An Overview of Estimating Causal Effects from Interval-Censored Data: G-Estimation Approach

T. I. Valappil[a], K. P. Singh[b], and A. A. Bartolucci[c]

a. *Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, AL 35294, USA (tvalappil@ms.soph.uab.edu)*

b. *Department of Epidemiology and Biostatistics, School of Public Health, UNT Health Science Center, Fort Worth, TX 76107, USA (ksingh@hsc.unt.edu)*

c. *Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, AL 35294, USA (albartol@uab.edu)*

Abstract: In a randomized or observational study, it is very important to control time-dependent covariates which can be confounders as well as intermediate variables, using appropriate statistical methods. The standard approaches may be biased when a covariate is influenced by treatment history and is also a determinant of subsequent outcome and treatment history. In this paper, we consider the problem of interval censoring in an observational study setup and discuss methods of controlling such covariates, in order to estimate the causal effect of a time-dependent treatment or an exposure on survival. In the presence of such covariates, the method of G-estimation allows for the appropriate adjustment and this method uses a new class of Structural Nested Failure Time (SNFT) models. This is based on the notion of counterfactual failure times estimated from interval censored data which can be reformulated as an incomplete or missing failure time data. The maximum likelihood estimates of the survival times are equivalent to solving Turnbull's self-consistency equations.

Keywords: Confounding Variable; Intermediate Variable; Sexually Transmitted Diseases, G-Estimation

1. INTRODUCTION

The standard strategy for estimating the effect of a time-varying treatment on an outcome is to model the probability of the outcome at time t as a function of past treatment history. This approach may be biased, when the time-dependent covariates are often both confounders and intermediate variables. This is true if one adjusts for the past history of time-dependent confounding covariates. Robin [1986] first proposed a novel approach for using observational data to estimate the causal effect of a time-varying treatment. This method, called G-estimation, uses a new class of structural nested failure time (SNFTM) models based on the notion of counterfactual failure times. For more details the reader is referred to Robin [1986, 1989 and 1992].

Interval censored data can be reformulated as an incomplete or missing failure time data and there are several approaches available in literature to estimate the missing data and the maximum likelihood estimates of the survival times are equivalent to solving Turnbull's self-consistency equations. Application of this methodology will be useful in analyzing follow-up data where the event of interest had occurred in an interval. In the context of causal inference, estimating the effect of treatments in the presence of confounding and intermediate variables had been a well discussed issue and the usual analysis methods to estimate these effects would produce biased results.

Robin [1986] promoted a model for causal inference based on potential outcomes if individuals receive different treatments under study. Commonly, the assumption is made that the outcome in one individual is independent of the treatment assignment and outcome in other individuals. The implication in much of the literature is that only properly randomized experiments can lead to useful estimates of causal effects. If taken as applying to all fields of research, this position is untenable. Even if the position that causal effects of

treatments can only be well established from randomized experiments is applied to the social sciences, in which there are only currently a few well-established causal relationships, its implication to ignore existing observational data may be counter- productive. Often the only immediately available data are observational (nonrandomized) and either the cost of performing the equivalent randomized experiment to test all treatments is prohibitive or there are (other) ethical reasons for which the treatment cannot be randomly assigned.

G-estimation is designed to estimate effects of generalized treatments (i.e. time-dependent exposures that may influence and be influenced by other time-dependent variables). Its practicality and robustness arise from the fact that it makes no assumption about the causal relations among the covariates. The only causal dependence that it models is of the exposure effect on the outcome. Along with that causal model, it employs another model for the regression of the study of exposure at each point in time on the exposure, covariate, and disease history of each subject up to that point in time.

2. STATISTICAL METHODS FOR ANALYSIS OF INTERVAL-CENSORED SURVIVAL DATA

Survival data render standard methods inappropriate because survival times are frequently censored. The survival time of an individual is censored when the end point of interest has not been observed for that individual. This may be because the data from a study are to be analyzed at a point in time when some individuals are still alive. Alternatively, the survival status of an individual may have been lost to follow-up. In some situations, however, the times of the events of interest may only be known to have occurred within an interval of time, $[L_i, R_i]$, where $L_i < T < R_i$. This can occur in a clinical trial (for example, when patients are assessed only at prescheduled visits). If the event has not occurred at one visit (time L) but has occurred by the following visit (time R), T is known only to be within the interval $[L_i, R_i]$. These are called interval censored data. Note that exactly observed, right and left censored data are special cases of interval censored data, with L = R for exactly observed data and R = ∞ for right censored and L = 0 for left censored observations.

