STATISTICS IN MEDICINE, VOL. 16, 485–496 (1997)

A MODEL FOR INTERVAL-CENSORED TUBERCULOSIS OUTBREAK DATA

PHILIP J. SMITH

Centers for Disease Control and Prevention, Division of Tuberculosis Elimination (E-10), 1600 Clifton Road NE, Atlanta, GA 30333, U.S.A.

THEODORE J. THOMPSON

Centers for Disease Control and Prevention, Division of Diabetes Translation (K-10), 1600 Clifton Road NE, Atlanta, GA 30333, U.S.A.

AND

JOHN A. JEREB

Centers for Disease Control and Prevention, Division of Tuberculosis Elimination (E-10) 1600 Clifton Road NE, Atlanta, GA 30333, U.S.A.

SUMMARY

As the incidence of tuberculosis (TB) has increased in the United States, occupationally acquired TB has increased among the health care workers (HCWs). This paper describes a model developed in response to the needs of an outbreak of multidrug-resistant TB. One of the goals of the outbreak investigation was to estimate the risk of tuberculin skin test (TST) conversion as a function of HCW job type and the period during which persons were employed over the study period. TST conversions were evaluated at periodic examinations and data are interval-censored. We present a generalized linear model that extends Efron's survival model for censored survival data to the case of interval-censored data. © 1997 by John Wiley & Sons, Ltd.

1. INTRODUCTION

In 1992 the Centers for Disease Control and Prevention conducted a retrospective study of HCWs to investigate occupational exposure to TB and TST conversion history. The study focused on 352 HCWs employed between 1978 and 1992 at a hospital in which an outbreak of multidrug-resistant TB was documented. Each of these HCWs has a negative initial TST result and subsequent testing during the course of their employment. Records for each employee included: the HCW's initial test date at the hospital; his or her job type; the dates on which the employee received a TST; and the outcome, which indicated whether the employee had 'converted', that is, developed a positive reaction to the TST that consisted of a 10 mm or more induration detected at the time of the clinic examination. A conversion was accepted as evidence of recent infection with *Mycobacterium tuberculosis*. Details of the outbreak have been reported by Jereb *et al.*[1] and a discussion of the history and accuracy of the TST are given by the American Thoracic Society[2] and Huebner *et al.*[3]

CCC 0277–6715/97/050485–12

Received July 1995
Revised February 1996

One of the purposes of the investigation was to characterize the risk of TST conversion as a function of a HCW's job type after correcting for increasing temporal trends in the prevalence of TB infection. A salient feature of the data from the investigation was that the exact dates of TB infection were unavailable. The only chronologic data available pertaining to the timing of the end point of interest were the TST dates, which defined the time intervals within which employees may have been infected. When this occurs, data are said to be 'interval-censored'.

In related research, Turnbull[4] describes estimation of the empirical distribution function for interval-censored data, Finkelstein and Wolfe[5] describe a semi-parametric model for interval-censored data, Finkelstein[6] presents a proportional hazards model, Self and Grossman[7] present rank tests, Brookmeyer and Goedert[8] describe a two-stage parametric model, Odell *et al.*[9] investigate the effect of using interval midpoints in lieu of event times, Diamond and McDonald[10] and Shiboski and Jewell[11] describe methods appropriate when subjects' current status is observed once, Dorey *et al.*[12] describe methods based on multiple imputation, Jewell *et al.*[13] give non-parametric methods that account for interval censoring, and Whitehead[14] describes methods that require identical observations times for each subject.

The methods we propose extend Efron's[15,16] semi-parametric models for censored survival data to the case where the end point may be interval- or right-censored. Our methods are appropriate when one follows each subject longitudinally and observes his/her TST status at each of several clinic visits spaced irregularly over the study period. Within this context, the clinic visit dates may differ among subjects. In this case we refer to the data as 'non-synchronized' interval-censored data. In contrast, the methods described by Whitehead require 'synchronized' interval-censored data, that is, clinic visit dates must be the same for each subject. In this case such methods are not entirely appropriate for the many studies in which one expects that the visit dates are non-synchronized.

Finkelstein and Wolfe[5] have proposed methods for non-synchronous interval-censored data. We note, however, that these methods have the undesirable feature that the number of nuisance parameters in their model grows proportionally with the number of observations. The methods described by Diamond and McDonald[10] share this feature. In an important and influential paper, Neyman and Scott[17] show that in this case maximum likelihood estimates need not be consistent. Even if they are consistent, they need not be efficient. McCullagh and Nelder (reference 18, pp. 245, 278) discuss this issue, also. In comparison, as the number of observations increase, the number of parameters in our model remains constant; we present results from one of numerous simulation studies we have conducted and that shows that regression effects are essentially unbiased.

