0.014
		CO	-0.000	0.036	0.035	0.952	0.001
		DS	-0.005	0.121	0.125	0.948	0.016
	6	FR	-0.005	0.150	0.147	0.955	0.022
		CO	0.002	0.031	0.032	0.944	0.001
		DS	-0.002	0.187	0.187	0.940	0.035
6	6	FR	-0.010	0.132	0.134	0.952	0.018
		CO	0.001	0.031	0.031	0.948	0.001
		DS	-0.005	0.153	0.159	0.945	0.025
	12	FR	-0.005	0.114	0.119	0.944	0.014
		CO	0.001	0.031	0.030	0.959	0.001
		DS	-0.000	0.124	0.127	0.942	0.016

Table 2

Simulation results for estimating the gamma shape slope parameter $\gamma_1 = 1.5$. Continuous covariate $Z \sim \mathcal{N}(0.5, 0.5^2)$, $m = 500$. Reported here are the bias, average standard error (ASE), empirical standard deviation (ESD), empirical coverage probabilities for 95% confidence intervals (ECP), and mean-square error (MSE) of the estimates of γ_1 . The estimation methods are the forward recurrence quasi-likelihood (FR) and discrete survival likelihood (DS), both of which make use of binary data which we assume is available; the continuously observed point process likelihood (CO) method represents the ideal case (i.e., the efficient estimator, if one had access to the exact event times).

τ	k	Method	Bias	ASE	ESD	ECP	MSE
2	4	FR	0.019	0.218	0.221	0.941	0.049
		CO	0.001	0.091	0.091	0.948	0.008
		DS	0.007	0.239	0.245	0.938	0.060
	6	FR	0.022	0.204	0.206	0.948	0.043
		CO	-0.006	0.091	0.088	0.961	0.008
		DS	0.016	0.216	0.220	0.939	0.049
	12	FR	0.034	0.190	0.193	0.949	0.038
		CO	0.003	0.091	0.090	0.956	0.008
		DS	0.025	0.194	0.195	0.947	0.039
	4	FR	0.017	0.197	0.206	0.940	0.043
		CO	-0.003	0.067	0.067	0.946	0.005
		DS	0.018	0.215	0.224	0.942	0.051
4	6	FR	0.018	0.181	0.184	0.933	0.034
		CO	0.002	0.067	0.066	0.953	0.004
		DS	0.014	0.190	0.197	0.947	0.039
	12	FR	0.020	0.166	0.169	0.940	0.029
		CO	0.001	0.067	0.069	0.944	0.005
		DS	0.012	0.167	0.167	0.954	0.028
	4	FR	0.014	0.196	0.197	0.954	0.039
		CO	0.002	0.056	0.056	0.942	0.003
		DS	0.015	0.221	0.218	0.955	0.048
	6	FR	0.016	0.176	0.174	0.956	0.031
		CO	-0.000	0.056	0.055	0.963	0.003
		DS	0.013	0.189	0.190	0.954	0.036
	12	FR	0.012	0.161	0.166	0.949	0.028
		CO	0.001	0.056	0.057	0.946	0.003
		DS	0.009	0.161	0.167	0.938	0.028

Table 3

Regression parameter estimates for the gamma renewal process model applied to the Checking In! study. The model assumes that the time between self testing events for the i th subject follows a gamma distribution with $\alpha_i = e^{\gamma_0 + \gamma_1 \text{trti}}$, $\mu_i = e^{\beta_0 + \beta_1 \text{trti}}$. Estimates, standard errors, and p-values are displayed in this table for each of the estimation methods: the discrete survival estimates are obtained by maximizing equation (3), while the forward recurrence estimates are obtained by maximizing (5).

Parameter	Discrete survival			Forward recurrence		
	Est.	SE	p-value	Est.	SE	p-value
γ_0	-2.310	0.169	0.0000	-2.134	0.138	0.0000
γ_1	2.220	0.202	0.0000	2.346	0.175	0.0000
β_0	2.309	0.331	0.0000	1.955	0.149	0.0000
β_1	0.549	0.347	0.1136	0.478	0.169	0.0048

Table 4

Group-specific parameter estimates for Checking In! study. Estimates and 95% confidence intervals refer to the shape and mean parameters of the gamma model assumed for the distribution of times between self-testing events.

Parameter	Group	Discrete survival			Forward recurrence		
		Est.	95% CI		Est.	95% CI	
Shape	Online ($e^{\hat{\gamma}_0}$)	0.099	0.071	0.138	0.118	0.090	0.155
	SMS ($e^{\hat{\gamma}_0 + \hat{\gamma}_1}$)	0.914	0.736	1.135	1.236	1.001	1.526
Mean (months)	Online ($e^{\hat{\beta}_0}$)	10.069	5.266	19.254	7.065	5.279	9.454
	SMS ($e^{\hat{\beta}_0 + \hat{\beta}_1}$)	17.434	14.197	21.410	11.389	9.715	13.352