Peto [1973] proposed a method to estimate survival curves from interval censored data which is analogous to the estimate desirable from right censored data by the life table techniques. Turnbull [1976] generalized the Kaplan-Meier estimator for analysis of interval-censored survival data. Using the idea of self-consistency, Turnbull constructed a simple algorithm. The algorithm converges monotonically to yield a maximum likelihood estimate (MLE) of a distribution function. Dempster *et al.* [1977] proposed a very elegant and comprehensive theory of maximum likelihood with incomplete data and developed the expectation-maximization (EM) algorithm and its properties. The method proposed by Turnbull [1976] can be viewed as an example of an EM algorithm. If we consider heavy interval censoring of the data, it can be viewed as missing data and the method of EM algorithm can be employed to estimate the survival time. The method of EM maps the maximum likelihood equation of the parameter estimates from the observed data likelihood to the maximum likelihood estimation based on a complete-data log likelihood function. An EM algorithm can be both conceptually and computationally simple, especially when the log-likelihood of the data under no censoring has a simpler functional form than the log-likelihood of the actual observed data.

Finkelstein and Wolfe [1986] proposed a semi-parametric model for interval censored data where the distribution of survival time, T, is nonparametric but the density of the vector of covariates, X, given T, follows a specified parametric model. Finkelstein [1986] generalized Cox's regression model for analysis of interval censored survival data. Also, methods for estimating a distribution function from interval-censored data have also been studied by other authors including Turnbull [1976] and Groeneboom [1991]. Examples of methods that relate to interval-censored data in models include Shiboski and Nocholas [1992], Jane and Louise [1998], Rebecca *et al.* [1999] and Finklestein [1986]. Flygare *et al.* [1985] and Odell *et al.* [1992] employed parametric models with interval censoring. Flygare *et al.* [1985] presented MLE techniques for estimating the two-parameter Weibull distribution from interval-censored data. The parameter MLEs were compared with the estimates obtained from midpoints of intervals.

Odell *et al.* [1992] studied the use of a Weibull based accelerated failure time regression model for left and interval censored data. Estimates from two methods, maximum likelihood method and midpoint interval method, were discussed. Their simulation studies indicate that for relatively large samples there are many instances when the MLEs are superior to the estimates obtained from the midpoints of intervals.

Self and Grossman [1986] proposed linear rank statistics

for testing the differences between groups when the data are interval censored. The test statistics are closely related to those proposed by Prentice [1978] for right censored data. Buckner and Messerer [1988] contrasted the Turnbull estimator with the conventionally used Kaplan-Meier estimator. In addition, these authors also used a parametric model for estimation or simulation of the delay times of complete remission diagnosis and relapse diagnosis. They discussed two possible consequences of the conventional approach: biased estimation and underestimation of the true error variance, which may lead to false positive results. Becker and Melbye [1991] developed a method for computing the non-parametric maximum likelihood estimate (NPMLE) of the survival curve from interval censored data by fitting a log-linear binomial model. The method gives the same results as the methods devised by Peto [1973] and Turnbull [1976] when the number of points where the survival curve estimated is not too big compared to the number of observations.

Sinha [1993] presented a nonparametric Bayes method for analyzing interval censored survival data. He used Monte Carlo algorithms, including data augmentation [Tanner and Wong, 1987] and Gibbs sampling [Geman and Geman 1984], to find posterior estimates of several quantities of interest. One of the advantages of the fully parametric approach is that the number of intervals can be regarded to be independent of the time points of observations so that the practical problems of convergence in fitting an algorithm become manageable. Becker and Melbye [1991] proposed a method in the setting of a fully parametric model for the intensity and discussed practical aspects of estimation of the survival curve and confidence intervals. He also proposed an excess risk model for the intensity and implemented it in the GLIM software. These authors proposed a class of score statistics that may be used in estimation and confidence procedures. Singh *et al.* [1988] proposed a parametric method for interval censored data using a generalized log-logistic based failure time model.