Compared to the methods of Diamond and McDonald,[10] Shiboski and Jewell,[11] and Jewell *et al.*,[13] our methodology allows for time-varying covariates and an easy assesment of the proportional hazards assumption.

Samuelsen and Kongerud[19] outline parametric methods for analysing interval-censored data when each individual's clinic visit times are unique. In a related analysis of current status data with parametric survival models, however, Ades and Nokes[20] show that results can depend sensitively upon the specific parametric assumptions made. In this regard, a semi-parametric method seems desirable.

In comparsion to the methods described by Self and Grossman[7] and Dorey *et al.*,[12] the model we present belongs to the class of generalized linear models[18,21] and thus we may apply seamlessly the many advances in this research area.[22–27]

In Section 2 we describe the generalized linear model and in Section 3 we illustrate the methods using data from the TB outbreak investigation. Section 4 presents a simulation study that investigates the efficacy of the method and Section 5 concludes with a discussion.

2. THE GENERALIZED LINEAR MODEL

In a sample of N subjects, let T_i denote the unobserved TST conversion time for the ith subject, t_{i0} denote the chronologic date at which the ith subject commenced employment at the hospital and commenced exposure. Also, let $t_{i1} < t_{i2} < t_{i3} < \ldots$ denote the distinct chronologic times following t_{i0} at which the ith subject is examined. Let $t_{i,n_i} = \min_j \{t_{ij}: T_i \leqslant t_{ij}\}$ provided one has observed that the ith subject converted during the study period, and $t_{i,n_i} = \max_j \{t_{ij}\}$, otherwise. Also, let $I_{ij} = (t_{ij-1}, t_{ij}]$ denote the jth interval between consecutive examinations of the ith subject, x_{ij} denote the ith subject's covariates during I_{ij}, and let

$$\pi_{ij} = P(T_i \in I_{ij} | T_i > t_{ij-1}) \tag{1}$$

denote the ith subject's hazard probability in interval I_{ij}.

Letting $y_{ij} = 1$ if $T_i \in I_{ij}$ and 0, otherwise, the likelihood is

$$\prod_{i=1}^{N} \prod_{j=1}^{n_i} \left[\pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{1 - y_{ij}} \right]. \tag{2}$$

For the proportional hazards model, the instantaneous hazard of conversion at time t for a subject with covariates x is

$$\pi(t, x) = \lambda_0(t) \exp\{\beta^{\mathrm{T}} x\} \tag{3}$$

where $\lambda_0(t)$ is the baseline hazard at time t and β is a vector of regression coefficients for estimation.

In this case

$$\pi_{ij} = 1 - \exp\{-\mathrm{e}^{\psi_{ij} + \beta^{\mathrm{T}} x_{ij}}\} \tag{4}$$

where $\psi_{ij} = \log(\Lambda_0(t_{ij}) - \Lambda_0(t_{i,j-1}))$, $\Lambda_0(t) = \int_{-\infty}^{t} \lambda_0(u) \mathrm{d}u$, and x_{ij} denotes a vector of time-varying covariates.

We may regard equation (2) as the product of $\sum n_j$ independent Bernoulli densities. In the parlance of generalized linear models, the random components of these densities are $\{y_{ij}\}$ with 'success' probabilities $\{\pi_{ij}\}$.

To identify our model further as a member of the family of generalized linear models, let

$$\eta_{ij} = \psi_{ij} + \beta^{\mathrm{T}} x_{ij} \tag{5}$$

denote the linear predictor that corresponds to the random component y_{ij}. For the proportional hazards model (4), the link function used to connect the linear predictor η_{ij} to $E(y_{ij}) = \pi_{ij}$ is the complementary log–log link

$$\log\{-\log(1 - \pi_{ij})\} = \psi_{ij} + \beta^{\mathrm{T}} x_{ij}.$$

In any analysis of the type of interval-censored data we describe, we must specify the parameters ψ_{ij} in some manner. For example, in a fully parametric proportional hazards analysis we can express ψ_{ij} as an analytic function that depends upon the particular distributional assumptions for unobserved event times. For example, if we assume that the distribution of unobserved event times has a Weibull distribution with baseline hazard function $\lambda_0(t) = \alpha \mu_0 t^{\alpha - 1}$, then $\psi_{ij} = \log\{\mu_0 [t_{ij}^{\alpha} - t_{ij-1}^{\alpha}]\}$.

In a semi-parametric proportional hazards analysis in which data are synchronized and examination times are the same for each subject, $\psi_{ij} = \psi_j$. In this case, we estimate the examination time effects by coding this effect as a factor variate in the linear predictor (5). Alternatively, we could use orthogonal polynomials to account for the examination time factor more parsimoniously.

