Published in final edited form as:

Stat Med. 2006 December 15; 25(23): 4053–4064. doi:10.1002/sim.2505.

Estimation of survival functions in interval and right censored data using STD behavioural diaries

Jaroslav Harezlak¹ and Wanzhu Tu^{2,3,*}

¹Department of Biostatistics, Harvard University School of Public Health, U.S.A.

²Division of Biostatistics, Indiana University School of Medicine, Indianapolis; IN 46202, U.S.A.

³Regenstrief Institute, U.S.A.

Summary

In investigations of sexually transmitted diseases (STDs), true infection times are often either interval censored or right censored. For such data, reliable estimation of the survival function is difficult to obtain. In this research, we propose a resampling based method for the estimation of the survival function using auxiliary behavioural information provided by daily diaries. By imputing the unknown infection time from a list of sexual encounter times recorded in the diaries, the proposed procedure can be easily implemented using existing estimation procedures for right censored data. Our simulations show that the proposed procedure consistently outperforms the existing *ad hoc* estimation methods and produces smaller mean integrated squared errors (MISE). We illustrate this method by analysing infection data obtained from an ongoing STD investigation of adolescent females.

Keywords

diary data; imputation; interval censoring; resampling; survival function estimation

1. Introduction

Interval censored data are common in studies involving infectious diseases. For example, in longitudinal investigations of sexually transmitted diseases (STDs), the infection status of a study subject is only available at pre-specified follow-up visits. The acquisition of the infectious pathogens, therefore, is only known to have occurred within a time interval while the exact infection time remains unobservable. Similar data situations occur frequently in studies of HIV/AIDS infections.

A number of methods have been proposed in statistical literature for the estimation of a survival function in the presence of interval censoring. Peto [1] proposed a non-parametric approach and Turnbull [2] derived a self-consistent algorithm for its estimation. Another method was considered by Finkelstein [3], where she proposed the use of a non-parametric maximum likelihood method to fit the Cox proportional hazard model. More recently, Kooperberg and Stone [4] used a spline technique to obtain a smoothed estimate of the survival function. Pan proposed an algorithm that iteratively imputes the unobserved event times from a pre-specified model [5,6], while several other authors considered Bayesian methods [7,8]. Lindsey and Ryan's [9] tutorial provides an excellent review of the literature. Taylor *et al.* [10] used multiple

*Correspondence to: Wanzhu Tu, Division of Biostatistics, Indiana University School of Medicine, RG4101, 1050 Wishard Boulevard, Indianapolis, IN 46202, U.S.A..

[†]E-mail: wtul@iupui.edu

imputation methods in the analysis of right-censored survival date. Their work has established a theoretical foundation for the use of the imputation methods in the context of survival analysis by drawing connections with the redistribution to the right algorithm and the standard Kaplan–Meier estimation.

From an epidemiological perspective, the primary pathway of the STD infection is sexual contact between partners. Therefore, the times of sexual intercourse represent the potential points of disease transmission. Recognizing this mechanism, STD researchers introduced diaries to record the subject's daily sexual activities, hoping that the diary data would supplement the diagnostic test results obtained at pre-determined clinical visits [11]. Intuitively, STD behavioural diaries contain important information about the potential times of infection, and therefore should not be disregarded in the estimation of the survival function. However, none of the aforementioned methods readily accommodates the auxiliary information provided by the diaries. For the self-consistent estimator, another practical complication is the requirement of randomness of observation intervals, which is rarely true for designed cohort studies that have pre-determined follow-up visits.

In the absence of appropriate estimation methods and computational algorithms, many analysts impute the unobserved infection time with an arbitrarily selected time point within the diagnostic interval. With the imputed data, survival function estimates are obtained from procedures for right censored data. Frequently used imputation points include the times of the first or the last recorded coital episodes within the diagnostic interval, or the mid-point between those two episodes. When diary data are not available, the beginning, the end, or the mid-point of the diagnostic interval are used for the imputation of the infection time. However, as observed by many authors, these *ad hoc* imputation methods often lead to biased estimates and questionable results [12,13].

In this paper, we propose a multiple imputation procedure that resamples (with replacement) the possible true infection times from the times of coital episodes recorded by the daily diaries. Our procedure is related to the multiple imputation approach proposed by Pan [5], but with three important distinctions: (1) it relies on the auxiliary information provided by the diary data instead of assuming a survival distribution; (2) it is not an iterative procedure in the sense that current estimates are improved upon previous estimates; and (3) it uses a different predictive distribution for the imputation of the unobserved infection times. The primary motivation of the proposed method is to better utilize the available information on potential infection times. A simulation study is conducted to assess the operating characteristics of the proposed method. We compare our method with several single imputation methods and the self-consistent estimator. Finally, as an illustration of our procedure, we consider a real-world example from an STD investigation that has motivated our research.

