Biostatistics (2004), **5**, 1, *pp.* 145–151
Printed in Great Britain

How special is a 'special' interval: modeling departure from length-biased sampling in renewal processes

GLEN A. SATTEN*
Centers for Disease Control and Prevention, Atlanta, GA, USA
GSatten@cdc.gov

FANHUI KONG, DAVID J. WRIGHT, SIMONE A. GLYNN, GEORGE B. SCHREIBER
National Heart, Lung, and Blood Institute Retrovirus Epidemiology Donor Study Coordinating Center, Westat, Rockville, MD, USA

SUMMARY

Length-biased sampling occurs in renewal processes when the probability that an interval is selected is proportional to the length of the interval. This can occur when intervals are selected because they contain an event that is independent of the renewal process and occurs with constant hazard. For example, if the times between donations for repeat blood donors are independent and identically distributed, and if the donor seroconverts to HIV (develops antibodies that indicate infection with human immunodeficiency virus), then the interval between the last HIV seronegative and first HIV seropositive test is expected to be longer than that donor's previous time intervals between donations. We develop hypothesis tests to determine if the relationship between the typical and length-biased intervals is as expected, or if there is departure from length-biased sampling. We further develop a regression method to determine if there are covariates that explain the departure from length-biased sampling. Our approach is motivated by the question of whether there is evidence that repeat blood donors who develop antibodies to HIV or other viral infections change their donation pattern in some way because of seroconversion.

Keywords: Human immunodeficiency virus; Infinite-dimensional nuisance parameter; Length-biased sampling; Renewal Process.

1. INTRODUCTION

In renewal theory, it is known that an interval that is tagged by the occurrence of another event is length-biased. For example, the times that a bus passes a certain bus stop may be a renewal process with inter-arrival times having mean μ and variance σ^2, but for an individual who arrives at the bus stop at some random time, the time between the last bus before that individual arrived at the bus stop and the arrival time of the next bus is longer. If the individual is equally likely to arrive at the bus stop at any time, then the 'special' interval between buses that contains her arrival at the bus stop has mean $\mu + \sigma^2/\mu$, because it is subject to length-biased sampling. This result follows because the chance that an interval

*To whom correspondence should be addressed.

between two buses brackets the arrival of the individual at the bus stop is proportional to the length of the interval (see e.g. Karlin and Taylor, 1975, p. 195).

The relationship between the duration of the special interval and the other intervals assumes independence between the arrival times of buses and the arrival time of the passenger at the bus stop. If the bus driver can sense waiting passengers and adjusts the bus arrival time (hopefully to benefit the passengers!) then the relationship between the duration of the 'special' interval $\mu + \sigma^2/\mu$ and the parameters specifying the distribution of the typical intervals is broken. (This relationship may also be broken if the rate at which passengers arrive at the bus stop is not uniform, but we assume here that this is not the case.) If the expected relationship does not hold, then the special interval may in fact be too special!

In this paper, we consider a methodology to determine if the duration of special intervals, relative to the duration of typical intervals, shows any evidence of dependence between the renewal process that generates the intervals and the (constant hazard) Poisson process that selects intervals as special. This question was motivated by the relationship between times of blood donation (which we assume follow a renewal process for each donor) and times at which donors become infected with human immunodeficiency virus (HIV), human T-lymphotropic virus (HTLV) and other viral markers (Schreiber *et al.*, 2002). Blood bankers and other investigators wanted to know if there was any evidence that repeat blood donors who become infected with any of these viruses adjust their donation patterns in any way, for example by returning to make a blood donation sooner than they might have otherwise in order to determine if they have become infected (blood donations are routinely tested for these and other viruses and donors are informed if their donations test positive). For this application, it is reasonable to assume that the hazard for seroconversion is constant over the time of the study. A renewal process is a convenient way to model repeated event data such as blood donations (Aalen and Husebye, 1991) and a number of models for the effect of covariates on recurrence times in a renewal process have been proposed (e.g. Chang and Wang, 1999).

