Journal of Medical Statistics and Informatics
ISSN 2053-7662 | Volume 4 | Article 1

HOAJ
Herbert Open Access Journals

Original Research Open Access

A conditional frailty model for bivariate interval-truncated failure time data: an application to a study on siblings diagnosed with schizophrenia

Rinku Sutradhar[1,2*] and Richard J. Cook[3]

*Correspondence: rinku.sutradhar@ices.on.ca

[1]Institute for Clinical Evaluative Sciences, Toronto, Ontario M4N 3M5, Canada.
[2]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario M5T 3M7, Canada.
[3]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada.

Abstract

Background: Paired failure time data often arise in medical studies involving familial information. When each member of a pair is subject to interval-truncation, there is lack of literature on methodologies for analyzing such data. The aim of this paper is to develop an approach for examining the association between paired failure times under the presence of interval-truncation.

Methods: A conditional frailty model is described and an expectation-maximization algorithm is developed for estimation of the model parameters. Simulation studies are conducted to examine the performance of the proposed algorithm, and an application of the methods is illustrated using data from a familial study on pairs of siblings diagnosed with schizophrenia.

Results: The results from the simulation studies show that all model parameters can be estimated with negligible bias, even under finite sample size. In the motivating dataset, our proposed model and method of estimation reveal a strong presence of dependence among times to diagnosis of schizophrenia within pairs of siblings.

Discussion: The proposed conditional frailty model is easy to implement since the expectation step of the algorithm is relatively straightforward. The maximization step also provides closed form expressions for the parameter estimates of the hazard function.

Keywords: Frailty model, interval-truncation, bivariate failure time data, monte carlo EM algorithm

Introduction

Interval truncated failure time data arise when only those individuals with a failure time that lies within a certain interval are observed. When a failure time falls outside this truncation interval, no information regarding the subject is available. One may view this as a screening procedure in which all observations outside the interval known as the truncation region are removed. It is the concept of "who is included" that distinguishes truncation from censoring. In the case of censoring, we have partial information on all observations in a sample in that subjects are known to have failure times greater (or less) than some value. However in the case of truncation, individuals with failure times outside the truncation interval are not considered for inclusion in the study. This important difference between truncation and censoring is reflected in the construction of the data likelihood.

Approaches for dealing with censoring in the presence of interval truncation were first discussed by Turnbull [1]. He proposed a method for obtaining the nonparametric maximum likelihood estimate of the distribution function under interval truncation that also accommodates left, right, and interval censoring. Turnbull described a self-consistent approach to estimation that turns out to be a version of the EM algorithm [2]. Frydman noted that the method proposed by Turnbull was not applicable in the case of both truncation and censoring, and suggested appropriate modifications [3]. Alioum and Commenges provided a further detailed correction of Turnbull's approach and offered an extension to regression analysis [4].

We are interested developing methodology for examining interval-truncated failure times in a bivariate setting. That is, when available information consists of paired failure times (e.g.,

doi: 10.7243/2053-7662-4-1

we consider families of size two), where, due to the observation scheme, paired failure times are known only if they both fall within their respective truncation intervals. Such data arises in settings where families are observed through family disease registries, for example, or where data are grouped by family by design. When paired data are collected subject to truncation, it is frequently of interest to investigate associations among failure times within families. The existing literature on methods for analyzing interval-truncated failure times in a bivariate setting is very limited and mostly refers to the case when only one component of the bivariate vector is subject to truncation. Exceptions to this assumption include the studies by van der Laan and Gurler [5,6], in which nonparametric estimation of the bivariate survival function is discussed where both components of the bivariate vector are randomly left or right-truncated, but not interval-truncated. This paper proposes a model and estimation approach for examining interval-truncated bivariate failure time data. Straight forward extensions are possible to deal with the multivariate setting.

Methods

Notation

Suppose a pair of individuals is included in a study if both members' event times are observed to fall in the calendar time interval $[L,R]$. Let Y_{ij} represent the year of birth for the jth member in the ith family and let X_{ij} denote the year of event for that individual. Let the age at the year of the event for the jth subject of the ith pair be denoted by $T_{ij}=X_{ij}-Y_{ij}$, and let $B_{ij}=[L_{ij},R_{ij}]=[L-Y_{ij},R-Y_{ij}]$ be the corresponding truncation interval in terms of age, where L_{ij} is the age at calendar time L and R_{ij} is the age at calendar time R. Although we observe T_{ij} if it belongs to its corresponding truncation interval B_{ij}, most information regarding familial association is available if two or more event times within a family meet truncation requirements.