Betensky *et al.* [1999] proposed a smooth hazard estimator for interval censored survival data using the method of local likelihood. The model is fitted using a local EM algorithm. The estimator is more descriptive than traditional empirical estimates in regions of concentrated information and takes on a parametric flavor in regions of sparse information. They derived two different standard error estimates for the smooth curve, one based on asymptotic theory and the other using the bootstrap method.

There are two broad reasons for modeling survival data. One objective of the modeling process is to determine which combination of potential explanatory variables affects the form of the hazard function. In particular, the effect that the treatment has on the hazard of death can be studied, as can the extent to which other explanatory variables affects the hazard function. Another reason for modeling the hazard function is to obtain an estimate of the hazard function itself for an individual. The resulting estimate could be particularly useful in devising a treatment regime or counseling patients about their prognosis.

In the parametric case, the failure time distribution is assumed known except for a few scalar parameters. The proportional hazard model, however, is nonparametric in the sense that it involves an unspecified function in the form of an arbitrary baseline hazard function. In consequence, this model is more flexible, but different approaches are required for estimation and testing. Suppose that data are available for n individuals, amongst whom there are r distinct death times and $(n-r)$ right-censored survival times. We will assume that only one individual dies at each death time, so that there are no ties in the data. The r ordered death times will be denoted by $t_{(1)} < t_{(2)} < \ldots\ldots\ldots\ldots < t_{(r)}$, so that $t_{(j)}$ is the j^{th} ordered death time. The set of individuals who are at risk at time $t_{(j)}$ will be denoted by $R(t_{(j)})$, so that $R(t_{(j)})$ is the set of individuals who are alive and uncensored at a time just prior to $t_{(j)}$. The quantity $R(t_{(j)})$ is called the risk set. Cox [1972] showed that the relevant likelihood function for the proportional hazard model is given by

$$L(\beta) = \prod_{j=1}^{r} \frac{e^{(\beta' X_{(j)})}}{\sum_{l \in R(t_{(j)})} e^{(\beta' X_{(l)})}} \qquad (1)$$

in which $X_{(j)}$ is a vector of covariates for the individual who dies at the jth ordered death time $t_{(j)}$ Peto [1973] proposed a nonparametric method for estimating the survival distribution based on interval censored data. Suppose T_i (i=1,2,.....n), the survival times for n patients, are independent random variables with S(t) as the survival function, the probability that the event of interest occurred beyond the time point t. This survival function can be estimated for interval censored data assuming that the follow-up visits are fixed. Let us assume that the follow-up visits occur at a finite number of M-1 known times $T_1, T_2, \ldots\ldots, T_{M-1}$, where (T_{j-1}, T_j), j=1,2,....M, represent the j^{th} interval. Define α_{ij} to be

the indicator variable such that α_{ij} =1 if the i^{th} subject has failed in the j^{th} interval, α_{ij} =0 otherwise. Let $X_j = \sum_{j=1}^{N} \alpha_{ij}$. This sum represents the total number of failures in the j^{th} interval. The survival time S(t) is discrete with the probability mass distributed only at $T_1, \ldots, T_{M-1}$. Therefore, $X = (X_1, \ldots\ldots, X_M)^T$ has a multinomial distribution with M parameters; $P = (P_1, \ldots, P_M)^T$, where P_j is the probability that a failure occurs in the j^{th} interval. We can write the log likelihood function as follows:

$$L = \sum_{j=1}^{K} X_j \log P_j \qquad (2)$$

where $\sum_{j=1}^{K} P_j = 1$ and $\sum_{j=1}^{K} X_j = N$

The statistical analysis becomes straightforward if we could determine the precise interval in which the failure had occurred. This is not usually the case in observational studies. For follow-up clinical trials conducted at the STD clinics or health departments, individuals may skip the follow-up visits frequently due to various reasons and cannot precisely determine the interval in which the infection had occurred, but can only estimate that the failure had occurred within one of several successive intervals. If L_i denotes the left and R_i denotes the right element of an interval for the i^{th} individual and $L_i < R_i$, i=1,2,...,N. Define the log likelihood function for this incomplete data as given by Valappil and Singh [1999].