2. Example

Bacterial STD infections by organisms *Neisseria gonorrhoeae* (NG), *Chlamydia trachomatis* (CT), or *Trichomonas vaginalis* (TV) are easily treatable with antibiotics, but the risk of reinfection is high. In a study on the behavioural epidemiology of recurrent sexually transmitted infections (STIs) in adolescents, female subjects between the ages of 14 and 17 who had been infected with one of the aforementioned organisms were approached at a county STD clinic. Eligible patients were enrolled after written informed consent. The informed consent statement was in compliance with U.S. Department of Human Services guidelines and was approved by a local institutional review board [14].

Upon enrolment, all patients received appropriate treatment of their infections and were scheduled for return visits at 1, 3, 5, and 7 months. At enrolment, subjects were given pocket size diaries and received instructions to record the occurrence of sexual intercourse, condom

use, condom failures, and an array of behavioural factors. Literature has shown that data collected from behavioural diaries have accuracy comparable to data collected using other methods, but diary data offer greater details of behaviours in their original time sequence [15]. At each follow-up visit, the subjects were interviewed and tested for the presence of NG, CT, and TV. Subjects who tested positive were promptly treated. They also received new diaries upon returning the completed ones. An important endpoint of the investigation was the timing of re-infections. In other words, the investigators were interested in estimating the survival function of recurrent STIs. This was of great clinical importance because it would help clinicians devise an STD testing strategy (for example, how often the subjects should be re-tested) for individuals at higher risk of repeated STIs.

We present data from a sample of five subjects in Figure 1. It should be noted that in this study, the true infection time was either interval censored or right censored. For example, as depicted in Figure 1, Subject 4 tested negative at her 1 month visit, but tested positive at her 5 month visit, indicating that an infection occurred at some point between 1 and 5 months. In our analysis, we are interested in time until the first infection by any of the STI's under consideration.

3. Methods

As discussed earlier, an STI occurs only when the host subject is engaged in unprotected or unsuccessfully protected sexual intercourse. With the auxiliary information provided by the behavioural diaries, we are able to limit the unknown infection time to those days on which such activities occurred, as recorded on 8 diary days during the 1–5 month interval for Subject #4. Barring further information on the reporting errors of condom use, we assume all unprotected and unsuccessfully protected sexual episodes lead to an STI with an equal probability. It should be noted that incorrect condom use or condom failures are not uncommon among adolescents [16–18]. Therefore in this application, we consider coital episodes with reported condom use as likely infection times, as illustrated by the case of Subject #4 in Figure 1. In situations where there is evidence suggesting unequal transmission probabilities, the resampling weights can be modified to reflect the differential STI risk.

Let T_i be the true failure (infection) time for the i th subject, where $i = 1, \dots, K$. Because of interval censoring, we do not directly observe T_i ; instead, we only know that the infection occurs inside a censoring interval $(L_i, U_i]$. Here L_i and U_i represent the negative pre-infection diagnostic test time and the positive post-infection diagnostic test time, respectively. Let C_i be the censoring time (the time at which the subject ends her diary) and $\Delta_i = I(U_i \leq C_i)$ be the positive test indicator. To understand the notation in the context of a typical STD investigation, we note that: (1) for an infected subject, $T_i < U_i \leq C_i$; (2) if a subject is tested positive at her first follow-up visit, $L_i = 0$ and U_i is the first follow-up visit time; and (3) for a censored subject, $C_i < U_i$.

In a typical STD investigation, the true infection time (T_i) of a subject is unknown. However, when a subject tests positive at a follow-up visit, we can use her diaries to trace the infection time to a set of coital episodes at which the infection could have been acquired. We denote such a set of possible infection times in an infected individual as $\mathbb{G}_i = \{E_{ij} : \Delta_i = 1 \text{ and } L_i < E_{ij} \leq U_i\}$, where E_{ij} is the j th coital episode time after L_i , $j = 1, \dots, n_i$. Theoretically, if a subject is not infected, the set of possible infection times should be empty, namely $\mathbb{G}_i = \emptyset$ if $\Delta_i = 0$. But for operational convenience, we let $\mathbb{G}_i = \{C_i : \Delta_i = 0\}$. To be more exact, we consider the true infection time for an infected subject i to be one of the coital episode times in $\mathbb{G}_i$; if the subject is censored, $\mathbb{G}_i$ will provide the censoring time. In other words, the elements of $\mathbb{G}_i$ form the support of the predictive distribution for the i th subject in our imputation scheme.

Apparently, disease incubation time could alter the composition of $\mathbb{G}_i$. With the known incubation time of a specific organism, we can easily include and exclude the potential infection times to accommodate the incubation period of the disease. But this is becoming increasingly unnecessary because newer DNA based assays, such as the polymerase chain reaction (PCR) tests, are able to detect infection that is still in the incubation phase. Therefore, without loss of generality, we treat the disease incubation period as 0. Also for simplicity, we consider only the first infection in this paper.