There are three challenges in these data. The first challenge involves the proper adjustment for length-biased sampling; the time between two blood donations that contains a seroconversion (development of detectable antibodies to a virus) event is a special interval and hence is subject to length-biased sampling. The second challenge is that each donor can have unique patterns of donation. While it may be reasonable to assume times of repeated donations follow a renewal process, the mean and variance are potentially different for each donor. The third challenge is that only events occurring in some fixed time window (the study period) are actually observed. Finally, we may also wish to examine the effects of covariates on any departure in the duration of the special interval from its expected value. In Section 2 we develop our notation, while our new estimators are presented in Section 3. In Section 4, we consider the effect of the sampling scheme and in Section 5 we analyze the repeat blood donor data described above.

2. DEFINITIONS AND NOTATION

Suppose for each of m persons we observe $n_i + 2$ events from a renewal process for $1 \leqslant i \leqslant m$. These $n_i + 2$ events define $n_i + 1$ inter-event intervals, of which one is special in the sense described in the introduction. Let X_{ij}, $1 \leqslant j \leqslant n_i$ denote the duration of the typical intervals and let Y_i denote the duration of the special interval. For the blood donation example, $n_i + 2$ is the total number of donations made by the ith donor during the study period, X_{ij} are the time intervals between donations at which the donor tested negative for viral antibodies, and Y_i is the time interval between the last negative and first positive donation. In the blood donor example the special interval is always the last interval, but we assume that the process would have continued uninterrupted if no seroconversion had taken place, so there is no significance attached to the special interval always being the last interval. Finally, suppose that we

observe a column vector of covariates $\boldsymbol{z}_i$ for each person that may explain whether or not the duration of the special interval Y_i is longer than expected given the values of the X_{ij}. We assume that for the ith person the mean renewal time is μ_i and the expected value of the square of the renewal time is ν_i. We assume that $\boldsymbol{\phi}_i \equiv (\mu_i, \nu_i, n_i, \boldsymbol{z}_i)$ are iid from some unspecified distribution. Then we have

$$E[X_{ij}|\boldsymbol{\phi}_i] = \mu_i$$

$$E[X_{ij}^2|\boldsymbol{\phi}_i] = \nu_i.$$

Assuming Y_i is subject to length-biased sampling, we would expect

$$E[Y_i|\boldsymbol{\phi}_i] = \frac{\nu_i}{\mu_i}. \tag{1}$$

We specify departures from (1) by writing

$$E[Y_i|\boldsymbol{\phi}_i] = \frac{\nu_i}{\mu_i}(\boldsymbol{\psi}^{\mathrm{T}} \cdot \boldsymbol{z}_i) \tag{2}$$

where $\boldsymbol{\psi}$ is a column vector of parameters that specifies deviations from the expected relationship (1). We assume that the first component of $\boldsymbol{z}$ is always 1, so that (1) holds for every donor when $\psi = \psi_0 \equiv (1, 0, \ldots, 0)^{\mathrm{T}}$. (We assume that if $\boldsymbol{z}$ has dimension d then values of $\boldsymbol{z}_i$ do not have lower dimension in the sense that there is not some d-dimensional vector $\boldsymbol{\tau} \neq 0$ such that $\boldsymbol{\tau}^{\mathrm{T}} \cdot \boldsymbol{z}_i = 0$ for every i.)

3. Inference on the duration of the special interval

To make inference on whether model (1) holds, define $S_i(\boldsymbol{\tau})$ by

$$S_i(\boldsymbol{\tau}) = Y_i \overline{X}_i - (\boldsymbol{\tau}^{\mathrm{T}} \cdot \boldsymbol{z}_i)\overline{X^2}_i$$

where $\boldsymbol{\tau}$ is an arbitrary column vector with d components, $\overline{X}_i = n_i^{-1} \sum_{j=1}^{n_i} X_{ij}$, and $\overline{X^2}_i = n_i^{-1} \sum_{j=1}^{n_i} X_{ij}^2$. Note that when $\boldsymbol{\tau} = \boldsymbol{\psi}$,

$$E[S_i(\boldsymbol{\psi})|\boldsymbol{\phi}_i] = \frac{\nu_i}{\mu_i}(\boldsymbol{\psi}^{\mathrm{T}} \cdot \boldsymbol{z}_i)\mu_i - (\boldsymbol{\psi}^{\mathrm{T}} \cdot \boldsymbol{z}_i)\nu_i = 0 \tag{3}$$

where we have used (2) and the definition of ν_i. As a result, $S_i(\cdot)$ is an estimating function for $\boldsymbol{\psi}$, and can be used to construct tests about and estimators of $\boldsymbol{\psi}$.