For the ith family with two members or siblings j=1,2, under the assumption of independent truncation, the joint failure times are sampled from a conditional distribution.

$$F\left(t_{i1},t_{i2}|\,T_{i1}\in B_{i1},T_{i2}\in B_{i2}\right)=P\left(T_{i1}\le t_{i1},T_{i2}\le t_{i2}|\,T_{i1}\in B_{i1},T_{i2}\in B_{i2}\right). \quad (1)$$

The likelihood for bivariate truncated failure times with observed data (t_{ij}, $B_{ij}=[l_{ij},r_{ij}]$; j=1,2, i=1,...n) can be expressed simply in terms of the bivariate survivor function:

$$L=\prod_{i=1}^{n}f\left(t_{i1},t_{i2}|\,T_{i1}\in B_{i1},T_{i2}\in B_{i2}\right)=\prod_{i=1}^{n}\left\{\frac{1}{P(T_{i1}\in B_{i1},T_{i2}\in B_{i2}}\times\frac{\partial^2}{\partial t_{i1}\partial t_{i2}}S^*(t_{i1},t_{i2})\right\}, \quad (2)$$

where

$$S^*\left(t_{i1},t_{i2}\right)=P(T_{i1}\ge t_{i1},T_{i2}\ge t_{i2})$$

and

$$P\left(T_{i1}\in B_{i1},T_{i2}\in B_{i2}\right)=S^*(l_{i1},l_{i2})-S^*(l_{i1},r_{i2})-S^*(r_{i1},l_{i2})+S^*(r_{i1},r_{i2}),$$

is the probability of inclusion in the truncation rectangle $B_i=B_{i1}XB_{i2}$.

Model formulation: joint distributions using a frailty model

A conditional formulation is a common approach adopted for modeling bivariate failure time data. Joint distributions are routinely formulated by assuming responses are conditionally independent given a common scalar random effect. Integrating the conditional joint distribution with respect to the unobserved random effect provides the joint distribution of the failure times. Suppose (T_1,T_2) are bivariate failure times for a particular pair. The joint survivor function under a conditional model can be written as

$$S^*\left(t_1,t_2\right)=P\left(T_1>t_1,T_2>t_2\right)=\int S_1(t_1|\alpha)S_2\left(t_2|\alpha\right)dG\left(\alpha;\phi\right), \quad (3)$$

and the marginal survivor function for T_j is

$$S_j^*(t)=P\left(T_j>t\right)=\int S_j(t|\alpha)\,dG\left(\alpha;\phi\right),j=1,2. \quad (4)$$

Paired failure times T_1 and T_2 are independent given the pair-specific random effect α, which has distribution $G(\cdot)$ indexed by parameter ϕ.

Various models have been proposed in the literature for the distribution of the random effects: one-parameter gamma distribution [8], the log-normal distribution [13], the positive stable distribution [9], and the inverse gamma distribution [10]. We assume the random effects arise from a one-parameter gamma distribution with mean 1 and variance ϕ^{-1}. Furthermore, consider a frailty model in which random effects act multiplicatively on the hazard rate for each member of a family. Conditional on the random effect, the survivor function for T_j is $S_j\left(t|\alpha\right)=e^{-\alpha\Lambda_{0j}^*(t)}$, where $\Lambda_{0j}^*(t)$ is the cumulative baseline hazard function for T_j (j=1,2) under the conditional model. These assumptions provide closed form expressions for the joint and marginal survivor function, respectively.

$$S^*(t_1,t_2)=\int\exp\{-\alpha\Lambda_{01}^*(t_1)\}\exp\{-\alpha\Lambda_{02}^*(t_2)\}\frac{\phi^\phi\alpha^{\phi-1}e^{-\phi\alpha}}{\Gamma(\phi)}\,d\alpha=\left(1+\frac{\Lambda_{01}^*(t_1)}{\phi}+\frac{\Lambda_{02}^*(t_2)}{\phi}\right)^{-\phi} \quad (5)$$

and

$$S_j^*(t)=\int\exp\{-\alpha\Lambda_{0j}^*(t)\}\frac{\phi^\phi\alpha^{\phi-1}e^{-\phi\alpha}}{\Gamma(\phi)}\,d\alpha=\left(1+\frac{\Lambda_{0j}^*(t)}{\phi}\right)^{-\phi},\ j=1,2. \quad (6)$$

Note that the marginal survivor function $S_j^*(t)$, under the conditional formulation, depends on both the parameter ϕ and the parameters of the conditional baseline hazard function $\Lambda_{0j}^*(t)$. Moreover, as $\Phi\to\infty$, $S_j^*(t)$ approaches $e^{-\Lambda 0j^*(t)}$ for j=1,2 and $S^*(t_1,t_2)$ approaches $e^{-(\Lambda_{01}^*(t)+\Lambda_{02}^*(t))}$. Thus, as the variance of the random effect distribution tends to zero, $S^*(t_1,t_2)$

doi: 10.7243/2053-7662-4-1

$= S_1^*(t_1)S_2^*(t_2)$, indicating that T_1 and T_2 are independent.

It is important to interpret the parameters of the model appropriately. Parameter ϕ measures both the lack of fit of the conditional hazard function, as well as the association among paired failure times. Large values of ϕ could indicate that standard survival models may be inadequate by failing to sufficiently account for variation in the data; this can arise simply from poor fit of a model. That is, large values of ϕ may be an indication of unaccounted for dependence, overdispersion, or other forms of model misspecification including violations of the proportional hazards assumption. Moreover, as the variance approaches zero, failure times among individuals within a pair approach independence.

Proposed estimation approach: an EM algorithm

Methodological complications arise when both failure times in a pair are observed subject to interval truncation. The inclusion of shared random effects among individuals within a pair, in addition to the presence of interval truncation for each member in the pair, results in a non-trivial likelihood function. Direct maximization of this likelihood function requires integrating the conditional joint distribution with respect to the unobserved random effect, which is not straightforward particularly if strong distributional assumptions are to be avoided. In addition, if a piecewise constant model is used, direct maximization becomes more challenging as the number of pieces (and hence parameters) increases. Instead of trying to directly maximize over a high dimensional space, the EM algorithm offers an alternative method of estimation that is sometimes slow to converge but avoids occasional divergence problems. We develop an EM algorithm for bivariate interval-truncated failure times under a conditional (random effect) formulation. We use the concept of "ghosts", which was first discussed by Turnbull [1] with the aim of obtaining a nonparametric estimate of a truncated distribution in the univariate setting. Here the notion of "ghost pairs" is used for the analysis of interval-truncated failure times in the bivariate case.

