STATISTICS IN MEDICINE
Statist. Med. 2004; **23**:3607–3621
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.1918

Analysis of clustered and interval censored data from a community-based study in asthma

Scarlett L. Bellamy[1,*,†], Yi Li[2,3], Louise M. Ryan[2,3], Stuart Lipsitz[4], Marina J. Canner[5] and Rosalind Wright[5]

[1]*Department of Biostatistics and Epidemiology, Center for Clinical Epidemiology and Biostatistics, 629 Blockley Hall/423 Guardian Drive, Philadelphia, PA 19103, U.S.A.*
[2]*Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, U.S.A.*
[3]*Dana-Farber Cancer Institute, 44 Binney Street, Boston, MA 02115, U.S.A.*
[4]*Department of Biometry and Epidemiology, Medical University of South Carolina, Charleston, SC 29425, U.S.A.*
[5]*Channing Laboratory, 181 Longwood Avenue, Boston, MA 02115, U.S.A.*

SUMMARY

Many authors in recent years have proposed extensions of familiar survival analysis methodologies to apply in dependent data settings, for example, when data are clustered or subject to repeated measures. However, these extensions have been considered largely in the context of right censored data. In this paper, we discuss a parametric frailty model for the analysis of clustered and interval censored failure time data. Details are presented for the specific case where the underlying time to event data follow a Weibull distribution. Maximum likelihood estimates will be obtained using commercially available software and the empirical efficiency of these estimators will be explored via a simulation study. We also discuss a score test to make inferences about the magnitude and significance of over-dispersion in clustered data settings. These methods will be illustrated using data from the East Boston Asthma Study. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: clustering; censoring; over-dispersion; survival analysis; frailty

1. INTRODUCTION

Both parametric and non-parametric methods are available for the analysis of interval censored data when observations are assumed to be independent. Classical textbooks for analysis of such time-to-event data are, for example, References [1, 2]. While methods have been developed

*Correspondence to: Scarlett L. Bellamy, Department of Biostatistics and Epidemiology, Center for Clinical Epidemiology and Biostatistics, 629 Blockley Hall/423 Guardian Drive, Philadelphia, PA 19103, U.S.A.
†E-mail: sbellamy@cceb.upenn.edu

Contract/grant sponsor: David and Lucile Packard Foundation; contract/grant number: NIMH (R01-MH61892)

Received January 2002
Accepted April 2004

for the analysis of correlated survival outcomes in settings where observations are either left or right censored, analysis methods for settings where observations are correlated and interval censored, as well as easy to implement practically, are not as well developed. Our proposed model extends the parametric failure time model to clustered and interval censored settings by introducing additive frailties to the linear predictor of the model. As presented, an additional attractive feature of this approach is that it can be readily implemented using existing commercial statistical computing software (e.g. SAS).

Our work is motivated by the East Boston Neighbourhood Health Center (EBNHC) Maternal-Child Lung (MCL) Study, a population-based longitudinal investigation of risk factors for respiratory illnesses and asthma in early childhood (see Reference [3] for a detailed study description). In the East Boston study, the time to the endpoints of interest (asthma development or other adverse respiratory outcomes) is known only to have occurred during time intervals surrounding regularly scheduled 'well-baby' clinic visits and follow-up telephone interviews. At the time of each interview, mothers or other primary care givers were asked questions about symptoms that had occurred since the last visit. For example, one question was phrased in the following way: 'Since we last saw you or spoke to you about *child's name*, have you been told by a doctor or nurse that he/she had asthma?' to determine if there had been a diagnosis of asthma since the previous contact. In this setting, observations can either be left censored (if subjects had the outcome of interest prior to first contact where questionnaire was administered) right censored (if subject never reported outcome of interest at any interview prior to their last recorded follow-up interview); or interval censored (if subject develops the outcome of interest in any interval between two successive interviews).

In addition to traditional risk factors such as maternal history of asthma and exposure to passive smoking, investigators in the East Boston Study were interested in studying the role of stress in the etiology of asthma [4]. Recent research has suggested that differential experiences of life stress may, in part, explain some of the disparities in the disease burden that exist between many racial/ethnic and socioeconomic groups in the United States. Thus investigators were interested in examining the influence of neighbourhood-level social characteristics on asthma risk. Neighbourhood-level information was obtained by geocoding residential addresses using the geographical information system (GIS) ArcView [5] and then linked to US Census data as well as the Boston Children and Families Database (BCFD) [6] which combines information from a number of city, state, federal and commercial databases from the Boston metropolitan area. Linking to BCFD provided an opportunity to create a clearer picture of neighbourhood conditions and neighbourhood disadvantage beyond the census data. For example, the inclusion of the Boston Police Department Computer Aided Dispatching (CAD) database in the BCFD, detailing the distribution of violent crimes/criminals, 911 calls and crime reports at the block-group level, offers an opportunity to explore various ecologic levels of urban stress, including violence exposure, to model the mean age to onset of asthma. Thus, the principal motivation for the extension of standard failure time methods in these analyses was to account for the clustered nature of the data in the sense that there may be common geographic exposures (at the block-group level) which may influence time to asthma development.

In this paper, we briefly describe the frailty model in the more general parametric setting where observations can be left, right or interval censored. We will derive functional forms of the general formulations in settings where we assume the underlying, unobserved failure times follow a flexible Weibull distribution. Next, we evaluate the proposed frailty model via a

simulation study and compare its performance to a naive analysis that ignores the clustering. We then formulate a score test for over-dispersion in this frailty setting and evaluate its performance via another simulation study. Finally, we present the results from an analysis of data from the East Boston Asthma Study and conclude with a brief discussion.