A test of the simple hypothesis $H_0 : \boldsymbol{\psi} = \boldsymbol{\psi}_0$ can be constructed by noting that under the null hypothesis, $\boldsymbol{S}(\boldsymbol{\psi}_0) := m^{-1} \sum_i S_i(\boldsymbol{\psi}_0)\boldsymbol{z}_i$ has an asymptotically multivariate normal distribution with variance–covariance matrix $m^{-1}\boldsymbol{\Sigma}_0$. We may estimate $\boldsymbol{\Sigma}_0$ by $\widehat{\boldsymbol{\Sigma}}_0 = \frac{1}{m} \sum_{i=1}^{m} S_i^2(\boldsymbol{\psi}_0)\boldsymbol{z}_i\, \boldsymbol{z}_i^{\mathrm{T}} - \boldsymbol{S}(\boldsymbol{\psi}_0)\boldsymbol{S}^{\mathrm{T}}(\boldsymbol{\psi}_0)$. A global test of the simple null hypothesis $\boldsymbol{\psi} = \boldsymbol{\psi}_0$ can then be constructed by comparing

$$G \equiv m\boldsymbol{S}^{\mathrm{T}}(\boldsymbol{\psi}_0)\widehat{\boldsymbol{\Sigma}}_0^{-1}\boldsymbol{S}(\boldsymbol{\psi}_0) \tag{4}$$

to the appropriate quantile of a chi-square test with d degrees of freedom.

The parameter vector $\boldsymbol{\psi}$ can also be estimated by solving the estimating equations

$$\sum_i S_i(\widehat{\boldsymbol{\psi}})\boldsymbol{z}_i = \boldsymbol{0} \tag{5}$$

that correspond to minimizing $\sum_i S_i^2(\boldsymbol{\tau})$ with respect to $\boldsymbol{\tau}$. The solution to (5) is given by

$$\widehat{\boldsymbol{\psi}} = \left(\sum_i \overline{X^2}_i \, \boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}} \right)^{-1} \cdot \left(\sum_i Y_i \overline{X}_i \boldsymbol{z}_i \right) \tag{6}$$

which gives a closed-form estimator of $\boldsymbol{\psi}$ so long as $\sum_i \overline{X^2}_i \, \boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}}$ is invertable. Standard theory for M-estimators (see e.g. Sen and Singer, 1993) shows that $\sqrt{m}(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi})$ has, asymptotically, a multivariate normal distribution with variance–covariance matrix $\boldsymbol{\Sigma}$, which can be estimated by $\widehat{\boldsymbol{\Sigma}}$ given by

$$\widehat{\boldsymbol{\Sigma}} = \left(\frac{1}{m} \sum_i \overline{X^2}_i \, \boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}} \right)^{-1} \cdot \left(\frac{1}{m} \sum_i S_i^2(\widehat{\boldsymbol{\psi}}) \boldsymbol{z}_i \, \boldsymbol{z}_i^{\mathrm{T}} \right) \cdot \left(\frac{1}{m} \sum_i \overline{X^2}_i \, \boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}} \right)^{-1}. \tag{7}$$

Tests of composite hypotheses can be constructed using Wald-like statistics. For example, testing whether k linear combinations of the components of $\widehat{\boldsymbol{\psi}}$ (denoted $\boldsymbol{C} \cdot \widehat{\boldsymbol{\psi}}$ where $\boldsymbol{C}$ is a $k \times d$ matrix of coefficients) are equal to some k-dimensional vector $\boldsymbol{\psi}_{0k}$ can be accomplished using the statistic

$$G' = m(\boldsymbol{C} \cdot \widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_{0k})^{\mathrm{T}} \{ \boldsymbol{C}^{\mathrm{T}} \cdot \widehat{\boldsymbol{\Sigma}} \cdot \boldsymbol{C} \}^{-1} (\boldsymbol{C} \cdot \widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_{0k}) \tag{8}$$

which has an (asymptotic) χ^2 distribution with k degrees of freedom. When $k = d$ the statistic G' is asymptotically equivalent to G given in equation (4).