Let $t_i=(t_{i1},t_{i2})$ denote the vector of observed failure times for the subjects of the ith pair and let $B_i=(B_{i1}\times B_{i2})=[L_{i1},R_{i1}]\times[L_{i2},R_{i2}]$ be the corresponding truncation rectangle. The observed data is represented as $X=(X_1,\ldots,X_n)$ with $X_i=(t_i,B_i)$ being the observed data for the ith pair (i=1,…,n). To define the complete data, we focus on the complement of the bivariate truncation rectangle B_i^c. **Figure 1** provides an illustration of the data for a particular pair of individuals. The truncation region is represented by the center rectangle and the point t_i is observed because both $T_{i1}\in B_{i1}$ and $T_{i2}\in B_{i2}$.

Due to truncation, the pair giving the observations $t_i=(t_{i1},t_{i2})$ may be considered a remnant of an unknown number of pairs from the same "birth cohort". Pairs are said to be in the same birth cohort if they have the same dates of birth. We define "ghost pairs" corresponding to pair i as pairs of individuals in the same birth cohort who were not observed because their failure times did not fall in the truncation region B_i. Thus these "ghost pairs" would have been observed had there been no truncation.

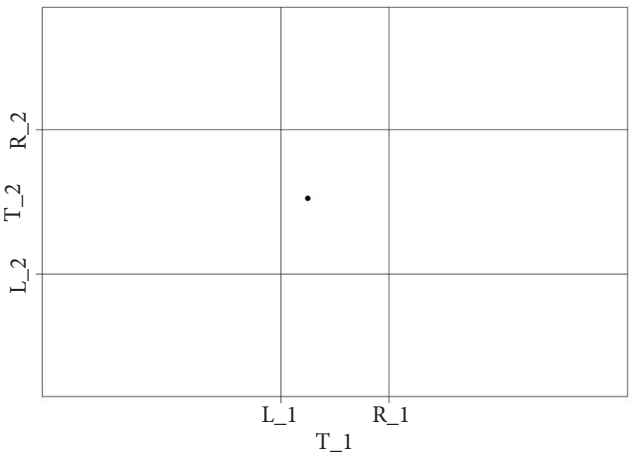

Figure 1. Bivariate truncation rectangle for a pair of failure times.

To define the complete data, let J_i be the number (unknown) of ghost pairs for the ith pair of observed event times. Suppose $u_i=(u_{i1},u_{i2})$ for $j=1,\ldots,J_i$ are the corresponding unknown failure times for these ghost pairs. These are the paired times that fall outside the truncation rectangle B_i for the ith observed pair, and are hence considered to arise from the unobserved ghosts. Times within an observed pair or ghost pair are considered independent given their corresponding random effect, arising from the distribution $G(\cdot)$ indexed by parameter ϕ. Here the complete data is given by $Y=(Y_1,\ldots,Y_n)$ with $Y_i=(X_i,\alpha_i,J_i,\{u_j,\alpha_j;j=1,\ldots,J_i\})$ being the complete data for the ith pair. Notice here we consider the times of the events for the ghost pairs, as well as the random effects for the observed pair and corresponding ghost pairs, as part of the complete data. The reason for this is that the logarithm of the resulting complete data likelihood is relatively convenient to work with and, as will be seen, maximization is particularly straight forward.

The complete data likelihood and corresponding log-likelihood for the ith pair of individuals are constructed as

$$L_{ci}\left(\theta;\mathbf{Y}_i\right)=f\left(t_{i1}|\alpha_i\right)f\left(t_{i2}|\alpha_i\right)g\left(\alpha_i\right)\times\prod_{j=1}^{J_i}\{f\left(u_{j1}|\alpha_j\right)f(u_{j2}|\alpha_j)g(\alpha_j)\} \quad (7)$$

and

$$l_{ci}\left(\theta;\mathbf{Y}_i\right)=\log f(t_{i1}|\alpha_i)+\log f\left(t_{i2}|\alpha_i\right)+\log g\left(\alpha_i\right)+\sum_{j=1}^{J_i}\{\log f(u_{j1}|\alpha_j)+\log f(u_{j2}|\alpha_j)+\log g\left(\alpha_j\right)\}, \quad (8)$$

 doi: 10.7243/2053-7662-4-1

respectively, where

$$f(t|\alpha)=\alpha\lambda_0^*(t)\exp\{-\alpha\int_0^t\lambda_0^*(w)dw\}. \quad (9)$$

We further assume a gamma distribution for the random effect distribution with p.d.f. given as

$$g(\alpha)=\frac{\phi^\phi}{\Gamma(\phi)}\alpha^{\phi-1}e^{-\phi\alpha},$$

where $E(\alpha)=1$ and $Var(\alpha)=\phi^{-1}$. Based on these assumptions, the complete data log-likelihood can now be re-expressed -the parameters of this function consist of the parameters of the baseline hazard functions and the parameter of the random effects distribution. These parameters will be estimated via an Expectation-Maximization algorithm, where a weakly parametric approach will be taken by assuming a piecewise constant model for the baseline hazard function. The mathematical details of this algorithm are provided in the Appendix. We show in the maximization step that the parameter estimates of the baseline hazard function have a closed simple form and the parameter estimate of the random effects distribution can be easily obtained using standard Newton-Raphson procedures.