Now the problem of maximizing the log-likelihood function is equivalent to solving self-consistency equations [Efron, 1967]. Applying the EM algorithm [Dempster *et al.* 1977] or solving the above self-consistency equations would yield the same result. Using the self-consistency algorithm, we can search for the maximum likelihood estimates.

3. THE METHOD OF G-ESTIMATION

Nearly 750,000 cases of gonorrhea and other STDs are reported annually in the United States [CDC, DSTDP/HIVP, 1995]. Another 700,000 unreported cases, mostly among teenagers and young adults, are believed to occur each year. Analyzing these type of data in estimating the effect of prevention treatments had been a challenge to statisticians due to confounding and intermediate factors which would modify the actual but unknown treatment effect [Figure 1].

To consider analyses of the interval censored data, using the method of G-estimation [Robins 1986, 1989 and 1992], consider a study of the effect of a treatment on the survival of patients. Let T_i be the survival time of the i^{th} subject, i=1,2...n. Let $E_{i,k}(t)$ record subject i's treatment at time t and visit k and $L_{i,k}(t)$ record various baseline, time- dependent and time-independent covariates at time t and visit k. Also let $\bar{E}(t^-)$ and $\bar{L}(t^-)$ be the recorded treatment and covariate history up to but not including time t. Let U represent the counterfactual "baseline" failure time random variable [Cox and Oaks, 1984; Robins 1986 and 1987, Rubin 1974], representing a subject's failure time had, possibly contrary to the fact, that the subject was not being treated.

In a randomized study, the treatment allocation is controlled by the randomization scheme. In an observational study, Robins [1986] considered two types of failure times to measure the causal effect; based on the assumption of no unmeasured confounders at each time t, the onset of treatment at t should be conditionally independent of the failure times, given the covariate history. Based on the assumption of no unmeasured confounders ,

$$U \perp\!\!\!\perp E(t) \mid \bar{E}(t^-), \bar{L}(t^-), T \le t, \qquad (3)$$

where $E(t)$ is the treatment rate at t, $\bar{L}(t^-)$ is the recorded covariate history up to but not including t, $\bar{E}(t^-)$ is the treatment history up to but not including time t and the symbol $\perp\!\!\!\perp$ is used here to show the independence.

If (3) is true, the change in treatment rate at t is independent of the baseline counter-factual failure time U, conditional on treatment history and history of all recorded covariates prior to t. The above assumption is the fundamental condition that allows one to draw causal inferences from the observational data [Robins, 1992]. It is also very important in an observational study to collect data on a sufficient number of covariates to make the assumption (3) is approximately true since the assumption of no unmeasured confounders is not always guaranteed to hold in an observational study. Under the assumption that (3) is true, let us consider that the treatment E(t) received at time t is dichotomous. That is

E(t) = 1, treated at time t; 0, otherwise. Therefore, according to the fundamental assumption of no unmeasured confounders, we can write (3) as follows:

$$\Delta(t \mid \bar{E}(t^-), \bar{L}(t^-), U) = \Delta(t \mid \bar{E}(t^-), \bar{L}(t^-)) \qquad (4)$$

The assumption (4) is the fundamental condition which if it is true, will allow us to draw causal inferences from observational data. In all observational studies this condition cannot be guaranteed to hold and is not empirically testable; under those circumstances, it is not appropriate to draw causal inferences. However, in a sequentially randomized study, at each time t, the treatment would be chosen at random; it is valid in this case to draw causal inferences without violating any assumptions.

To estimate the causal effect of a treatment, it is imperative that we test the hypothesis of no treatment effect. Therefore, we define the null hypothesis of no treatment effect on survival time as,

$$U=T \qquad (5)$$

If this causal null hypothesis is true, it implies that the subject's observed and baseline failure times are equal irrespective of the treatment effect. In conjunction with our earlier assumption of "no unmeasured confounders" (3), the causal null hypothesis implies that, conditional on past treatment history and other covariate history, the hazards of treatment at time t do not depend on observable failure time T. Therefore, we can define the extended model that adds a term

$$\Delta(t \mid \bar{E}(t^-), \bar{L}(t^-), T) = \Delta(t \mid \bar{E}(t^-), \bar{L}(t^-)) \qquad (6)$$