In our approach for specifying ψ_{ij}, we note that (by the mean value theorem) for a suitably chosen value of m'_{ij} ($t_{ij-1} < m'_{ij} \leqslant t_{ij}$)

$$\begin{aligned}\psi_{ij} &= \log\{\Lambda_0(t_{ij}) - \Lambda_0(t_{ij-1})\}\\ &= \log\{w_{ij}\lambda_0(m'_{ij})\}\\ &= \log(w_{ij}) + \log(\lambda_0(m'_{ij}))\end{aligned} \tag{6}$$

where $w_{ij} = t_{ij} - t_{i,j-1}$ denotes the width of I_{ij}.

For the usual case in which the distribution of event times is unknown, we empirically model the nuances of the unknown function $\log(\lambda_0(m'_{ij}))$ in (6) by an appropriately chosen smooth function of the midpoint of I_{ij}, m_{ij}. In survival analyses of head and neck cancer data and gamma-ray burst data in which we assume that we know event times exactly, Efron[15,16] uses this identical approximation for the logarithm of the difference of integrated hazard functions. Specifically, when this function is a second-degree polynomial, an approximation to the linear predictor (5) is

$$\eta_{ij} = \log(w_{ij}) + \alpha_0 + \alpha_1 m_{ij} + \alpha_2 m_{ij}^2 + \beta^{\mathrm{T}} x_{ij} \tag{7}$$

where the term $\log(w_{ij})$ represents an offset in the linear predictor.

To obtain maximum likelihood estimates of the regression parameters in (7), their standard errors, and deviance tests, one can use the iteratively reweighted least squares alogrithm implemented in many standard statistical software packages. It is possible to investigate departures from the proportional hazards assumption by examining the significance of interactions between the midpoint and covariate effects.

2.1. Estimation of the survival curve

To obtain a smooth estimate of the survival curve, we must choose a small value $\omega > 0$ to define a fine grid of hypothetical examination times $a'_j = \min_i\{t_{i0}\} + jw$. These define new intervals $I'_j = (a'_{j-1}, a'_j]$ with width w and midpoints $m'_j = (a'_{j-1} + a'_j)/2, j = 1, 2, \ldots$. We can estimate the hazard probabilities of the new intervals using

$$\hat{\pi}_{x',j} = 1 - \exp\{-\exp(\hat{\eta}_{x',j})\}$$

where $\hat{\eta}_{x',j}$ denotes the linear predictor (7) evaluated at the maximum likelihood estimates of the regression parameters for the specified covariate vector x'.

The maximum likelihood estimate of the probability of being free from infection beyond time a'_j is

$$\hat{s}_{x',j} = \prod_{k=1}^{j-1}(1 - \hat{\pi}_{x',k}) \tag{8}$$

$j = 2, \ldots$, where $\hat{s}_{x',1} = 1$, by definition.

Letting $\hat{\mathscr{I}}$ denote the expected information for the regression parameters of the linear predictor (7), the Appendix shows that a Taylor series approximation for the standard error of $\hat{s}_{x',j}$ is

$$\mathrm{SE}(\hat{s}_{x',j}) = \hat{s}_{x',j}\left[\left(\sum_{k=1}^{j-1} v_{x',k}\right)\hat{\mathscr{I}}^{-1}\left(\sum_{k=1}^{j-1} v_{x',k}\right)^{\mathrm{T}}\right]^{1/2} \tag{9}$$

where $v_{x',k} = \exp\{\hat{\eta}_{x',k}\}x'$.

Table I. Final TST results for HCWs, 1978–1992

	Negative	Positive
Nursing staff (N)	49	36
Housekeeping (H)	8	5
Food service (FD)	16	8
Others with frequent exposure (OF)	28	10
Laboratorians (L)	13	4
Radiology personnel (R)	10	3
Others with intermediate exposure (OI)	43	12
Secretaries (S)	17	3
Others with negligible exposure (ON)	80	7

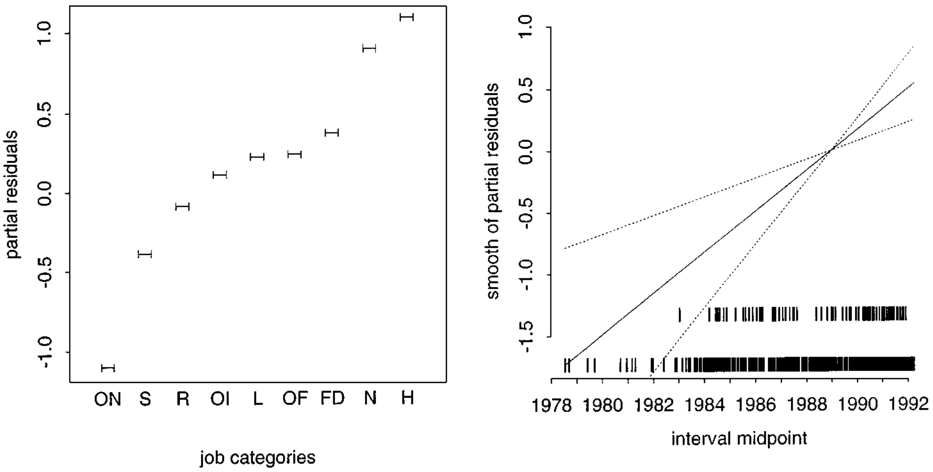

Figure 1. Partial residual plots for job category and interval midpoint generalized linear model