Evaluating performance of proposed methodology via simulations

We conducted simulations for investigating the performance of our proposed EM algorithm. We generated random effects $\alpha_1,\ldots,\alpha_n$ from a gamma distribution with p.d.f.

$$g(\alpha)=\frac{\phi^\phi\alpha^{\phi-1}e^{-\phi\alpha}}{\Gamma(\phi)},$$

where $E(\alpha)=1$ and $Var(\alpha)=\phi^{-1}$. Bivariate failure times (T_1,T_2) were simulated from the joint distribution

$$S^*(t_1,t_2)=\int S(t_1|\alpha)S(t_2|\alpha)dG(\alpha),$$

with

$$S(t|\alpha)=e^{-\alpha(\psi t)^\gamma},$$

where the cumulative baseline hazard function $\Lambda_0^*(t)=(\psi t)^\gamma$ had a Weibull form with shape parameter γ and scale parameter ψ^{-1}. We generated n=1000 pairs of failure times. The median of the marginal distribution

$S^*(t)=\int S(t|\alpha)dG(\alpha)=[1+(\psi t)^\gamma/\phi]^{-\phi}$ was specified at n=40. Note that T_1 and T_2 were taken to have the same marginal distribution. The true value of the parameter ϕ was assumed to be 1/0.25 and 1/0.75 ($\log\phi$=1.386 and 0.288, respectively), and the true value of the shape parameter γ was taken to be 1.0 and 1.2. Each failure time was truncated over the same interval $B=[L,R]$. The left endpoint L was taken to be 0.0 and the right endpoint R was calculated at the 95th and 90th percentiles of the true marginal distribution. In summary, we generated data under 8 scenarios, each of which was simulated M=100 times.

The value of ψ was expressed in terms of ϕ and γ. It was obtained from the marginal survival function as follows:

$$0.5=S^*(\eta)=[1+\frac{(\psi\eta)^\gamma}{\phi}]^{-\phi},$$

which gave

$$\psi=\left[\frac{(0.5^{-1/\phi}-1)\phi}{\eta^\gamma}\right]^{1/\gamma}.$$

The truncation percentiles were calculated in a similar manner. They depended on both parameters ϕ and γ:

$$R=\left[\frac{((1-p)^{-1/\phi}-1)\phi}{\psi^\gamma}\right]^{1/\gamma}=\left[\frac{((1-p)^{-1/\phi}-1)}{(0.5^{-1/\phi}-1)}\right]^{1/\gamma}*\eta,$$

where $P(T\le R)=p$ for p=0.95 and 0.90. Upon generating uniform random variables U~ U(0,1) and V~ U(0,1), the bivariate failure times were derived from $S(t|\alpha)=e^{-\alpha(\psi t)^\gamma}$ as

$$t_1=\left[-\frac{\log(1-u)}{\alpha\psi^\gamma}\right]^{1/\gamma}$$

and

$$t_2=\left[-\frac{\log(1-v)}{\alpha\psi^\gamma}\right]^{1/\gamma}.$$

To accommodate for bivariate interval truncation, (t_1,t_2) was retained only if it belonged to rectangle B_1XB_2, that is if both times satisfied their truncation requirements. This procedure was continued until n=1000 pairs were generated from the truncation region.

The generated data of paired failure times were fit assuming a gamma distribution for the random effects and a cumulative baseline hazard function of the following forms: (i) 1-piece or exponential model (true model when γ=1) (ii) 2-piece model, and a (iii) 4-piece model.

The cut points were determined based on the percentiles of the true marginal distribution. That is,

$$q=\left[\frac{((1-p)^{-1/\phi}-1)\phi}{\psi^\gamma}\right]^{1/\gamma}=\left[\frac{((1-p)^{-1/\phi}-1)}{(0.5^{-1/\phi}-1)}\right]^{1/\gamma}*\eta,$$

where $P(T\le q)=p$. For a 2-piece model, the cut point was computed for p=0.50, and for a 4-piece model, the cut points were computed for p=0.25,0.50 and 0.75. Parameter estimates under the bivariate truncation conditions were obtained using the proposed EM algorithm with ghost pairs discussed above. Note that the the crude rates of events during each of the piecewise constant intervals from our simulated data were taken to be the initial values of the piecewise constant rate parameters for the EM algorithm, and the tolerance level for convergence was specified at 1e-04.

doi: 10.7243/2053-7662-4-1

Application of proposed methodology to data on siblings with schizophrenia

L.S Penrose (1945) discussed a survey on the cases of familial mental illness, conducted in 1944 by the Division of Psychiatric Research for the Ontario Department of Health. An investigation was made on all families in which two or more members had been admitted to a mental hospital in Ontario. The records date back to 1926, and thus a period of 18 years was covered. Of the families that were admitted to a mental hospital, only those families in which two siblings were admitted for a diagnosis of schizophrenia were included in our study. The age at diagnosis of schizophrenia was taken to be the age at first admission to a mental hospital for schizophrenia.