2. WEIBULL FRAILTY MODEL

Suppose V_{ij}, the failure time of interest for the jth subject ($j=1,\ldots,n_i$) from the ith cluster ($i=1,\ldots,N$), is known only to have occurred in an observed time interval denoted by lower and upper time points (L_{ij}, U_{ij}), respectively. Now suppose conditional on an unobserved frailty (b_i), observations from cluster i are independent. Then the likelihood contribution for the ith cluster can be expressed as the product of differences of the (conditional) survivorship functions evaluated at the observed lower and upper time point:

$$L_i(\boldsymbol{\beta}, p, \theta) = \int_{b_i} \prod_{j=1}^{n_i} \{S_{ij}(L_{ij}|b_i) - S_{ij}(U_{ij}|b_i)\} g(b_i)\, \mathrm{d}b_i \tag{1}$$

where $g(b_i)$ is the assumed density function for the unobserved frailties. Using this model formulation, both left and right censoring can be expressed as special cases of interval censoring where the observed time interval for left censored observations is $(0, U_{ij})$ and (L_{ij}, ∞) for right censored observations. Using the convention that $S_{ij}(L_{ij}|b_i) - S_{ij}(U_{ij}|b_i) = f_{ij}(L_{ij}|b_i)$ if $L_{ij} = U_{ij}$, this formulation can also accommodates exact failure times. If we assume the underlying failure times follow a Weibull (λ, p) distribution, then $S_{ij}(t|b_i) = \exp\{-H_0(t)\mathrm{e}^{-p(\boldsymbol{\beta}'\mathbf{X}_{ij}+b_i)}\}$, where $H_0(t) = (\lambda t)^p$, and $\lambda = \exp(-\beta_0)$. Here, we are using the parameterization that is consistent with the accelerated failure time (AFT) model formulation used in the LIFEREG procedure in SAS, although we could have just as easily formulated the likelihood in the proportional hazards (PH) setting by adding the frailty term to the (linear) log-hazard portion of the model. Glidden and Vittinghoff [7] propose a gamma frailty modelling approach in clustered survival settings from multicentre clinical trials via a PH model. Again, this is a nice feature of the Weibull model formulation. Glidden and Vittinghoff also support the use of frailty models (e.g. a gamma frailty) as a practical and appealing tool in clustered data settings. Note the survivorship functions are subscripted by ij to denote the possible dependence on a $(k+1)\times 1$ vector of covariates ($\mathbf{X}_{ij}$) for the jth subject in the ith cluster. Thus, our interest is estimating the corresponding vector of covariate effects ($\boldsymbol{\beta}$) corresponding to $\mathbf{X}_{ij}$. Although other distributional forms for the unobserved frailties can be assumed, for computational simplicity we assume the frailties follow a Normal($0, \theta$) distribution.

As the expression for cluster-specific likelihood contributions in (1) has no closed form solution, we will approximate this integral numerically (via Gaussian quadrature) before proceeding to maximize the marginal likelihood function. The full data likelihood is then approximated by taking the product (over i) of the cluster-specific likelihood contributions since observations are assumed to be independent across clusters (i.e. $L(\boldsymbol{\beta}, p, \theta) = \prod_{i=1}^{N} L_i(\boldsymbol{\beta}, p, \theta)$). Maximum likelihood estimates (MLEs) for $\boldsymbol{\Omega} = (\boldsymbol{\beta}', p, \theta)'$, are calculated using Newton–Raphson optimization for the corresponding (approximate) full data log-likelihood function. The Hessian matrix of second derivatives at the final step of the Newton–Raphson procedure can be inverted to estimate the variances of estimated model parameters. A reasonable set of starting values to begin the iterative optimization procedure for finding MLEs for $\boldsymbol{\Omega} = (\boldsymbol{\beta}', p, \theta)'$ can

be obtained by using standard model fitting methods (not adjusting for the possible dependence of observations within clusters) for the appropriate parameters of interest and arbitrarily choosing a starting value for the variance of the frailty term. In the Weibull setting, we can obtain starting values using the LIFEREG procedure [8] for $\boldsymbol{\beta}'$, and p and arbitrarily choosing a starting value of, say, 0.1 for θ. The LIFEREG procedure in SAS simply fits a parametric model (user specified) to a set of generally censored survival data assuming observations from different units are independent.

We wrote a model fitting macro based on the general algorithm presented previously using the integrated matrix language (IML) in SAS [9] calling the NLPNRA function to derive MLEs based on Newton–Raphson as a numerical approximation method. The NLPNRA function also facilitates appropriate restrictions on estimating parameters of interest allowing us to restrict p and θ to be positive. Alternatively, the NLMIXED procedure in SAS, Version 8 can be used to find MLEs for the proposed Weibull frailty model via the general methodology we have presented, adding to the potential application of these methods to other clustered, interval censored time-to-event data analyses. The NLMIXED procedure finds MLEs for a range of non-linear, mixed effects models by maximizing an approximization to the full data likelihood function, integrated over the random effects. Specifically, NLMIXED uses adaptive Gaussian quadrature (a slightly modified version of Gaussian quadrature which we used in our SAS macro) to approximate the cluster-specific likelihood contributions from (1). Our macro utilizes Newton–Raphson as a reasonable method of obtaining MLEs for the model parameters of interest, in contrast with NLMIXED which allows users to specify other optimization techniques (e.g. quasi-Newton methods and the Nelder–Mead simplex method) [10]. When we compared the results from our SAS macro to results obtained from the NLMIXED procedure (using 10 quadrature points), they were identical. We believe this is a nice practical result which will allow for such models to be fit in other clustered, survival settings. We have included the SAS code for fitting a very simple model to the East Boston data using NLMIXED in the appendix.