If the above is an instantaneous rate process, we can test it by specifying a time-dependent Cox proportional hazards model:

$$\Delta(t \mid \bar{E}(t^-), \bar{L}(t^-), T) = \lambda_0(t) e^{\beta^* W_{i,k}(t) + \Sigma \alpha_k L_{i,k}(t)} \qquad (7)$$

where $\lambda_0(t)$ is an unspecified baseline hazard function. $W_{i,k}(t)$ given by $\int_0^t E_{i,k}(u)\,du$ is the cumulative exposure to the treatment prior to t, L_i's are real valued functions of the baseline covariates and β and α are unknown parameters. If odds model is correctly specified, an asymptotic α - level Cox partial likelihood score, Wald or likelihood ratio test of the hypothesis $\theta = 0$ in the θT to $\beta^* W_{i,k}(t)$ in (7) is referred to as a G-test by Robins [1992] The G-test is a generalization to time-dependent treatments and confounders of the effect of a single time-independent treatment.

In the case of interval censored data, to study the random variable S, which is the true time to response, and its relationship to covariates X, the likelihood function is given by Valappil and Singh [1999] and is defined as,

$$\mathcal{G} = \prod_{i=1} Pr(X_i = x_i, S_i \in (L_i, R_i]) \qquad (8)$$

The survival times are ordered in ascending order of magnitude and we define an indicator function α_{ij} such that, $\alpha_{ij} = 1$ if $S_j \in A_i$ or 0 otherwise. As discussed in Turnbull [1976], the MLE will be unique only up to an equivalence class and the likelihood function can be written as follows:

$$\mathcal{G} = \prod_{i=1} \sum_{j=1}^{m} \alpha_{ij} Pr(X_i = x_i, S_i = si) \qquad (9)$$

As discussed earlier, the failure time of the i^{th} subject is denoted by T_i if we assume a continuous distribution F_θ where θ is our parameter of interest. Suppose we only know that T_i has occurred in an interval (L_i, R_i) but have no information on the exact time of occurrence. If we assume that each subject is visited or screened randomly as a stochastic process h(t), then the pairs (L_i, R_i), where T_i is known to occur, are each realizations of the process h(t). The log-likelihood is a function of θ where the visiting process h(t) is independent of T_i for the N subjects can be written as:

$$\mathcal{G} = \sum_{i=1}^{N} \log[F(R_i) - F(L_i)] + h_p(t) \qquad (10)$$

where $h_p(t)$ involves the parameters of describing the process h(t). The visiting process does not have any effect on the duration of T_i. Since T_i is interval censored, the complete information about F can only be inferred from the interval. The only way we can reduce the loss of information is by controlling the width of the interval. In health related data these situations are very difficult to achieve. This can be illustrated by an example. Let us consider the case when h(t) is a homogeneous Poisson process with intensity function $\lambda(t) = \lambda$.

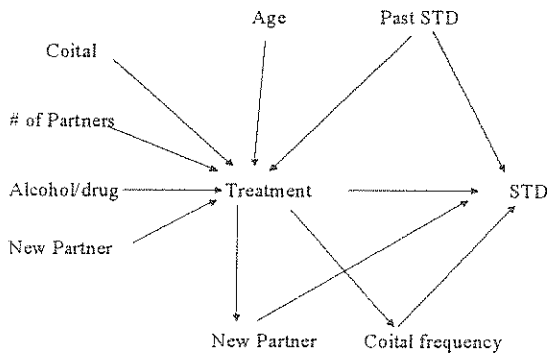

Figure 1: Confounding and Intermediate Variables: The effect of treatment on STD.

Let us assume that the time intervals between successive visits are independent and identically distributed as an exponential distribution; with that assumption, the probability density function of (L_i, R_i) is given by

$$Pr(L_i = l, R_i = r) = [\lambda e^{-\lambda r} I_{(l=0,r>0)} + \lambda^2 e^{(-\lambda(r-l))} I_{(0<l<r)}](F(r) - F(l)) \quad (11)$$