The primary objective was to determine whether there was an association within pairs in diagnosis times among schizophrenics. Estimating the cumulative hazard function and the marginal survival function of the event times were also of interest. Various forms for the baseline hazard function were examined, including: (i) Weibull with shape parameter γ and scale parameter ψ^{-1} i.e., $\Lambda_0^*(t)=(\psi t)^{\gamma}$ and (ii) piecewise constant models with two, three, and four pieces i.e.,

$$\Lambda_0^*(t)=\sum_{r=1}^{R}\rho_r w_r(t),$$

where ρ_r is the rth unknown rate parameter corresponding to the rth piecewise constant interval, and $w_r(t)$ is the length of the intersection of interval $(0,t)$ and the rth piecewise constant interval. When aiming for a flexible weakly parametric model using piecewise constant rates, the selection of cut points should be considered carefully. The choice of cut points $a_1,\ldots,a_{R-1}$ are typically determined either i) a priori (for example, if there is prior evidence indicating the event rate may begin to change after a specific time) or ii) based on the percentiles from the estimated marginal survivor function of the data. It may also be useful to begin with a small number of cut points (say 2) and then examine the effect on parameter estimates as the number of cut points increases. In our data, the breakpoints for the piecewise constant models were selected based on the estimated marginal survivor function under the Weibull fit. The two-piece model had a breakpoint at t=20; the three-piece model had breakpoints at t=20 and 40; and the four-piece model had breakpoints at t=16,28, and 48. Upon constructing the complete data log-likelihood, the parameters of the model were estimated using the EM algorithm with ghost pairs under the conditional formulation, as proposed above.

Results

Results from simulations

Results from the simulation study under the true value γ=1 and γ=1.2 are provided in **Tables 1** and **2**, respectively. The mean and simulation-based standard error are given for the estimates of the model parameters. Interest lied in examining how the average bias and standard error for ϕ and ρ vary with respect to (i) the number of pieces (ii) the amount of truncation , and (iii) the true value of the variance of the random effect distribution. Recall that as the variance ϕ^{-1} approaches zero, there was no association among failure times within a pair.

From **Table 1**, when increasing the number of pieces in the model of the baseline hazard function, there was no notable effect on the bias of the estimates, however the standard error increased. The same occurred as the degree of truncation

Table 1. Summary of analysis of interval-truncated bivariate failure time data under a conditional model with marginal survivor function $S^*(t;\gamma,\psi,\phi)=\left[1+(\psi t)^{\gamma}/\phi\right]^{-\phi}$ ($\gamma=1.0,\eta=40,\psi=0.0189$ and log ψ=-3.967 if ϕ^{-1}=0.25, ψ=0.0227 and log ψ=-3.784 if ϕ^{-1}=0.75.

				5%						10%			
		1-Piece		2-Piece		4-Piece		1-Piece		2-Piece		4-Piece	
		Mean+	SE++	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
--	--	1.382	0.075	1.383	0.089	1.390	0.097	1.385	0.081	1.387	0.123	1.383	0.099
ϕ^{-1}=0.25	log ρ_1	-3.967	0.035	-3.968	0.037	-3.972	0.077	-3.966	0.040	-3.966	0.041	-3.974	0.075
(log ϕ=1.386)	log ρ_2	--	--	-3.965	0.028	-3.966	0.037	--	--	-3.961	0.074	-3.967	0.036
--	log ρ_3	--	--	--	--	-3.962	0.040	--	--	--	--	-3.966	0.043
--	log ρ_4	--	--	--	--	-3.946	0.209	--	--	--	--	-3.928	0.355
--	log ϕ	0.278	0.118	0.283	0.132	0.278	0.134	0.261	0.182	0.266	0.198	0.244	0.241
ϕ^{-1}=0.75	log ρ_1	-3.785	0.044	-3.785	0.045	-3.782	0.059	-3.785	0.048	-3.785	0.047	-3.780	0.063
(log ϕ=0.288)	log ρ_2	--	--	-3.781	0.068	-3.789	0.057	--	--	-3.785	0.076	-3.788	0.065
--	log ρ_3	--	--	--	--	-3.784	0.072	--	--	--	--	-3.786	0.089
--	log ρ_4	--	--	--	--	-3.782	0.125	--	--	--	--	-3.765	0.194

+Average estimated value over all simulations; ++Empirical standard error

doi: 10.7243/2053-7662-4-1

Table 2. Summary of analysis of interval-truncated bivariate failure time data under a conditional model with marginal survivor function $S^*(t;\gamma,\psi,\phi)=\left[1+(\psi t)^{\gamma}/\phi\right]^{-\phi}$ ($\gamma=1.2,\eta=40,\psi=0.0198$ and log ψ=-3.921 if ϕ^{-1}=0.25, ψ=0.0231 and log ψ=-3.768 if ϕ^{-1}=0.75.