4. A NOTE ON SAMPLING

In many studies (including the blood donor study), only those events that occur within a certain time interval are observed. We will refer to this situation as 'conditional sampling'. We show here that the results in Section 3, corresponding to unconditional sampling, are also valid for conditional sampling. Suppose that only intervals between renewal events that occur within some time interval $[0, \mathcal{T}]$ are observed. Further, suppose that the first renewal event after time 0 for the ith individual occurs at time τ_{0i}. Then, treating τ_{i0} as the zero of time for the ith individual, only those events that occur before time $T_i = \mathcal{T} - \tau_{0i}$ are seen. In particular, if only persons with 'special' intervals are included in the analysis, the special interval must have concluded before time T_i. For the blood donation example, donors are enrolled at their first donation after time 0 (time τ_{i0}) and only donations made before the termination of the study (time $\mathcal{T}$) are included. The interval between the last donation and $\mathcal{T}$ is not included because HIV serostatus is not observed at time $\mathcal{T}$. The interval between 0 and τ_{0i} is also not used even though serostatus at time 0 can be inferred from a negative HIV test at time τ_{0i}, as a stationarity assumption is required to use this interval and its mean duration is not the same as the typical intervals.

To proceed, we first condition on $n_i + 1$, the number of intervals observed by time T_i. Among persons who have experienced n_i standard intervals and one special interval, the joint distribution of these intervals $X_1, X_2, \ldots, X_{n_i}, Y$ is given by

$$f(x_1, x_2, \ldots, x_{n_i}, y) = \frac{\left[\prod_{k=1}^{n_i} g(x_k) \right] y \, g(y) \; I\left[y + \sum_{k=1}^{n_i} x_k < T_i \right]}{D(T_i)}$$

where g is the density function for the renewal process and

$$D(T_i) = \int \ldots \int \left[\prod_{k=1}^{n_i} g(x_k) \right] y \, g(y) \; I\left[y + \sum_{k=1}^{n_i} x_k < T_i \right] \mathrm{d}x_1 \, \mathrm{d}x_2 \ldots \mathrm{d}x_{n_i} \, \mathrm{d}y.$$

Table 1. *Results of analysis of blood donor data from REDS study: evidence for a departure in length of the seroconversion interval from its expected value for two viral diseases*

Virus (seroconverters)	Mean, median (Range) of n_i	Mean of $\overline{X}_i$ (days)	Mean of Y_i (days)	G (p-value)	ψ (95% confidence interval)
HIV (49)	4.73, 4 (2 − 38)	209	450	6.95 (0.008)	1.42 (1.05, 1.79)
HTLV (32)	3.5, 3 (2 − 17)	341	440	1.00 (0.32)	0.61 (0.18, 1.05)

Note that the X_j and Y are no longer independent due to the term $I\left[y + \sum_{j=1}^{n_i} x_i < T_i\right]$, and that $E(X_j|n_i, T_i)$ will not be the same as in the unconditional renewal process. However, note that $E(X_j|n_i, T_i) = E(X_k|n_i, T_i)$. Further,

$$E(X_j Y|n_i, T_i) = \frac{\int \ldots \int x_j \, y^2 \left[\prod_{k=1}^{n_i} g(x_k)\right] g(y) \, I\left[y + \sum_{k=1}^{n_i} x_k < T_i\right] \mathrm{d}x_1 \, \mathrm{d}x_2 \ldots \mathrm{d}x_{n_i} \, \mathrm{d}y}{D(T_i)}$$

and

$$E(X_j^2|n_i, T_i) = \frac{\int \ldots \int x_j^2 \, y \left[\prod_{k=1}^{n_i} g(x_k)\right] g(y) \, I\left[y + \sum_{k=1}^{n_i} x_k < T_i\right] \mathrm{d}x_1 \, \mathrm{d}x_2 \ldots \mathrm{d}x_{n_i} \, \mathrm{d}y}{D(T_i)}$$

and hence $E(X_j Y - X_j^2|n_i, T_i) = 0$ under conditional sampling. Taking further expectation with respect to the distribution of $n_i|T_i$ gives $E(X_j Y - X_j^2|T_i) = 0$. If we redefine $\phi_i \equiv (\mu_i, \nu_i, n_i, \boldsymbol{z}_i, T_i)$ it is easy to see that conditional sampling inherits the iid structure on ϕ_i assumed in Section 3. As a result, the inference procedures presented in Section 3 are applicable to studies conducted using a conditional sampling scheme without further modification.