2.1. Simulation study

To evaluate the performance of our proposed model, we conducted a simulation study. First, we simulated the frailties (b_i) from a Normal$(0, \theta)$ distribution, then conditional on the frailty, we simulated three independent time points (a failure time (V_{ij}) and two observation times (L_{ij} and U_{ij})) for subject j in cluster i. The observation times were simulated from exponential distributions with different means (i.e. $L \sim \text{Exponential}(\alpha_1)$ and $U \sim \text{Exponential}(\alpha_2)$ such that $\alpha_1 < \alpha_2$) while the failure times were generated from the frailty distribution function from equation (1). If, by chance, $u < l$, then $l = \min(l, u)$ and $u = \max(l, u)$. Censoring (left, right, interval and no censoring/exact) was defined based on the sequential orientation of the simulated failure times and observation times. Failure times are left censored if $v_{ij} < l_{ij}$; right censored if $v_{ij} > u_{ij}$; interval censored if $l_{ij} < v_{ij} < u_{ij}$; and observed exactly if $l_{ij} = v_{ij} = u_{ij}$. The population parameters α_1, α_2, and λ were chosen such that approximately one-third of the observations were in each censoring pattern (left, right or interval) and no failures were observed exactly. Finally, we generated a single binary covariate from a random Bernoulli distribution. All true population parameters for the simulation study are listed in Table I, however, the primary practical interest will be to examine the properties associated with estimating the covariate effect (true $\beta = 0.3$).

Table I. Population parameter inputs for simulation study.

Parameter	Input
α_1	0.5
α_2	2.0
λ	1.5
p	1.5
β	0.3
θ	0.025, 0.25
No. of clusters	50, 100
Cluster size	3, 5

This particular combination of α_1, α_2, and λ results in approximately one-third of the observations in each of the censoring patterns (left, right or interval) for each simulated sample.

Table II shows the average estimated model parameters from the frailty model from 1000 simulated data sets. These results suggest that the estimates are close to the true population parameters of interest which is consistent with results from other clustered data settings for non-survival outcomes. The empirical standard deviation of the estimated parameters seems to become smaller as the number of clusters increases. This result is likely due to favorable asymptotic properties of MLEs (e.g. improved efficiency in larger samples). In terms of the variance (standard error) estimates associated with the proposed frailty model, the method of using the Hessian matrix at the final iteration of Newton–Raphson seems to give reasonably good variance estimates in the sense that the standard deviation of the estimators seems roughly equal to the average standard error associated with the estimator. The variability associated with estimating p (empirically) seems to be larger than the variability associated with estimating λ ($\exp(-\beta_0)$), given the same amount of data (see Table II and Figure 1). The coverage probabilities, defined as the percentage of 95 per cent confidence intervals constructed for each parameter of interest for each simulated dataset which contained the true value of the population parameter of interest, for the model parameters are presented in Table II. The coverage seems reasonable except in the simulations where the true variance of the frailties (θ) was small, but improves in simulation studies where we move away from the boundary of parameter space for θ.

From Table III we see, empirically, although the unadjusted parameter estimates appear to be fairly robust in estimating covariate effects, the unadjusted method seems to systematically under-estimate the standard errors of these estimates, again, comparing the mean standard error estimate to the empirical standard deviation of the estimator. The exception seems to be in the estimator of p in which the unadjusted method seems to systematically over-estimate the standard error of the estimator. A similar result has been observed in other clustered data settings (e.g. References [11, 12]).

It is clear from Figure 1 that inferences based on results from the unadjusted LIFEREG procedure can be misleading. What is interesting is how it seems that although the model is misspecified in the unadjusted LIFEREG setting, the estimated covariate effect seems unbiased. So if one is simply interested in inferences based on the covariate effects, the unadjusted analysis seems to provide unbiased estimates of these effects, but the associated variances are likely not valid based on the empirical results of our simulation study. The naive model which

Table II. Average estimated model parameters for (β, λ, p, and θ) and 95 per cent confidence interval coverage probabilities from Weibull frailty model.

	Clusters	Cluster size	$\hat{\beta}$ (SE) empirical SD coverage	$\hat{\lambda}$ (SE) empirical SD coverage	$\hat{p}$ (SE) empirical SD coverage	$\hat{\theta}$ (SE) empirical SD coverage
$\theta = 0.25$	50	3	0.318 (0.206) 0.227 0.927	1.514 (0.163) 0.149 0.967	1.557 (0.218) 0.242 0.945	0.242 (0.129) 0.132 0.902
		5	0.308 (0.158) 0.171 0.944	1.512 (0.141) 0.111 0.985	1.529 (0.156) 0.162 0.954	0.236 (0.095) 0.083 0.944
	100	3	0.307 (0.145) 0.160 0.923	1.507 (0.114) 0.103 0.974	1.519 (0.148) 0.160 0.935	0.229 (0.090) 0.087 0.922
		5	0.303 (0.112) 0.118 0.933	1.507 (0.100) 0.078 0.990	1.514 (0.109) 0.117 0.929	0.237 (0.067) 0.058 0.949
$\theta = 0.025$	50	3	0.305 (0.176) 0.197 0.923	1.525 (0.124) 0.136 0.922	1.585 (0.191) 0.219 0.937	0.038 (0.041) 0.056 0.541
		5	0.309 (0.137) 0.161 0.920	1.515 (0.098) 0.106 0.939	1.528 (0.138) 0.141 0.951	0.029 (0.030) 0.036 0.646
	100	3	0.311 (0.125) 0.144 0.913	1.510 (0.088) 0.093 0.938	1.539 (0.129) 0.136 0.950	0.035 (0.033) 0.041 0.646
		5	0.297 (0.097) 0.105 0.943	1.509 (0.069) 0.075 0.924	1.514 (0.096) 0.104 0.931	0.026 (0.023) 0.026 0.742

Results from 1000 simulated samples, truth: $\lambda = 1.5$, $p = 1.5$, and $\beta = 0.3$. SE based on Hessian (information) matrix from optimization procedure.