3. ANALYSIS OF THE TB OUTBREAK DATA

Table I lists the numbers of HCWs whose TST converted to positive over the course of their employment. Figure 1 shows the smoothed partial residual plots (solid line) and 95 per cent confidence bands (dashed lines) of our generalized linear model that includes job type and the interval midpoint polynomial as main effects.

The upper row of jittered tick marks in the right-most plot in the Figure 1 indicates the midpoints of censoring intervals within which conversions were observed. The lower row of jittered tick marks corresponds to midpoints of subjects' censoring intervals regardless of whether a conversion was observed. In all, there are 1035 different intervals in our data set, the median width being 1·25 years. These tick marks show that few conversions occurred from 1978–1983. Thereafter the rate at which conversions occurred increased as time progressed. This observation is congruent with the increasing TB admissions at the index hospital.[1]

Table II lists the relative risks for several HCW job types where the reference group consists of HCWs with negligible exposure to patients (for example, accountants). This table shows that for nurses, housekeeping personnel and other HCWs who receive frequent exposure to hospitalized persons with TB, the risk of infection is significantly elevated. Also, for persons deemed

Table II. Chronologically adjusted relative risks and 95 per cent confidence intervals for HCW job types

	Lower	Relative risk	Upper
Housekeeping (H)	2·9	8·9	28
Nursing staff (N)	3·3	7·4	16
Food service (FD)	1·6	4·4	12
Others with frequent exposure (OF)	1·5	3·8	9·9
Laboratorians (L)	1·1	3·8	13
Others with intermediate exposure (OI)	1·3	3·4	8·4
Radiology personnel (R)	0·72	2·8	10
Secretaries (S)	0·53	2	7·8

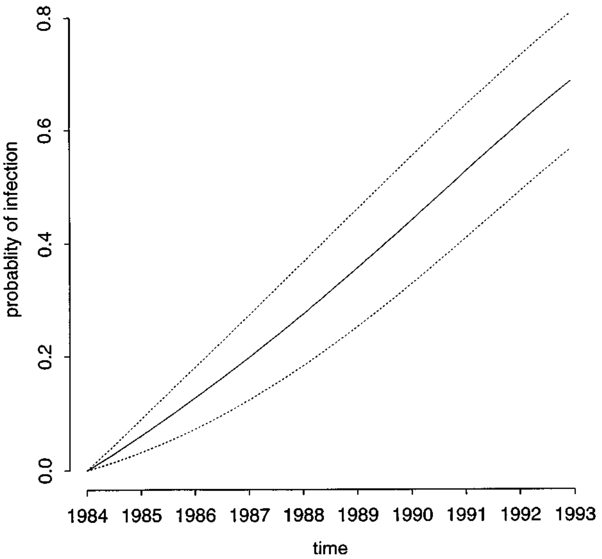

Figure 2. Estimated infection probabilities and 95 per cent confidence bounds for a person starting employment on 1/1/1984, nursing staff (N)

to have intermediate exposure (for example, dieticians, laboratorians who analyse potentially dangerous TB cultures), there is a significantly increased risk of conversion.

Figure 2 illustrates the risk of infection of a hypothetical person who commenced employment in 1984 at the hospital under study. This plot shows that if the person was employed as a nurse in the hospital, we estimate the probability of infection to have increased to nearly 0·80 by 1992.

4. A SIMULATION STUDY

To assess the efficacy of our model for interval-censored data we conducted several large Monte Carlo studies to estimate bias, variability, mean squared error, and loss of efficiency resulting from observations being interval-censored.

In these simulations, we generated event times to follow a Weibull distribution. The Weibull density function is

$$f(t) = \alpha\lambda(\lambda t)^{\alpha-1} \exp - (\lambda t)^{\alpha}$$

where α is a shape parameter and λ is a scale parameter such that $\lambda = \exp\{-\beta^{T}x\}$ where x is a p-vector of covariates and β is a vector of regression parameters for estimation. Letting $\mu = -\ln\lambda = \beta^{T}x$, $\alpha = 1/\sigma$, and $U = (\ln(T) - \mu)/\sigma$, the survival function of the random variable U is

$$S_U(u) = \mathrm{e}^{-\mathrm{e}^{u}}.$$

In simulating the event times, we assumed that $t = 0$ was the starting point and the objective was to estimate treatment effects at $t = 1$ under different assumptions about the hazard. In this case,

$$\beta^{T}x = \beta_0 + \beta_1 x_1$$

where x_1 is a binary predictor indicating treatment group ($x_1 = 1$ corresponding to treatment and $x_1 = 0$ corresponding to control).