Now let us consider the case of fixed follow-up visits because fixed follow-up visit data are more popular in public health research. In a cohort study of a specific outcome, it will be difficult to keep track of the outcome on a continuous basis. In most of the studies the subjects were followed up weekly, biweekly, or monthly and it makes easier if the intervals between visits are constant over time and the variability associated with the follow-up time can be controlled. Let the interval between each visit be incremented by δ. Therefore, under a fixed follow-up scheme R_i becomes $L_i + \delta$. If we assume T_i follows an exponential distribution with mean $\frac{1}{\theta}$ then the log likelihood function can now be written as:

$$\vartheta(\theta) = \sum_{i=1}^{N} \log[e^{(-\theta L_i)} - e^{(-\theta(L_i+\delta))}] \quad (12)$$

The MLE of θ is given by Valappil and Singh [1999]. Usually, follow-up studies cover a fixed period of time and is decided at the beginning of the study. The survival time of an individual is said to be censored if the end point of interest has not been observed during the study time due to lost to follow-up or other reasons. Consider the situation where there are N subjects and k follow-up visits. As we assumed earlier, each visit is incremented by. Thus, if the event does not occur for the ith subject t after k visits, the i^{th} observation will be right censored at $k\delta$. With the introduction of right censoring, the MLE of θ is given by Valappil and Singh [1999].

There are several other approaches including multiple imputation [Rubin,1987] or midpoint imputation method to estimate the survival time using interval censored data. In the case of mid-point imputation, if we could assume that each follow-up interval is incremented by δ, (L_i, R_i) will be replaced by $L_i + \delta/2$ and we can easily get the MLE associated with the log-likelihood function and it is given by Valappil and Singh [1999]. There are several issues regarding the use of mid-point and other imputation methods and its impact on causal inference.

4. DISCUSSION

In standard time-to-event or survival analysis, occurrence times of the event of interest are observed exactly or are right- censored, meaning that it is only known that the event occurred after the last observation time. There are numerous methods available for estimating and testing the censored survival data. In some situations, however, the times of the events of interest may only be known to have occurred within an interval of time. Interval censoring arises naturally whenever individual subjects or experimental units are observed only occasionally during follow up times decided in to the study design stage. Other examples of interval censored data would be the case of human immunodeficiency virus (HIV). Important events such as infection, seroconversion (first appearance of detectable antibodies) are ascertainable only by laboratory analysis. Therefore the specific times of infection are not known. Several approaches are currently available for fitting the proportional hazard model to interval-censored or grouped data.

The interval-censored data can be reformulated as missing failure time data. When faced with missing values, imputation is a generic term for filling in missing data with plausible values. In a multivariate dataset, each missing value may be replaced by the observed mean for that variable or may be substituted by some sort of predicted value from a regression model. Expectation-Maximization (EM) algorithm is a general technique for finding maximum likelihood estimates for parametric

models when the data are not fully observed. There are many statistical problems which may not appear to involve missing data, but which can be reformulated as missing data problems. Interval-censored data fit into this category.

Multiple imputation [Rubin, 1987] is another technique in which each missing value is replaced by m>1 simulated values. The *m* sets of imputations reflect uncertainty about the true value of the missing data. The task of generating multiple imputations has been problematic until recent years. No straightforward, general purpose algorithm has been available for generating proper multiple imputations in a multi variate setting, but currently using the fastest computers, this is possible using the technique of iterative simulation. The method of self-consistency algorithm proposed by Turnbull is another approach to estimate the survival time using interval censored failure time data. Turnbull [1976] has proved that maximizing the log likelihood function is equivalent to using the self-consistency equations.

In conclusion, the conventional analytic methods may be inappropriate for the analysis of complex causal relations in follow-up studies especially when the data is interval-censored. G-estimation provides an important enhancement to the validity of such analyses. This research developed an approach for the analysis of interval-censored survival times, which employs the self-consistency algorithm in using the G-estimation procedure.

5. ACKNOWLEDGEMENT

We would like to extend our sincere thanks to Drs. Maurizio Macaluso, Marshall Joffe and James Robins for their invaluable advice and help.

6. REFERENCES

Betensky, R. A., J. C. Lindsey, L. M. Ryan and M. P. Wand, Local EM Estimation of the hazard function for interval-censored data, *Biometrics*, 55, 238-245, 1999.

Buckner, G, and D. Messerer, Remission duration:an example of interval-censored observations, *Statistics in Medicine*, 7, 1139-1145, 1988.