does not adjust for the frailties yields biased estimates for population parameters associated with the underlying Weibull distribution (λ and p). These results hold, in general, for other simulation scenarios we considered, with the severity of the bias in estimating parameters of interest lessening in simulations where the variance of the frailty was assumed to be small ($\theta = 0.025$). If the primary goal of inference is not estimation, but, say is prediction (i.e. estimate the survival probability at, for example 4 years, for a particular combination of covariates), then predicted survival rates based on the unadjusted model can be biased and misleading, despite the reasonable estimates of covariate effects. This is due to the fact that these predicted rates, in addition to being a function of the estimated covariate effects, are also a function of the additional parameters from the underlying failure time model, for which the unadjusted model provides biased estimates. As expected, the empirical variability associated with the unadjusted analysis (LIFEREG) is less than the empirical variability associated with the adjusted analysis. This is likely due to the extra source of variability associated with the frailties which is ignored in the unadjusted analysis. These general results hold in simulations with both 50 and 100 clusters, but empirical differences in the amount of variability in the adjusted and unadjusted analysis declines as the number of clusters increases. Again, this is

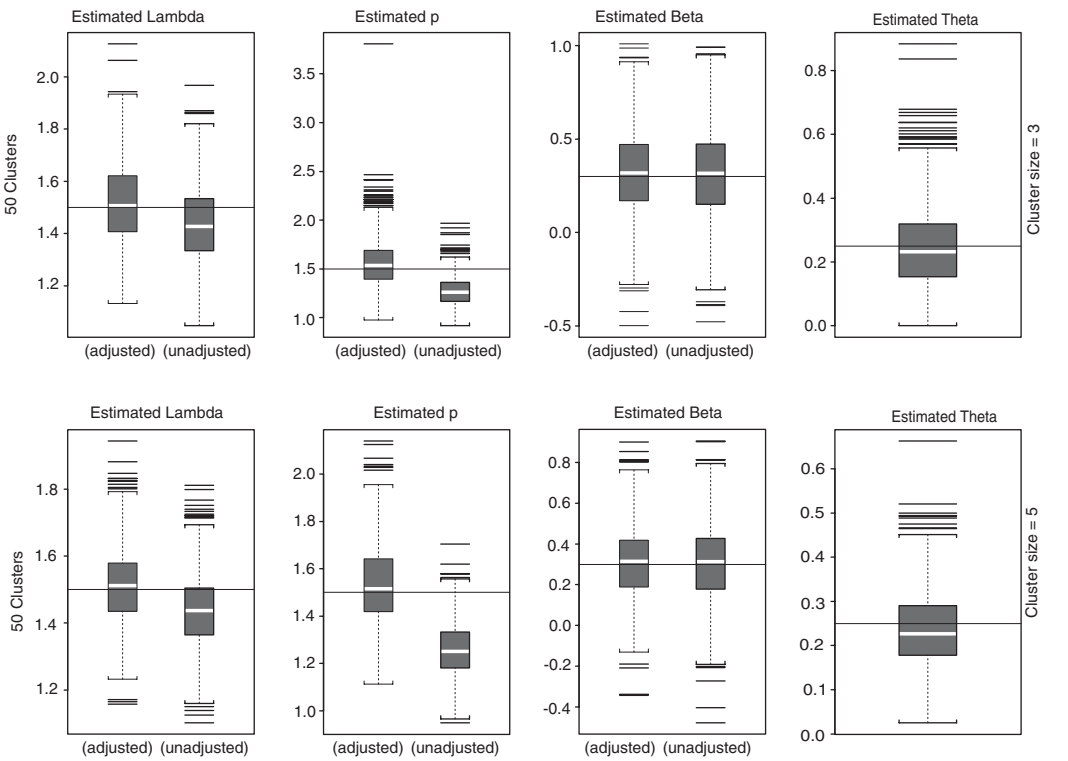

Figure 1. Empirical distribution of estimated Weibull frailty model parameters for 1000 simulated samples (50 clusters; $\theta = 0.25$). The box plots labelled 'adjusted' show results for the Weibull frailty model, while the 'unadjusted' results are from the naive, unadjusted model.

likely an asymptotic result (i.e. the variability decreases as the number of clusters increases). Finally, since the empirical standard error of our estimators seem close to the average of the estimated standard errors, we believe the estimated standard errors of the parameters of interest (based on the Hessian matrix at the final iteration of the Newton–Raphson method for finding MLEs) is reasonable.

3. SCORE TEST FOR OVER-DISPERSION

In this section, we present a score test for over-dispersion, i.e. performing the hypothesis test $H_0 : \theta = 0$ vs $H_a : \theta > 0$, where θ is the variance of the frailty. Under the null hypothesis, standard methodology which assumes independent observations are most appropriate. Because the null hypothesis lies on the boundary of the parameter space, standard likelihood-based inference is not appropriate. Lin [13] and Li and Lin [14] suggest a score test that is more suitable when testing such hypotheses that involve testing variance components. Therefore, we can derive a score test for over-dispersion (i.e. quantify the magnitude of variance component of the random effect.) using the Laplace method to approximate the integral in (1), under

Table III. Average estimated model parameters from unadjusted Weibull model.

	Clusters	Cluster size	$\hat{\beta}$ (SE) empirical SD	$\hat{\lambda}$ (SE) empirical SD	$\hat{p}$ (SE) empirical SD
$\theta = 0.25$	50	3	0.315 (0.090) 0.230	1.436 (0.102) 0.141	1.269 (0.225) 0.146
		5	0.304 (0.070) 0.183	1.436 (0.079) 0.107	1.259 (0.174) 0.111
	100	3	0.302 (0.063) 0.163	1.434 (0.072) 0.098	1.258 (0.158) 0.100
		5	0.296 (0.049) 0.125	1.432 (0.056) 0.075	1.249 (0.123) 0.074
$\theta = 0.025$	50	3	0.305 (0.076) 0.197	1.508 (0.088) 0.131	1.513 (0.196) 0.173
		5	0.309 (0.059) 0.161	1.504 (0.069) 0.104	1.482 (0.153) 0.128
	100	3	0.311 (0.054) 0.144	1.495 (0.063) 0.090	1.482 (0.139) 0.117
		5	0.297 (0.042) 0.106	1.498 (0.049) 0.075	1.473 (0.108) 0.093

Results from 1000 simulated samples, truth: $\lambda = 1.5$, $p = 1.5$, and $\beta = 0.3$.