We chose five values of σ, $\sigma \in \{0{\cdot}65, 0{\cdot}8, 1, 1{\cdot}25, 1{\cdot}54\}$, representing realistic values in TB applications that correspond to conditions that range from rapidly increasing to rapidly decreasing hazard rates. We also selected probabilities at $t = 1$ for both treatment groups. For example if we assume that 30 per cent of individuals on treatment will get the event by $t = 1$ but that only 10 per cent of those assigned to control will get the event, then

$$\beta_0 + \beta_1 = -\sigma \log(-\log(0{\cdot}7))$$

and

$$\beta_0 = -\sigma \log(-\log(0{\cdot}9)).$$

One can obtain solutions for the regression parameters, β_0 and β_1, by solving these simultaneous equations for a given value of σ.

We ran 500 repetitions for each simulation at each of six samples sizes $n \in \{25, 50, 100, 200, 400, 800\}$ per group for rates 5–10 (that is 5 per cent for the treatment group and 10 per cent for the control group), 10–30, 30–50, 30–70 and 80–90. We took censoring interval endpoints as the deciles of the distribution that corresponds to the control group's Weibull distribution. In this regard, the design of our simulation corresponds to the design of the many tuberculosis relapse trials conducted by the Centers for Disease Control and Prevention.

To investigate the efficacy of the model for interval-censored data, we assumed the linear predictor was of the form given by (7), in which $\log(\lambda_0(m'_{ij}))$ is approximated by a second-degree polynomial in the interval midpoints. Also, we used a second model in which we estimated treatment effects using a parametric proportional hazards model in which we know the event times exactly and the underlying distribution of times to the event corresponded to a Weibull distribution.

The purpose of the second model was to estimate the relative loss in efficiency that results from estimation of treatment effects when the data are interval-censored, as compared to when the event times are known exactly along with the analytic form of their distribution. We obtained maximum likelihood estimates for the Weibull model by the method of successive relaxation described by Aitkin *et al.*[21]

Figure 3 gives graphical summaries that depict how the relative bias of estimated treatment effects depends upon sample size, σ, and the true treatment effect, β_1 for the model for interval censoring and the Weibull model for exact event times. Both models show negative relative bias when the sample size per group is small and that the bias becomes negligible as the sample size per group increases to levels that one would expect in well designed clinical trials.

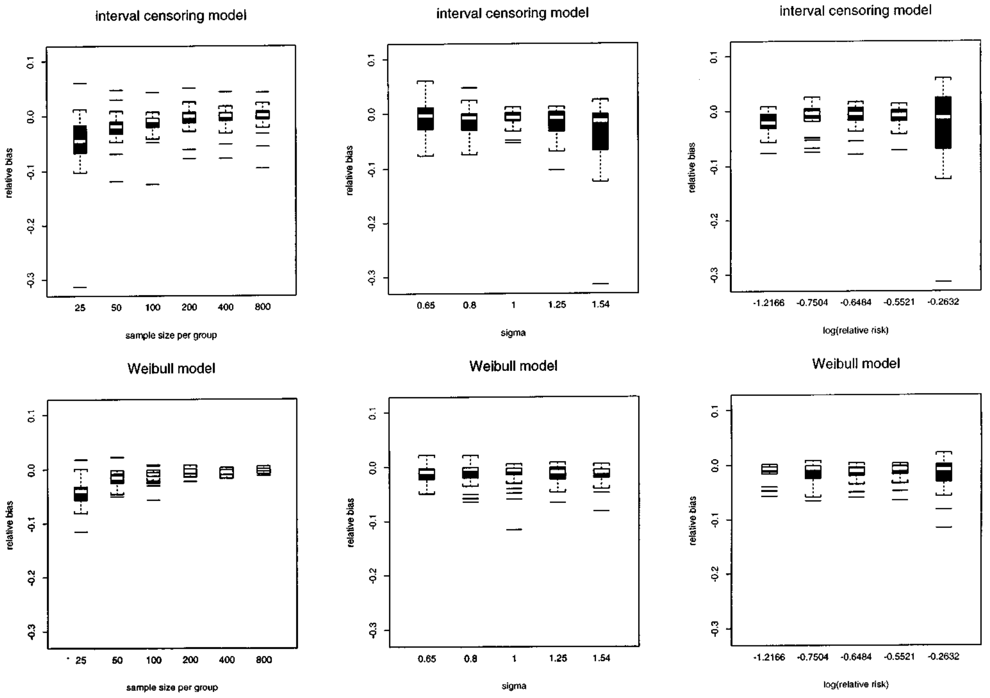

Figure 3. Simulation results: relative bias

For the case in which event times are exponentially distributed, $\sigma = 1$, and the midpoint of each interval is exactly equal to the suitably chosen value of m' that satisfies the mean value theorem (6). In this case, it may not be surprising that the relative bias is negligible for the model that accounts for interval censoring in Figure 3. For other more extreme values of σ, the relative bias of both models is negligible, although the variability is greater for the model that accounts for interval censoring. Similarly, relative bias of both models is negligible for different values of the treatment effect β_1.