5. The retrovirus epidemiology donor study

The retrovirus epidemiology donor study (REDS) has collected a database of information on 6.8 million non-autologous (i.e. not for the subsequent use of the donor) blood donations made between 1991 and 1997 at five blood centers (Schreiber *et al.*, 2002). Up to seven years of follow-up data is available for each donor, including data on times of donations, donor demographics and results from routine laboratory screening for HIV, HTLV and other viruses. Only donations made during the course of the REDS study are included in the study, corresponding to the conditional sampling scheme discussed in Section 4. Although repeat blood donors are a very low-risk population (Lackritz *et al.*, 1995) and persons acknowledging risk factors for HIV and other viral diseases are deferred from making donations, a small number of donors do become infected with one or more of these viruses. It is of interest to know if seroconversion, which occurs on average about 20–35 days after infection (Lackritz *et al.*, 1995; Busch and Satten, 1997; Sternberg and Satten, 1999), has any effect on donation patterns. These effects could be conscious (e.g. donors making more frequent donations at times they engage in risk behaviors in order to obtain serologic test results routinely performed on all blood donations) or unconscious (e.g. delay of donations due to development of flu-like symptoms of primary HIV infection which occur just before seroconversion (Busch and Satten, 1997)).

In Table 1, we show data from 49 donors who seroconverted to HIV and 32 donors who seroconverted to HTLV (Schreiber *et al.*, 2002). Donors seroconverting to either virus have negative-to-positive

Table 2. *Effect of covariates on HIV seroconversion interval, REDS study*

Covariate	G' ($\psi_2 = 0$) (d.f.)	p-value	$\widehat{\psi}_1$ (95% C.I.)	$\widehat{\psi}_2$ (95% C.I.)
Number of Prior Donations	2.70 (1)	0.10	1.56 (1.09, 2.03)	−0.08 (−0.17, 0.02)
Acknowledged Risk Factor	1.38 (1)	0.24	1.35 (0.97, 1.74)	0.58 (−0.39, 1.55)

interdonation intervals (Y_i) that are longer than their average negative-to-negative interdonation intervals ($\overline{X}_i$) and hence would naively appear to be examples of seroconversion resulting in a delay of the post-seroconversion donation. However, the interdonation interval that contains the seroconversion is 'special', and hence is expected to be longer. When we analyze the data for persons seroconverting to each virus using the chi-square test G given in equation (4) (conducting separate analyses for each virus and with no additional covariates, i.e. $d = 1$), we discover that only donors seroconverting to HIV appear to delay their post-seroconversion donation (p-values were calculated using the tail area of a chi-square distribution with one degree of freedom). In fact, we estimate $\psi < 1$ for donors seroconverting to HTLV, meaning that they actually may return somewhat sooner than expected (although the 95% confidence interval includes the null value 1).

We were also interested in testing the effect of two covariates on the length of the interdonation interval containing the seroconversion. These two covariates are the number of prior negative donations made during the study period, and whether the donor acknowledges a risk factor for HIV at the counseling session where the donor is informed of their HIV-positive test result. Because there are so few seroconverters, we considered each covariate separately. For each analysis $d = 2$, $\boldsymbol{\psi} = (\psi_1, \psi_2)^{\mathrm{T}}$ and $\boldsymbol{z}_i = (1, z_{i2})^{\mathrm{T}}$, where z_{i2} was either the number of previous negative–negative interdonation intervals available for the ith seroconverter minus 1 or an indicator of denial of a risk factor (0 if risk was acknowledged, 1 if risk factors were denied). Thus, for the first analysis, ψ_1 corresponds to the departure from expected length for a person who has made only three blood donations in the study period, while for the second analysis ψ_1 corresponds to the departure from expected length for persons who acknowledge HIV risk factors. We tested $\psi_2 = 0$ using G' given in equation (8) with $k = 1$, $C = (0, 1)$ and $\psi_{0k} = 0$. These results are shown in Table 2, along with point estimates of ψ_1 and ψ_2 and 95% confidence intervals for ψ_1 and ψ_2 obtained using (7). From Table 2, we see that the number of previous seronegative donations does not explain the increased length of time between the last negative and first positive blood donations for donors who seroconvert to HIV, but whether donors acknowledge a risk factor for HIV has a larger (although non-significant) effect. In particular, donors who deny risk have only a slightly longer seroconversion interval than expected (the 95% confidence interval for ψ_1 includes 1); while ψ_2, the increase in duration due to persons who acknowledge a risk factor at their post-HIV-positive donation interview is not significant, it should be recalled that in the overall analysis there was a significant increase in HIV seroconversion interval, and that the sample size for this analysis is very small. Finally, it is also worth noting that all donors have denied risk factors for HIV prior to donation.