the null hypothesis, and then compute an efficient estimate of the score vector for the model parameters of interest. More specifically, we approximate this integral using a two-term Taylor series expansion of the integrand about $b_i = 0$ (the theorized mean of the random effect, b_i). They show that such a test can be calculated using the following efficient score test statistic:

$$U_{0,i}(\boldsymbol{\Omega}_0, 0) = \left.\frac{\partial l_i}{\partial \theta}\right|_{\theta=0} = \frac{1}{2}\left\{ \left.\frac{\partial^2 \log f_i(\boldsymbol{\Omega}_0, b_i)}{\partial b_i^2}\right|_{b_i=0} + \left(\left.\frac{\partial \log f_i(\boldsymbol{\Omega}_0, b_i)}{\partial b_i}\right|_{b_i=0} \right)^2 \right\} \tag{2}$$

where $\boldsymbol{\Omega}_0 = (\boldsymbol{\beta}', p)'$.

Additionally, they propose the following efficient variance estimate for the score test statistic based on results from Cox and Hinkley [15].

$$\mathrm{Var}(U_{\theta,i}(\boldsymbol{\Omega}_0, 0)) = I_{\theta\theta,\boldsymbol{\Omega}_0} = I_{\theta\theta} - I_{\theta\boldsymbol{\Omega}_0'} I^{-1}_{\boldsymbol{\Omega}_0\boldsymbol{\Omega}_0'} I_{\boldsymbol{\Omega}_0\theta} \tag{3}$$

where

$$I_{\theta\theta} = \sum_{i=1}^{n} E\left(\left(\frac{\partial l_i}{\partial \theta} \right)^2 \right)$$

$$I_{\theta\boldsymbol{\Omega}_0'} = \sum_{i=1}^{n} E\left(\left(\frac{\partial l_i}{\partial \theta} \right) \left(\frac{\partial l_i}{\partial \boldsymbol{\Omega}_0} \right)' \right)$$

$$I_{\boldsymbol{\Omega}_0\boldsymbol{\Omega}_0'} = \sum_{i=1}^{n} E\left(\left(\frac{\partial l_i}{\partial \boldsymbol{\Omega}_0} \right) \left(\frac{\partial l_i}{\partial \boldsymbol{\Omega}_0} \right)' \right)$$

Table IV. Power/size of score test for over-dispersion in Weibull AFT model.

	Rejection rates ($\alpha = 0.05$)						
	$\theta = 0$	$\theta = 0.025$	$\theta = 0.10$	$\theta = 0.25$	$\theta = 0.50$	$\theta = 1.00$	$\theta = 1.50$
50 Clusters							
Cluster size = 3	3.80	4.45	6.85	12.30	27.85	57.75	78.00
Cluster size = 5	2.95	3.75	8.45	22.50	51.65	88.45	98.00
100 Clusters							
Cluster size = 3	4.15	3.80	7.50	20.75	48.55	87.85	98.15
Cluster size = 5	3.95	5.20	12.45	41.10	84.80	99.55	100.0

Results from 2000 simulated samples.

and

$$\frac{\partial l_i}{\partial \boldsymbol{\Omega}_0} = \begin{pmatrix} \dfrac{\partial l_i}{\partial \beta_0} \\ \vdots \\ \dfrac{\partial l_i}{\partial \beta_k} \\ \dfrac{\partial l_i}{\partial p} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial l_i}{\partial \beta'} \\ \dfrac{\partial l_i}{\partial p} \end{pmatrix}$$

This is a variance estimate of the score which accounts for additional uncertainty associated in estimating the other model parameters $\boldsymbol{\Omega}_0 = (\boldsymbol{\beta}', p)'$ and the properly standardized score has an asymptotically standard normal distribution; that is

$$Z = U_\theta(\boldsymbol{\Omega}_0, 0) \Big/ \sqrt{I_{\theta\theta, \boldsymbol{\Omega}_0}} \xrightarrow{\text{d}} \text{Normal } (0, 1)$$

We have included a discussion of the proposed score test understanding there is considerable debate in the literature as to the appropriateness of such tests in settings where data are indeed clustered/correlated (e.g. when data are clustered by design in, say, group randomized trials). This is especially true when the magnitude of the clustering is small since most tests are ill-powered to detect small levels of over-dispersion.

3.1. Simulation study to evaluate size/power of score test

In order to evaluate our proposed score test, we simulated data under a similar version of the model described previously when $\theta = 0$ and when $\theta > 0$. In this simulation study we generated a single covariate (x_{ij}) from a Normal (0,1). Recall, in the previous simulation setting x_{ij} was simulated from a random Bernoulli distribution. This slight modification to the previous model formulation is irrelevant to the performance of the score test and to the results presented in Table IV.

The first column of Table IV shows that the size of the score test is smaller than expected, especially in settings where there are a nominal number of clusters. Also, the score test only

Table V. East Boston asthma failure time statistics.

LRI Tertiles	N	Mean T_1 (SD)	Mean T_2 (SD)
Interval Censored Obs			
0	13	3.066 (1.203)	3.964 (1.370)
1	10	2.287 (1.653)	2.900 (1.721)
2	49	1.781 (1.450)	2.414 (1.567)
Right Censored Obs			
0	178	4.371 (1.863)	—
1	141	4.467 (1.696)	—
2	134	4.713 (1.549)	—

appears to achieve modest power when there are either a large number of clusters or in settings where the level of over-dispersion is large.

To further examine the asymptotic size properties of the proposed score test statistic, which seem to be quite small for smaller numbers of clusters (see Table IV), we increased the number of clusters to 200, 500 and 1000 clusters and repeated the simulation study (under the null hypothesis; $\theta = 0$ and for 2000 simulated samples). The simulation study indicates that the proposed score test has a type I error rate that remains smaller than expected, even when the total number of clusters was increased to 200 and 500 clusters (4.55 and 3.68 for 200 clusters and cluster size equal to 3 and 5, respectively; and 3.88 and 3.75 for 500 clusters and cluster size equal to 3 and 5, respectively). The type I error rate does not reach the desired level (≈ 5 per cent) until the number of clusters is quite large (1000 clusters and cluster size $= 3$). This phenomenon has been observed previously in other likelihood based, variance components testing settings (see, for example, Reference [16]). Further research to develop tests with better size properties could, therefore, be a valuable contribution to the literature.