In general, our results that suggest negligible relative bias when one models interval-censored data by smoothing the unknown underlying hazard function semi-parametrically are concordant with the results of statistical research conducted by Lin[28] who investigated the bias of regression effects in categorical models for right-censored data when one models the unknown underlying hazard function semi-parametrically.

Figure 4 gives graphical summaries that depict how that root mean squared error (RMSE) of estimated treatment effects depends upon sample size, σ, and the true treatment effect, β_1 for the model for interval censoring and the Weibull model for exact event times. Both models show that the RMSE decreases rapidly as sample size per treatment group increases. Also, this figure shows that median RMSEs for the model for interval-censored data and the Weibull model are comparable for different values of σ and β_1, although the variability in RMSEs for the moel for interval-censored data is slightly greater. Over all conditions, the model for interval-censored data tended to be 11 per cent efficient as compared to the model in which the event times and their distribution are known.

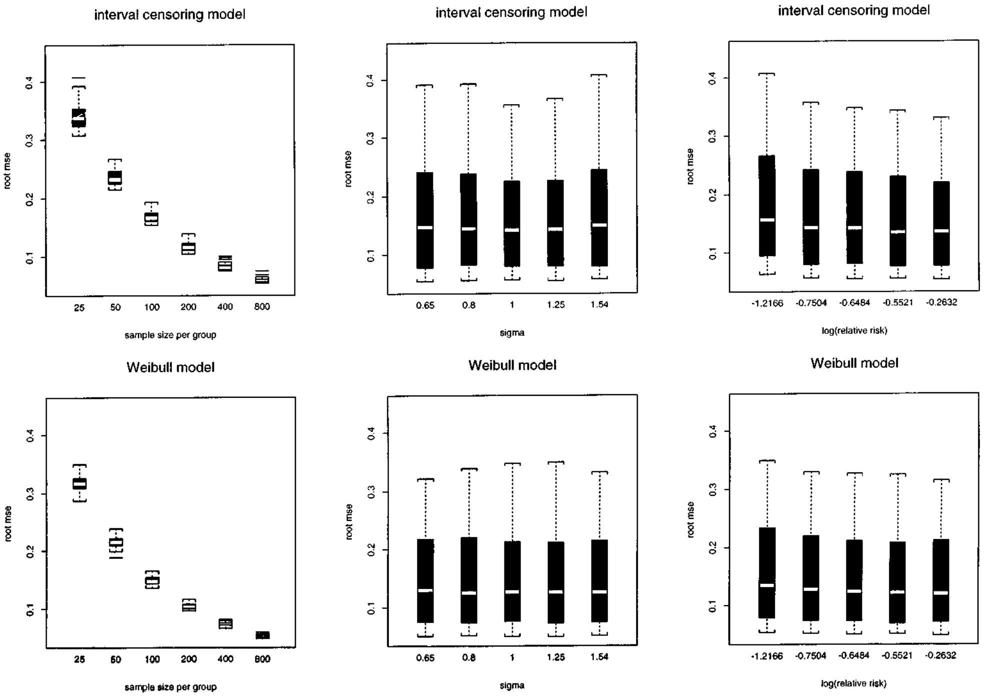

Figure 4. Simulation results: efficiency

5. DISCUSSION

Specification of the proportional hazards model (3) provides a starting point form which one may easily derive a generalized linear model for interval-censored data. One consequence of the proportional hazards assumption is that survival curves expressed on the scale of the linear predictor are parallel and cannot cross. By allowing the modelling of $\log(\lambda_0(\cdot))$ as an emprically-chosen function of time, one may assess this consequence by investigating interactions between time and other predictors. Also, one achieves a good measure of parsimony in an otherwise over-specified model.

A further consequence of the proportional hazards assumption (3) is that the link function that connects the mean of the random component to its associated linear predictor is the complementarty log–log transformation. When the prevalence rate is less than 0·2, the complementary log–log and logit transformations are essentially the same. In this common situation, choice of the link function is not a pivotal issue and one may specify it depending on whether one wishes to estimate odds ratios or relative risks, provided, respectively, by the logit and complementary log–log links. For the former case, one may estimate hazard probabilities by computing the anti-logit of the linear predictor (7) evaluated at the maximum likelihood estimates of the regression parameters and obtain standard errors of the survival probabilities (8) with use of (9) where $v_{x',k} = \hat{\pi}_{x',k}(1 - \hat{\pi}_{x',k})x'$.