				5%						10%			
		1-Piece		2-Piece		4-Piece		1-Piece		2-Piece		4-Piece	
		Mean+	SE++	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
--	$\log\phi$	2.091	0.223	1.676	0.165	1.429	0.056	1.601	0.287	1.576	0.169	1.414	0.082
ϕ^{-1}=0.25	$\log\rho_1$	-4.018	0.037	-4.043	0.037	-4.336	0.078	-4.103	0.045	-4.070	0.044	-4.347	0.073
($\log\phi$=1.386)	$\log\rho_2$	--	--	-3.810	0.061	-3.969	0.037	--	--	-3.877	0.089	-3.983	0.035
	$\log\rho_3$	--	--	--	--	-3.819	0.044	--	--	--	--	-3.856	0.035
	$\log\rho_4$	--	--	--	--	-3.915	0.128	--	--	--	--	-3.903	0.138
	$\log\phi$	0.909	0.161	0.576	0.153	0.396	0.044	0.888	0.304	0.677	0.227	0.334	0.066
ϕ^{-1}=0.75	$\log\rho_1$	-3.861	0.037	-3.888	0.042	-4.015	0.059	-3.879	0.048	-3.887	0.046	-4.050	0.048
($\log\phi$=0.288)	$\log\rho_2$	--	--	-3.663	0.063	-3.738	0.056	--	--	-3.693	0.075	-3.781	0.047
	$\log\rho_3$	--	--	--	--	-3.589	0.071	--	--	--	--	-3.669	0.047
	$\log\rho_4$	--	--	--	--	-3.497	0.131	--	--	--	--	-3.724	0.168

+Average estimated value over all simulations; ++Empirical standard error

increased. Furthermore, a larger value for the variance of the random effect distribution resulted in an increase in both the bias and the standard error of the estimate for ϕ. **Table 2** showed a strong decreasing trend in the bias of the estimate for ϕ as the number of pieces increased. Furthermore, a rise in the degree of truncation resulted in a notable increase in the standard error of the estimate for ϕ. Note that the magnitude of the bias of the estimates for all parameters in **Table 1** and **Table 2** were in fact very small, irrespective of any change in conditions, and the standardized bias also indicated that the bias was negligible throughout.

Results from application to data on siblings with schizophrenia

The data was comprised of 173 pairs of siblings, in which each member of a pair was hospitalized with schizophrenia between the years 1926 and 1944. The age at first hospitalization was considered as the age of diagnosis (the event time). This observation was left and right truncated between the age at 1926 and the age at 1944, respectively. The ages at first hospitalization for schizophrenia for the first-born and second-born siblings are represented by T_1 and T_2, respectively. The ages at 1926 for the first-born and second-born siblings are denoted by L_1 and L_2, respectively. And the ages at 1944 for the first-born and second-born siblings are represented by R_1 and R_2, respectively. Note that in this particular case, the length of the truncation interval is the same across individuals, and thus the area of the truncation square is the same across pairs. Data for the first 20 pairs of siblings can be found in **Table 3**. The data is further illustrated in **Figure 2**. The x-axis provides the age of the first-born sibling and the y-axis represents the age of the second-born sibling. The truncation intervals for individuals in a pair create a truncation square, and their event times are given by the corresponding point, which is only observed because it falls within their truncation square.

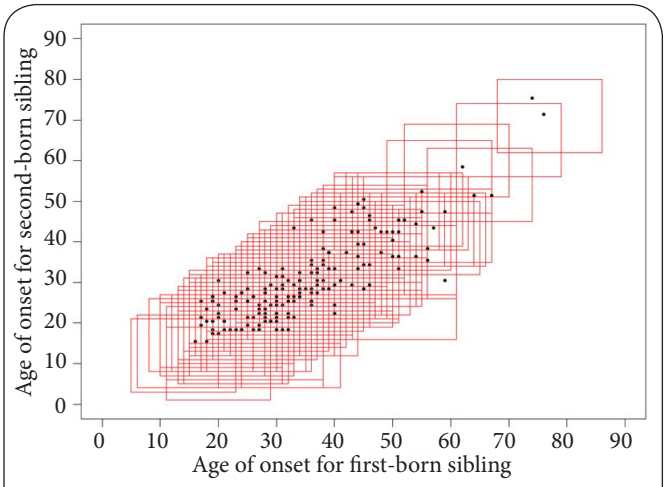

Figure 2. Truncation rectangles for sibling pairs with schizophrenia.

Table 4 provides the results from our analysis. The estimates and their standard errors are given for the parameters of the Weibull, 2-piece, 3-piece, and 4-piece models, along with the corresponding observed log-likelihood at the maximum. Once these estimates were obtained, the 25th, 50th, and 75th percentiles under each model were computed by solving

$$[1+\phi\Lambda_0^*(t_q)]^{-1/\phi}=1-q$$

doi: 10.7243/2053-7662-4-1

Table 3. Schizophrenia Data (first 20 sibling pairs).