4. EAST BOSTON ASTHMA STUDY ANALYSIS

We return now to the application which motivated the proposed Weibull frailty model formulation. Participants in the East Boston Asthma Study were largely white (50 per cent) and Hispanic (44 per cent) children (49 per cent males and 51 per cent females) who range in age from infancy to 6 years old. In this subset of data, we were interested in exploring the relationship between mean time to asthma onset and the number of lower respiratory tract infections the children experience in their first year of life (LRI). There were 753 total observations with number of lower respiratory tract infections ranging from 0 to 16. All observations were either right or interval censored. For each analysis (unadjusted and adjusted), we used the subset of complete data for LRI and census block-group information from 525 observations in 52 census block-groups; we observed 72 cases of asthma in this subset ($72/525 = 13.7$ per cent). Table V summarizes the empirical failure times (physician diagnosis of asthma) for the complete subset of data.

The empirical results presented in Table V seem fairly consistent with the current pediatric asthma literature as half of all cases of asthma are diagnosed by age 3, and two-thirds are diagnosed by age 5 [17, 18]. Empirically, the mean time to physician diagnosis of asthma

Table VI. East Boston asthma study results.

	Unadjusted analysis		Adjusted analysis	
	Estimate	SE	Estimate	SE
Intercept	3.834	0.276	3.877	0.290
LRI	−0.275	0.041	−0.288	0.044
Scale	0.906	—	0.892	0.096
θ	—	—	0.119	0.159

seems to systematically decrease as the number of respiratory tract infections increases. This empirical pattern of increasing asthma morbidity as a function of increased number of respiratory tract infections in the first year of life appears to be consistent in both the unadjusted and adjusted analyses.

Table VI indicates only a small to modest level of over-dispersion ($\hat{\theta} = 0.119$) which we did not find to be statistically significant based on the computation of the proposed score test presented previously ($p = 0.77$). Based on our simulations, however, it is possible that the test lacks adequate power to detect a true effect, given the size of our study (52 census block-groups) and the estimated modest level of over-dispersion. To address the potential bias of the proposed score test this small sample settings (i.e. small number of clusters), we constructed a bootstrap estimate of the p-value associated with the score test for over-dispersion for the East Boston data [19]. For the bootstrap analysis, we took 1000 bootstrap re-samples of the East Boston data set (re-sampling individuals and all corresponding covariates). For each of these 1000 re-samples, we computed a score test statistic and computed a bootstrap p-value based on a permutation test of bootstrapped re-samples. Specifically, the bootstrap p-value is the proportion (out of the 1000 bootstrap re-samples) of score test statistics that were greater than the computed score test statistic from the original sample. In the case of East Boston, the bootstrap p-value $= 0.951$.

The highly significant regression coefficient associated with LRI in the unadjusted analysis remains highly significant in the adjusted analysis ($p < 0.001$). Because this coefficient is negative, both models suggest the mean time to asthma diagnosis decreases as the number of LRIs increase. This result is consistent with current pediatric asthma literature and with the observed empirical results presented in Table V. The magnitude of this coefficient is very similar in both the adjusted and unadjusted analysis. This result is consistent with the results we observed for covariate effects in our simulation study. One can see that even in this simple analysis, the unadjusted method seems to under-estimate the variance associated with the estimates of the model parameters of interest. Finally, since the level of over-dispersion observed in the data was not statistically significant, the estimates for the parameters in the model are quite similar. Again, this result seems to be consistent with the results observed in our simulation study.

For illustration purposes only, we estimated predicted survival probabilities for children (1–5 years old) with varying number of respiratory tract infections in their first year of life (0, 5, 10, and 15). In actuality, there are likely a number of additional covariates which influence the asthma diagnosis outcomes that we have not considered in our models. However, our simple model serves to illustrate the practical application of the proposed methodology as a prediction tool in clustered survival settings with interval censored survival times. The predicted survival