Becker, N.G. and M. Melbye, Use of log-linear model to compute the empirical survival curve from interval-censored data, with application to data on tests for HIV positivity, *Australian Journal of Statistics*, 33, 125-133, 1991.

CDC, DSTD/HIVP, Rosenberg, P.S., Scope of the AIDS epidemic in the United States, *Science* 270, 1372-1377, 1995.

Cox, D.R., Regression models and life-tables, *Journal of American Statistical Association* B, 34, 187-220, 1972.

Cox, D.R. and D. Oaks, *Analysis of Survival Data*. London: Chapman and Hall, 1984.

Dempster, A.P., N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data, *Journal of the Royal Statistical Society* Series B, 37, 1-38, 1977.

Efron, B., The two sample problem with censored data. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 4, 831-853, 1967.

Finkelstein, D. M., A proportional hazards model for interval-censored failure time data, *Biometrics*, 42, 845-854, 1986.

Finkelstein, D. M. and R. A. Wolfe, Isotonic regression analysis of interval-censored failure time data using EM algorithm, *Communication on Statistics*, A, 15, 8, 2493-2505, 1986.

Flygare, M.E., J. A. Austin, and R. M. Buckwalter, Maximum likelihood estimation for the 2-parameter Weibull distribution based on interval-data, *IEEE Transactions on Reliability*, R-34, 57-59, 1985.

Geman, S and D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-74, 1984.

Groeneboom, P., Nonparametric maximum likelihood estimation for interval censoring and deconvolution, *Technical Report*, 378, Department of Statistics, Stanford University, 1991.

Jane C. L. and M. R. Louise, Tutorial in Biostatistics methods for interval-censored data, *Statistics in Medicine*, 17, 219-238, 1998.

Odell, P.M., K. M. Anderson and R. B. D'Agostino, Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time data, *Biometrics*, 48, 951-959, 1992.

Peto, R., Experimental survival curves for interval censored data, *Applied Statistics*, 22, 86-91, 1973.

Prentice, R.L., Linear rank tests with right-censored data, Biometrica, 65, 167-179, 1978.

Rebecca A.B., J. C. Lindsey, L. M. Ryan and M. P. Wand, Local EM estimation of the hazard function for interval-censored data, *Biometrics*, 55, 238-245, 1999.

Robins, J. M., A new approach to causal Inference in mortality studies with a sustained exposure period, Applications to control of the healthy worker effect,

Mathematical Modeling, 7,1393-1512, 1986.

Robins, J.M., Addendum to a new approach to causal inference in mortality studies with a sustained exposure period, Applications to control of the healthy worker effect, *Computers for Mathematical Application*, 14, 923-945, 1987.

Robins, J.M., The control of confounding by intermediate variables, *Statistics in Medicine*, 8, 679-701, 1989.

Robins,J.M., Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, 79, 321-334, 1992.

Rubin, D.B., Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology*, 66(5), 688-701, 1974.

Rubin, D.B., Multiple Imputation for nonresponse in Surveys, New York, John Wiley, 1987.

Self, S.G. and E. A. Grossman, Linear rank tests for interval-censored data with application to PCB levels in adipose tissue of transformer repair workers, *Biometrics*, 42, 521-530, 1986.

Shiboski, S.C. and P. J. Nicholas, Statistical analysis of the time dependence of HIV infectivity based on partner study data. *Journal of American Statistical Association*, 87, 360-372, 1992.

Sinha, D., Nonparametric Bayesian analysis of interval-censored survival data. Paper presented at the ENAR Biometric Society Spring Meetings, Philadelphia, PA, March 21-24, 1993.

Singh, K.P.,C. Lee and E.O. George, On a generalized log-logistic model for analysis of censored survival data, *Biometrical Journal*, 37, 843-850, 1988.

Tanner, M.A. and W. H. Wong, The calculation of posterior distribution by data augmentation, *J. American Statistical Association*, 82, 528-540, 1987.

Turnbull, B.W., The empirical distribution function with arbitrarily grouped, censored and truncated data, *Journal of Royal Statistical Society Series* B, 38, 290-295, 1976.

Valappil, T. and K. P. Singh, Assessing causal effects in epidemiological studies, Manuscript, Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, USA, 1999.