6. CONCLUSION

An interval in a renewal process is 'special' if it has been tagged by the occurrence of another event. If this second event occurs with constant hazard, then the duration of the 'special' interval is length biased. We have developed a simple approach to determining if the relationship between the duration of

the 'special' interval and previous 'typical' intervals departs from what is expected under length-biased sampling.

We have additionally shown that our new approach is applicable even if sampling is conditional on events seen in a finite time period. This result may seem unexpected initially, as the duration of renewal events is distorted by conditional sampling. Further, the expected value of a single renewal interval under conditional sampling is a function of the total number of observed events. However, conditional on the number of observed events, we have shown that each interval (including the special interval) is 'squeezed' in such a way that the relationship between special and typical interval durations is preserved.

Finally, assuming that interdonation intervals of repeat blood donors in the REDS study follow a renewal process, we have determined that repeat blood donors that become infected with HIV (but not HTLV) appear to delay their first HIV-positive donation. Further, regression analysis shows that there is evidence that this effect is restricted to persons who are aware that they have factors that may have put them at risk for acquiring HIV infection.

REFERENCES

AALEN, O. O. AND HUSEBYE, E. (1991). Statistical analysis of repeated events forming renewal processes. *Statistics in Medicine* **10**, 1227–1240.

BUSCH, M. P. AND SATTEN, G. A. (1997). Time course of viremia and antibody seroconversion following HIV exposure. *The American Journal of Medicine* **102**, 117–124.

CHANG, S. H. AND WANG, M. C. (1999). Conditional regression analysis for recurrence time data. *Journal of the American Statistical Association* **94**, 1221–1230.

KARLIN, S. AND TAYLOR, H. M. (1975). *A First Course in Stochastic Processes*. San Diego, CA: Academic.

LACKRITZ, E. M., SATTEN, G. A., ABERLE-GRASSE, J., DODD, R. Y., RAIMONDI, V. P., JANSSEN, R. S., LEWIS, W. F., NOTARI, E. P. AND PETERSEN, L. R. (1995). Estimated risk of HIV transmission by screened blood in the United States. *The New England Journal of Medicine* **333**, 1721–1725.

SATTEN, G. A. (1997). Steady-state calculation of the risk of HIV infection from transfusion of screened blood from repeat donors. *Mathematical Biosciences* **141**, 101–113.

SCHREIBER, G. B., GLYNN, S., SATTEN, G. A., KONG, F., BUSCH, M. P., TU, Y. AND KLEINMAN, S. (2002). HIV seroconverting donors delay their return: screening test implications. *Transfusion* **42**, 414–421.

SEN, P. K. AND SINGER, J. M. (1993). *Large Sample Methods in Statistics: an Introduction with Applications*. New York: Chapman and Hall.

STERNBERG, M. AND SATTEN, G. A. (1999). Discrete-time nonparametric estimation for semi-Markov models of chain-of-events data with interval censoring and truncation. *Biometrics* **55**, 514–522.

ZUCK, T. F., THOMSON, R. A. AND SCHREIBER, G. B. *et al.* (1995). The retrovirus epidemiology donor study (REDS): rationale and methods. *Transfusion* **35**, 994–951.

[*Received March 21, 2002; first revision May 16, 2003; second revision September 11, 2003; accepted for publication September 15, 2003*]