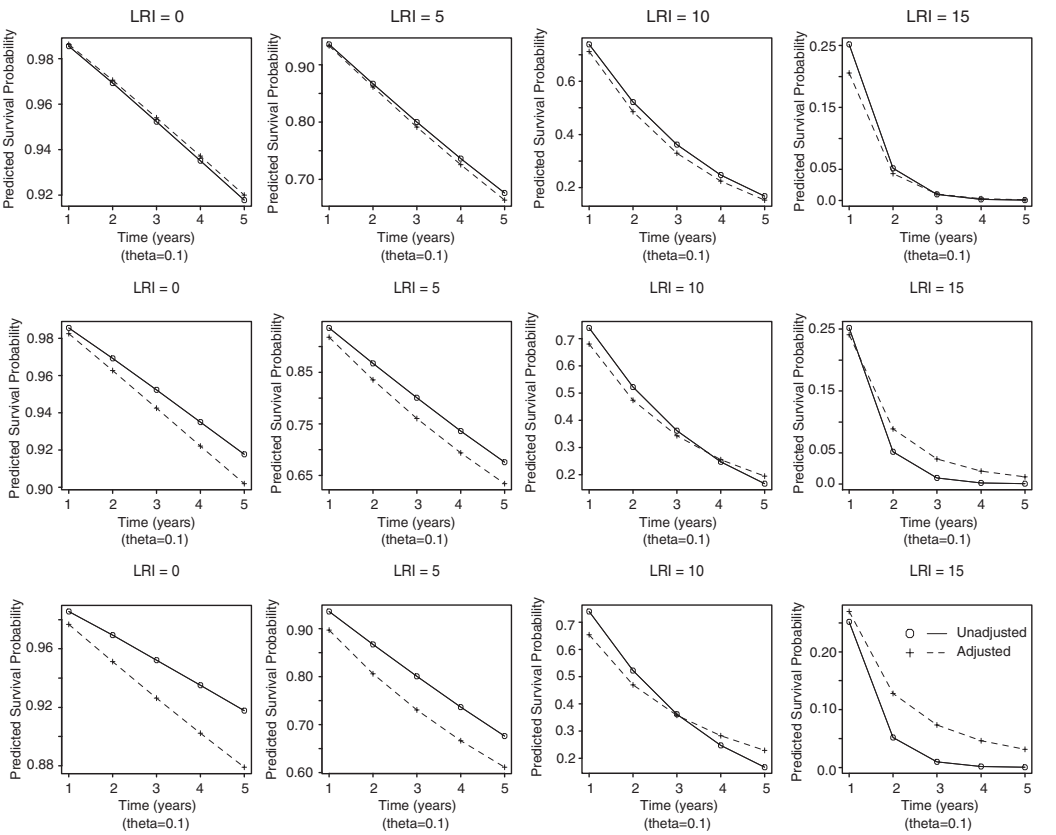

Figure 2. Predicted survival probabilities at ages 1–5 for unadjusted and adjusted model estimates presented in Table VI. Columns vary the covariate LRI (number of lower respiratory tract infections in the first year). Rows vary θ (variance of the normal frailty).

probabilities for the adjusted model were obtained by assuming the estimated model parameters (Table VI) were fixed and known and then using the QUAD call function in SAS/IML to numerically approximate (Gaussian quadrature) cluster-specific likelihood contributions of the survivorship function. Figure 2 is a plot of these estimated survival probabilities based on the parameter estimates from both the unadjusted (LIFEREG) analysis and the adjusted frailty (NLMIXED) analysis presented in Table VI. The top row of survival curves assumes the variance of the frailties is equal to 0.10 (roughly equivalent to the estimated frailty variance from the adjusted Weibull frailty model). From these plots we can see that in this case where there is only a small level of over-dispersion ($\theta = 0.10$), the predicted survival probabilities from the unadjusted and adjusted analyses are roughly equivalent. However, as the level of over-dispersion increases the differences in the survival curves becomes more pronounced. This is especially true, as indicated by the widening gap between the unadjusted and adjusted survival curves, for fewer numbers of lower respiratory tract infections in the first year of life (e.g. LRI = 0 and 5). From the estimated survival curves there appears to be a crossing of the curves as the number of LRIs increases (e.g. LRI = 10 and 15). So although the survival

probability looks poorer in the unadjusted analysis, initially, the survival probability becomes more optimistic over time for this model compared to the unadjusted model.

5. DISCUSSION

In this paper we have extended the parametric model formulation for time-to-event data to accommodate a frailty term in clustered data settings where observations are left, right and interval censored. Adding the frailty term to the model seemed a natural extension of the parametric method of analysis. Although we have motivated these ideas in the context of a Weibull distribution for the underlying failure times of interest, these methods can be easily adapted to other parametric forms for both the underlying assumed conditional distribution of the failure times of interest and for the assumed distribution of the frailties. The Weibull is a flexible model for survival data for a variety of reasons including the fact that it can be formulated as both an accelerated failure time model and proportional hazards model. In more general AFT settings, one can accommodate a frailty to the linear predictor part of the model, while in more general PH settings, the frailty can be added to the log-hazard ratio making this frailty approach quite general and easily extendible to other parametric AFT and PH model formulations.

We presented detailed algorithms for obtaining parameter estimates for the model of interest as well as methods for obtaining variance/standard error estimates for the parameters of interest. We also formulated a score test for over-dispersion in this censored, clustered data setting. The performance of the model fitting algorithm and the score test were evaluated via simulation studies.

The fact that we can use the NLMIXED procedure in the latest version of SAS to obtain parameter estimates and variance estimates is an attractive feature of our proposed method and has greatly broadened the range of frailty models that can be fit to interval censored survival data. In this parametric setting, nested models can be compared via their likelihoods to assess model fit and non-nested models can be compared via other popular regression diagnostics, say AIC and BIC (e.g. Reference [20]). The SAS procedure NLMIXED is quite useful in accommodating a wide range of theorized parametric models. One simply needs to formulate the (log)likelihood function and specify a parametric form for the frailties.

In both the simulation study and the analysis of the East Boston Asthma Study, we observed fairly robust estimates of covariate effects in the unadjusted model despite the model misspecification and the variance associated with these estimates seem to be under-estimated. This result has been observed in other non-linear mixed model settings. Further theoretical exploration of this result would be a valuable contribution to the clustered, survival literature. On the other hand, our proposed Weibull frailty model not only seems to provide consistent estimates of the population parameters of interest, but also seems to provide fairly accurate estimates of the variance associated with these estimates as well comparing the average of the standard errors associated with the population parameter estimates to the empirical standard deviation of the estimates. So the method of using the information matrix at the final step of Newton–Raphson seems to be a reasonable method of computing variances for the parameter estimates.

The proposed score test for over-dispersion appears to be biased in small sample settings (small numbers of clusters) and does not achieve the expected size until the number of clusters becomes quite large. Others have proposed various forms of the likelihood ratio test for over-

dispersion as an alternative to the score test presented here, however we believe the score test is computationally simpler since one only needs to compute functions of the likelihood under the null hypothesis. Specifically, the likelihood ratio test requires additional computations of the likelihood under the alternative hypothesis. The observed bias and non-optimal size of the score test is an interesting result that deserves further theoretical exploration. Finally, since the estimated level of over-dispersion was low in the East Boston data, the power to detect such a small level of over-dispersion was likely compromised by the small number of clusters (i.e. there were only 52 census block-groups represented in the subset of data from the study).