Pair ID	T_1	L_1	R_1	T_2	L_2	R_2
1	19	18	36	20	11	29
2	32	26	44	30	24	42
3	45	45	63	50	38	56
4	27	21	39	23	17	35
5	27	18	36	18	8	26
6	44	36	54	36	33	51
7	40	25	43	22	17	35
8	34	21	39	32	19	37
9	33	23	41	21	4	22
10	30	18	36	26	12	30
11	23	18	36	26	18	36
12	27	20	38	22	4	22
13	46	34	52	45	32	50
14	28	19	37	20	18	36
15	44	43	61	39	38	56
16	46	33	51	29	26	44
17	32	30	48	26	24	42
18	30	13	31	21	6	24
19	36	36	54	35	33	51
20	39	36	54	37	34	52

for t_q, where q=0.25,0.50,0.75. Furthermore, the natural log of the odds ratio and its standard error were computed at t=21.0 under each model. The focus on the age of 21 was determined by our clinical collaborators, who define this age to be an indicator of early versus late schizophrenia onset; it is also believed that genetic factors are strongly associated with onset of schizophrenia prior to the age of 21 [14]. Note that the natural log of the odds ratio at time t was calculated as:

$$\log\left(OR\left(t\right)\right)=\log\left[\frac{P(T_2>t|T_1>t)/P(T_2\le t|T_1>t)}{P(T_2>t|T_1\le t)/P(T_2\le t|T_1\le t)}\right],$$

which was simply a function of marginal and joint survival functions. The estimate of the variance of the log($OR(t)$) estimator was computed using a sandwich estimator. The results of the odds ratio under all models led to similar conclusions. Specifically, under a Weibull baseline hazard model, the 95% confidence interval for the log($OR(t$=21)) was (0.080, 3.780), indicating a strong presence of association among diagnosis times for schizophrenia within pairs of siblings.

The results in **Table 4** were based on the joint modeling of paired failure times. For comparison purposes, however, similar measures may be obtained based on independence assumptions among event times within a pair. With a naive matched-pairs analysis under a binary response (age of onset ≤21 or age of onset >21), the estimate of the log($OR(t$=21)) is 0.788 with standard error 0.381, and the 95% confidence interval for the log($OR(t$=21)) was (0.035, 1.525). Furthermore, McNemar's test of marginal association for such matched-pair data gave a chi-square test statistic of 3.781 with a resulting p-value of 0.052. These naive results agreed with our more formal analysis, indicating a notable presence of association among diagnosis times for schizophrenia within pairs of siblings.

An illustration of the performance of the piecewise constant models can be seen from the plot of the estimated marginal survivor functions given in **Figure 3**. The 95% confidence bands are also provided under the Weibull fit at years 20, 40, and 60. The results from the piecewise models agree with that from the Weibull model. However, during the first two decades in which very few events are observed to occur, the four-piece model provides a more suitable fit than the two-piece or three-piece model.

Although the piecewise model fit could be carried out using direct maximization, the EM algorithm was used to avoid divergence problems often arising when there is a large number of pieces. In the case of a 2-piece model, the computing

Table 4. Results of the schizophrenia data analysis under a conditional formulation.

	Weibull		2-Piece (0,20],(20,∞)		3-Piece (0,20],(20,40],(40,∞)		4-Piece (0,16],(16,28],(28,48],(48,∞)	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
ϕ	3.789	1.007	0.262	0.950	0.226	0.722	1.713	0.719
γ	4.993	0.510	--	--	--	--	--	--
ψ	0.032	0.004	--	--	--	--	--	--
ρ1	--	--	0.010	0.009	0.011	0.007	7e-04	0.009
ρ2	--	--	0.030	0.027	0.035	0.020	0.026	0.033
ρ3	--	--	--	--	0.041	0.030	0.067	0.075
ρ4	--	--	--	--	--	--	0.096	0.142
l_{obs}	-971.580	--	-989.391	--	-989.181	--	-986.748	--
$t_{0.25}$	27.680	--	23.099	--	21.909	--	27.360	--
$t_{0.50}$	40.263	--	38.261	--	34.693	--	42.027	--
$t_{0.75}$	69.091	--	68.288	--	56.495	--	89.198	--
log (OR)	1.930	0.944	0.296	0.098	0.267	0.099	1.279	0.098

doi: 10.7243/2053-7662-4-1

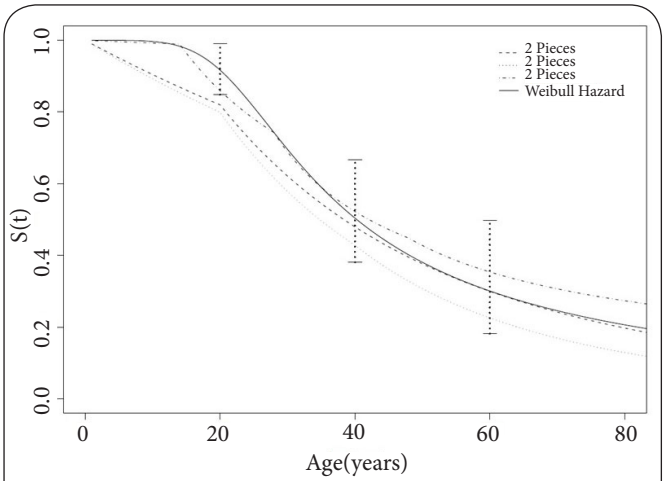

Figure 3. Estimated marginal survivor functions under various fits to the schizophrenia data based on a conditional model.

time using a standard maximization function was faster than implementing an EM approach. However, if one wanted to fit an 8-piece or 10-piece model, the EM algorithm was attractive since it did not require obtaining the derivative functions, and thus avoided possible problems with convergence.

Discussion

This paper was concerned with inference for bivariate failure time data that were interval-truncated. A conditional (random effect) formulation was discussed for modeling the marginal and joint distribution of the failure times, where particular interest lied in measuring the association within pairs. Under the conditional model, the marginal distributions were determined by both the variance parameter of the random effect distribution and the parameters of the conditional hazard function. To minimize the need for strong distributional assumptions and facilitate a likelihood analysis based on parametric models, the baseline hazard functions were assumed to be of a piecewise constant form.