Finally, it is important to realize that there are situations in which any model that accounts for interval censored data will give poor results. For example, if the hazard is rapidly increasing (or rapidly decreasing) and the duration of time between observations is very long, there will be great uncertainty about the time at which the event actually occurred. In this situation, almost all of the 'signal' pertaining to the hazard is lost. Our experience indicates that even parametric models for interval censored data perform poorly in this situation. In follow-up studies such as relapse trials, this result underscores the importance of an appropriate specification of time between scheduled observations.

For a wide range of hazards realistically expected in TB research, however, the numerous simulations that we have conducted indicate that our model provides estimated regression effects with small or negligible relative bias.

We note that using equation (7) is equivalent to using the mean value theorem to approximate ψ_{ij}. One referee commented that an intuitive diagnostic for the adequacy of this approximation would be to examine the magnitudes of α_1 and α_2 relative to α_0 since α_1 and α_2 determine whether the hazard rate is not constant from interval to interval. If $r_1 = \alpha_1/\alpha_0$ and $r_2 = \alpha_2/\alpha_0$ are both large, then one could guess that 'long' w_{ij} intervals may be problematic. An alternative approximation for ψ_{ij} is based on the trapezoidal rule; the values of both of the endpoints of each interval are used along with interval widths to approximate ψ_{ij}. Within this context, separate polynomials may be used for the right and left interval endpoints. We note, however, that even in situations in which the hazard is rapidly increasing (or rapidly decreasing) the approximation using the mean value theorem can be improved by transforming the time scale using logarithms. Empirical research conducted by Lin[28] has shown that this technique compresses the intervals' widths and improves the approximation, and leads to greater accuracy of regession estimates.

Finally, we have outlined methods that enable us to estimate smooth survival curves. In principle, this method can be used to estimate survival curves for any specified set of clinic visit times over which predictors may vary. For the case in which clinic visit times are highly non-synchronous, alternative estimates of a smooth survival curve may be obtained using methods described by Tanner and Wong,[29,30] and Anderson and Senthilselvan.[31,32]

APPENDIX

Using Taylor series,

$$\operatorname{var}\{\ln(\hat{s}_{x',j})\} \approx \hat{s}_{x',j}^{-2} \operatorname{var}\{\hat{s}_{x',j}\}. \tag{10}$$

Also,

$$\operatorname{var}\{\ln(\hat{s}_{x',j})\} \approx \left(\sum_{k=1}^{j-1} v_{x',k}\right) \hat{\mathscr{I}}^{-1} \left(\sum_{k=1}^{j-1} v_{x',k}\right)^{\mathrm{T}} \tag{11}$$

where $v_{x',k} = \partial(1 - \hat{s}_{x',k})/\partial\beta$ and $\hat{\mathscr{I}}$ is the Fisher's information matrix for the estimated regression parameters.

Combining (10) and (11) yields

$$\operatorname{SE}(\hat{s}_{x',j}) = \hat{s}_{x',j} \left[\left(\sum_{k=1}^{j-1} v_{x',k}\right) \hat{\mathscr{I}}^{-1} \left(\sum_{k=1}^{j-1} v_{x',k}\right)^{\mathrm{T}}\right]^{1/2}.$$

ACKOWLEDGEMENTS

The author gratefully thanks David Eaves, Bradley Efron, John Karon, Ann Lanner, James Lindsey, Glen Satten, Ian White, and Matt Zack for helpful comments made during the course of this research. Also, we thank an anonymous referee for comments that greatly enabled us to improve our manuscript.