APPENDIX A

A.1 SAS NLMIXED code

```
proc nlmixed data=asthma qpoints=10;
   parms beta0=3.91953 beta1=-0.28539 p=0.92786 theta=0.01;
   bounds p > 0, theta > 0;

   ebetaxb=exp(-(beta0 + beta1*1ri1 + b));
   lambda=exp(-beta0);
   s_1 = exp(-(t1*ebetaxb)**(1/p));
   s_u = exp(-(t2*ebetaxb)**(1/p));
   f_t = ((lambda*p)*(lambda*t1)**(p-1))*ebetaxb**1/p;

   /* ctype (censoring type): 1=exact 2=left 3=right 4=int */
   if ctype=1 then lik = f_t;
      else if ctype=2 then lik = 1 - s_u;
         else if ctype=3 then lik = s_1;
            else lik = s_1 - s_u;

   llik=log(lik);
   model y ~ general(llik);
   random b ~ normal(0,theta) subject=clusidz;
run;
```

A.2 Results

The parameter estimates are shown in Table AI.

Table AI. Parameter Estimates.

Parameter	Estimate	Standard error	DF	*t* value	Pr > |*t*—	Lower	Upper
beta0	3.8768	0.2895	51	13.39	<0.0001	3.2956	4.4580
beta1	-0.2875	0.04362	51	-6.59	<0.0001	-0.3751	-0.1999
p	0.8915	0.09619	51	9.27	<0.0001	0.6983	1.0846
theta	0.1188	0.1588	51	0.75	0.4576	-0.1999	0.4375

ACKNOWLEDGEMENTS

Funding for this work by the first author was provided by the David and Lucile Packard Foundation as well as NIMH (R01-MH61892). This paper was improved tremendously by the thoughtful comments and suggestions of its independent reviewers.

REFERENCES

1. Cox DR, Oakes D. *Analysis of Survival Data*. CRC Press: London, 1984.
2. Collett D. *Modelling Survival Data in Medical Research*. Chapman & Hall: London, 1994.
3. Hanrahan J, Tager I, Segal M, Tosteson T, Castille R, Vunakis HV, Weiss S, Speizer F. The effect of maternal smoking during pregnancy on early infant lung function. *American Review of Respiratory Disease* 1992; **145**:1129–1135.
4. Wright RJ, Rodriguez M, Cohen S. Review of psychosocial stress and asthma: an integrated biopsychosocial approach. *Thorax* 1998; **53**:1066–1074.
5. Environmental Systems Research Institute, 1996.
6. Carlton S. *A Users Guide for the Boston Children and Families Database.* Boston Persistent Poverty Project, The Boston Foundation, 1995.
7. Glidden DV, Vittinghoff E. Modelling clustered survival data from multi-centre clinical trials. *Statistics in Medicine* 2004; **23**:369–388.
8. SAS Institute Inc. *SAS/STAT Users Guide*, *Version 6*, (4th edn.) SAS Institute Inc.: Cary, North Carolina, 1990.
9. SAS Institute Inc. *SAS/IML Software*: *Changes and Enhancements*, *through Release 6.11*. SAS Institute Inc.: Cary, North Carolina, 1995.
10. SAS Institute Inc. *SAS/STAT Users Guide*, *Version 8*. SAS Institute Inc.: Cary, North Carolina, 1999.
11. Bellamy SL, Gibberd R, Hancock L, Howley P, Kennedy B, Klar N, Lipsitz S, Ryan L. Analysis of dichotomous outcome data for community intervention studies. *Statistical Methods in Medical Research* 2000; **9**:135–159.
12. Zeger SL, Liang K-Y, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988; **44**:1049–1060.
13. Lin X. Variance component testing in generalized linear models with random effects. *Biometrika* 1997; **84**: 309–326.
14. Li Y, Lin X. Testing random effects in uncensored/censored clustered data with categorical responses. *Biometrics* 2003; **59**:25–35.
15. Cox DR, Hinkley DU. *Theoretical Statistics.* Chapman & Hall: London, 1974.
16. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993; **80**:27–38.
17. Barbee R, Dodge P, Lebowitz M, Burrows B. The epidemiology of asthma. *Chest* 1985; **87**:21S–25S.
18. Croner S, Kjellman N. Natural history of bronchial asthma in childhood. *Allergy* 1992; **47**:150–157.
19. Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* 1986; **1**:54–77.
20. Belsley DA, Kuh E, Welsch RE. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley: New York, 1980.
21. Adler NE, Boyce WT, Chesney MA, Folkman S, Syme SL. Socioeconomic inequalities in health: no easy solution. *Journal of the American Medical Association* 1993; **269**:3140–3145.
22. Attar BK, Guerra NG, Tolan PH. Neighborhood disadvantage, stressful life events and adjustment in urban elementary-school children. *Journal of Clinical Child Psychology* 1994; **23**:391–400.
23. Diez-Roux A. Bringing context back into epidemiology: variables and fallacies in multi-level analysis. *American Journal of Public Health* 1998; **88**:216–222.
24. Donner A, Klar N. Statistical considerations in the design and analysis of community intervention trials. *Journal of Clinical Epidemiology* 1996; **49**:435–439.
25. Kennedy JE, Gentle WJ. *Statistical Computing*. Dekker: New York, 1980.
26. Manatunga AK, Oakes D. Parametric analysis for matched pair survival data. *Lifetime Data Analysis* 1999; **5**:371–387.
27. Pearce N, Beasley R, Burgess C, Crane J. *Asthma Epidemiology*: *Principles and Methods*. Oxford University Press: Oxford, 1998.
28. Piekett KE, Pearl M. Multi-level analyses of neighborhood socioeconomic context and health outcomes: a critical review. *Journal of Epidemiology and Community Health*, in press.
29. Smith LA, Finkelstein JA. The impact of sociodemographic factors on asthma. In *Asthma's Impact on Society: The Social and Economic Burden*, Weiss KB, Buist AS, Sullivan SD (eds). Marcel Dekker, Inc.: New York, 2000; 219–243.
30. Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society*, *Series B* 1976; **38**:290–295.