The literature offers limited approaches in estimation when both failure times in a pair are observed subject to interval truncation. This paper developed an EM approach for estimation based on a conditional formulation. The idea of "ghosts" was introduced in the bivariate setting so that the complete data log-likelihood could be easily maximized. The formulation was convenient to implement since the E-step was relatively straightforward, and it also provided closed form expressions for the parameter estimates of the piecewise constant conditional hazard function in the M-step. Results from the simulation studies indicated that all model parameters could be estimated with negligible bias in finite samples. The data simulated for **Table 1** was based on an exponential baseline hazard function (as the shape parameter γ=1), which is why more pieces were not needed to improve model fit. On the other hand, as the baseline hazard function was not constant for the data simulated in **Table 2**, we see that increasing the number of pieces provided a better approximation to the true marginal survivor function. In the motivating schizophrenia data set, our proposed model and method of estimation revealed a strong presence of dependence among schizophrenia diagnosis times within pairs of siblings. In addition, since the rate of onset for schizophrenia was quite low in the early years, the four-piece model was more sensitive to this shift and was able to detect this shift, thus providing different results than the two-piece and three-piece models. Moreover, as the estimate of the shape parameter under the Weibull distribution was much greater than 1.0, it was reasonable that increasing the number of pieces provided a better fit, as seen by the plots of the estimated marginal survivor functions in **Figure 3**.

Although we considered the scenario where there was no censoring, our proposed methods could be naturally extended to handle the case where the event times were interval-censored within their respective truncation intervals. We also considered the scenario in which there were paired failure times; our conditional approach could be easily extended to handle a multivariate setting with more than 2 individuals within a group.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Authors' contributions	RS	RJC
Research concept and design	✓	✓
Collection and/or assembly of data	✓	✓
Data analysis and interpretation	✓	✓
Writing the article	✓	✓
Critical revision of the article	✓	✓
Final approval of article	✓	✓
Statistical analysis	✓	✓

Acknowledgement and funding

This work was supported by the Institute for Clinical Evaluative Sciences. Richard Cook is a Canada Research Chair in Statistical Methods for Health Research. We would also like to thank Dr. Janice Husted from the University of Waterloo, Canada, and Dr. Anne Bassett from the University of Toronto, Canada, for their valuable input and guidance regarding the data.

Publication history

EICs: Jimmy Efird, East Carolina University, USA.
Max K. Bulsara, University of Notre Dame, Australia.
Received: 17-Nov-2015 Final Revised: 31-Dec-2015
Accepted: 13-Jan-2016 Published: 29-Jan-2016

References

1. Turnbull B.W. **The empirical distribution function with arbitrarily grouped, censoredand truncated data**. *Journal of the Royal Statistical Society B*. 1976; **38**:290-295. | Article
2. Dempster A.P, Laird N and Rubin D.B. **Maximum likelihood from incomplete data via the EM algorithm**. *Journal of the Royal Statistical Society B*. 1977; **39**:1-38. | Pdf

doi: 10.7243/2053-7662-4-1

3. Frydman H. **A note on nonparametric estimation of the distribution function frominterval-censored and truncated observations**. *Journal of the Royal Statistical Society B*. 1994; **56**:71-74. | Article
4. Alioum A and Commenges D. **A proportional hazards model for arbitrarily censored and truncated data**. *Biometrics*. 1996; **52**:512-24. | Article | PubMed
5. Van der Laan M.J. **Nonparametric estimation of the bivariate survival function with truncated data**. *Journal of Multivariate Analysis*. 1996; **58**:107-131. | Article
6. Gurler U. **Nonparametric bivariate estimation with randomly truncated observations**. *Handbook of Statistics*. 2004; **23**:195-207.
7. Andersen P.K, Borgan O, Gill R.D and Keiding N. **Statistical Models Based on Counting Processes**. New York: Springer-Verlag. 1993.
8. Clayton D.G. **A model for association in bivariate lifetables and its application inepidemiological studies of familial tendency in chronic disease incidence**. *Biometrika*. 1978; **65**:141-151. | Article
9. Hougaard P. **Survival models for heterogeneous populations derived from stable distributions**. *Biometrika*. 1986; **73**:387-396. | Article
10. Hougaard P. **A class of multivariate failure time distributions**. *Biometrika*. 1986; **73**:671-678. | Article
11. Penrose LS. **Survey of cases of familial mental illness**. *Eur Arch Psychiatry Clin Neurosci*. 1991; **240**:315-24. | Article | PubMed
12. Crow T.J. **A note on "Survey of Cases of Familial Mental Illness" by L.S. Penrose**. *European Archives of Psychiatry and Clinical Neuroscience*. 1991; 507-517. | Article
13. Duchateau L and Janssen P. **The Frailty Model**. New York: Springer-Verlag. 2008.
14. White F, Livesey D and Hayes B. **Developmental Psychology: From Infancy to Development, 3rd edition**. Pearson Australia. 2013.

Citation:
Sutradhar R and Cook RJ. **A conditional frailty model for bivariate interval-truncated failure time data: an application to a study on siblings diagnosed with schizophrenia**. *J Med Stat Inform*. 2016; **4**:1.
http://dx.doi.org/10.7243/2053-7662-4-1