REFERENCES

1. Jereb, J. A., Lkevens, R. M., Privett, T. D., Smith, P. J., Crawford, J. T., Sharp, V. L., Davis, B. J., Jarvis, W. R. and Dooley, S. W. 'Tuberculosis in health care workers at a hospital with an outbreak of multidrug-resistant *Mycobacterium tuberculosis*', *Archives of Internal Medicine*, **155**, 854–859 (1995).
2. American Thorasic Society. 'The tuberculin skin test', *American Review of Respiratory Disease*, **124**,(3), 356–363 (1981).
3. Huebner, R. E., Schein, M. F. and Bass, J. B. 'The turberculin skin test', *Clinical Infectious Diseases*, **17**, 968–975 (1993).
4. Turnbull, B. W. 'The empirical distribution function with arbitrarily grouped, censored, and truncated data', *Journal of the Royal Statistical Society*, *Series B*, **38**, 290–295 (1976).
5. Finkelstein, D. M. and Wolfe, R. A. 'A semi-parametric model for regression analysis of interval-censored failure time data', *Biometrics*, **41**, 933–945 (1995).
6. Finkelstein, D. M. 'A proportional hazards model for interval-censored failure time data', *Biometrics*, **42**, 845–854 (1986).
7. Self, S. G. and Grossman, E. A. 'Linear rank tests for interval-censored failure data with application to PCB levels in adipose tissue of transformer repair workers', *Biometrics*, **42**, 521–530 (1986).
8. Brookmeyer, R. and Goedert, J. J. 'Censorng in an epidemic with an application to hemophilia-associated AIDS', *Biometrics*, **45**, 325–335 (1989).
9. Odell, P. M., Anderson, K. M. and D'Agostino, R. B. 'Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model', *Biometrics*, **48**, 951–959 (1992).
10. Diamond, I. D. and McDonald, J. W. 'Analysis of current status data', in *Demographic Applications of Event History Analysis*, Clarendon Press, Oxford, 1992, pp. 231–252.
11. Shiboski, S. C. and Jewell, N. P. 'Statistical analysis of the time dependence of HIV infectivity based on partner study data', *Journal of the American Statistical Association*, **87**, (418), 360–372 (1992).
12. Dorey, F. J., Little, R. J. A. and Schenker, N. 'Multiple imputation for threshold-crossing data with interval censoring', *Statistics in Medicine*, **12**, 1589–1603 (1993).
13. Jewell, N. P., Malani, H. M. and Vittinghoff, E. 'Nonparametric estimation for a form of doubly-censored data, with application to two problems in AIDS', *Journal of the American Statistical Association*, **89**, 7–18 (1994).
14. Whitehead, J. 'The analysis of relapse clinical trials, with application to a comparison of two ulcer treatments', *Statistics in Medicine*, **8**, 1439–1454 (1989).
15. Efron, B. 'Logistic regression, survival analysis, and the Kaplan–Meier curve', *Journal of the American Statistical Association*, **83**,(402), 414–425 (1988).
16. Efron, B. and Petrosian, V. 'Survival analysis of the gamma-ray burst data', *Journal of the American Statistical Association*, **89**,(426), 452–462 (1994).
17. Neyman, J. and Scott, E. L. 'Consistent estimates based on partially consistent observations', *Econometrika*, **16**, 1–32 (1948).
18. McCullagh, P. and Nelder, J. A. *Generalized Linear Models*, 2nd edn, Chapman and Hall, New York, 1989.
19. Samuelsen, S. O. and Kongerud, J. 'Interval censoring in longitudinal data of respiratory symptoms in aluminum potroom workers: a comparison of method', *Statistics in Medicine*, **13**, 1771–1780 (1994).
20. Ades, A. E. and Nokes, D. J. 'Modeling age- and time-specific incidence from seroprevalence', *American Journal of Epidemiology*, **137**,(9), 1022–1043 (1993).
21. Aitkin, M., Anderson, D., Francis, B. and Hinde, J. *Statistical Modelling in GLIM*, Clarendon Press, Oxford, 1989.
22. Pregibon, D. 'Logistic regression diagnostic', *Annals of Statistics*, **9**, 705–724 (1981).
23. Carroll, R. J., Spiegelman, C. H., Lan, K. K. G., Bailey, K. T. and Abbott, R. D. 'On errors-in-variables for binary regression', *Biometrika*, **71**, (1), 19–25 (1984).
24. Smith, P. J. and Heitjan, D. 'Testing and adjusting for departures from nominal dispersion in generalized linear models', *Journal of the Royal Statistical Society*, *Series C*, **42**,(1), 31–41 (1993).
25. Dellaportas, P. and Smith, A. F. M. 'Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling', *Journal of the Royal Statistical Society*, *Series C*, **42**,(3), 443–460 (1993).
26. Zeger, S. L. and Karim, M. R. 'Generalized linear models with random effects; a Gibbs sampling approach', *Journal of the American Statistical Association*, **86**,(413), 79–86 (1988).
27. Hastie, R. J. and Tibshirani, R. J. *Generalized Additive Models*, Chapman and Hall, Oxford, 1990.
28. Lin, Chen-Sheng, 'A comparison of categorical models for right censored data', Ph.D. Dissertation, University of Michigan, Ann Arbor, 1994.

29. Tanner, M. A. and Wong, W. H., 'The estimation of the hazard function from randomly censored data by the Kernel method', *Annals of Statistics*, **11**, 989–993 (1983).

30. Tanner, M. A. and Wong, W. H. 'Data-based nonparametric estimation of the hazard function with applications to model diagnostics and exploratory analysis', *Journal of the American Statistical Association*, **79**, 174–182 (1984).

31. Anderson, J. A. and Senthilselvan, A. 'Smooth estimates for the hazard function', *Journal of the Royal Statistical Society*, *Series B*, **42**, 322–327 (1980).

32. Anderson, J. A. and Senthilselvan, A. 'A two-step regression model for hazard functions', *Applied Statistics*, **31**, 44–51 (1982).