In this research, the infection time of a study subject is considered right censored if she remained uninfected at her last clinical visit. Since the clinical visit times are pre-scheduled at 1, 3, 5 and 7 months according to the study design, the censoring time would be independent of the true infection time, i.e. $T_i \perp\!\!\!\perp C_i$ for all i . We use the following algorithm to get an estimate and to construct pointwise confidence bands around the estimate of survival function $\hat{S}(t)$.

Algorithm

1. For the b th resampled data set, we impute uniformly one infection (or censoring) time for the i th subject from the set $\mathbb{G}_i$ and denote it as $X_i^{(b)}$. If a subject i is infected, $X_i^{(b)} = E_{ij}$ for some j (i.e. the time of one of the coital events in the infection interval). If a subject i is censored, $X_i^{(b)} = C_i$. When this process is completed for all K subjects, we have a complete right censored data set $\{X_i^{(b)}, \Delta_i^{(b)}; \text{ for } i = 1, \dots, K\}$ for the b th repetition.
2. With the data generated from Step 1, we compute the Kaplan–Meier survival curve estimate $\hat{S}^{(b)}(t)$ for the current data set. Let q be the number of infections; $t_1^{(b)}, t_2^{(b)}, \dots, t_q^{(b)}$ the resampled infection times; $N_r^{(b)}$ the number of infections at time $t_r^{(b)}$; $R(t_r^{(b)})$ the number of subjects at risk at time $t_r^{(b)}$, then we have $\hat{S}^{(b)}(t) = \prod_{r=1}^q \{1 - N_r^{(b)} / R(t_r^{(b)})\}$.
3. We repeat Steps 1–2 B times and combine the Kaplan–Meier estimates obtained for each data set:

$$\tilde{S}(t) = \frac{1}{B} \sum_{b=1}^B \hat{S}^{(b)}(t) \quad (1)$$

According to the proposed algorithm, if the failure times for all the infected subjects were known, the set $\mathbb{G}_i$ would contain only the true infection time, and the resulting estimate $\tilde{S}(t)$ would be the same as the Kaplan–Meier estimate for the right censored data. A key feature of our imputation method is to impute an interval censored infection time from a predictive distribution based on an infected subject's coital history, instead of imputing from the infection times of other subjects. Specifically, we impute the interval censored infection time of the i th subject by resampling uniformly the coital times within the infection interval. In other words, subject i cannot be infected at time $E_{i'k}$ if $i' \neq i$, unless $E_{ij} = E_{i'k}$ for some $j \in [1, \dots, n_i]$ and $k \in [1, \dots, n_{i'}]$.

Following the rules established in References [19, 20], we note that to construct confidence bands for $\tilde{S}^{(b)}(t)$, one needs to take into account two different sources of variation: the uncertainty of the estimation process itself due to the original sampling variation from the target population and the variability associated with the resampling process. Letting $V^{(b)}(t)$ denote the variance of $\tilde{S}^{(b)}(t)$, we write $\text{Var}(\tilde{S}(t)) = V^{(b)}(t) + (1 + B^{-1})V_{\text{bet}}(t)$, where $V^{(b)}(t) = \Sigma V^{(b)}(t)/B$ estimates the within-imputation variance and $V_{\text{bet}}(t) = \Sigma \{\hat{S}^{(b)}(t) - \tilde{S}^{(b)}(t)\}^2 / (B - 1)$ estimates the between-imputation variance.

An Splus program for the computation of the proposed estimator can be obtained at website <http://www.biostat.iupui.edu/Tech%20Papers/index.html>. For those who are interested in testing and extending the proposed procedure, we have made available a simulated data set that is similar in structure to the real study data. Interested readers are welcome to contact the authors for further details of the computational procedure and the study data.

4. Simulations

We conduct a comparative simulation study to evaluate the performance of the proposed method against several currently used single imputation methods. All simulations are executed by the use of S-Plus 6.1 (Insightful Corp., Seattle, WA).

The data generating mechanism is designed to mirror the motivating STD investigation. Details on the generation of visit, censoring, and coital episode times are described below:

1. For the i th subject, we first generate four follow-up visit times $V_{i1}, \dots, V_{i4}$, from either a uniform distribution $V_{iv} \sim \text{Uniform}(60 \times v - 5, 60 \times v + 5)$ or a triangular distribution $V_{iv} \sim \text{Triangular}(60 \times v - 20, 60 \times v + 20)$, for $i = 1, 2, \dots, n$ and $v = 1, 2, 3, 4$. In other words, we assume that there are four follow-up visits, around 60, 120, 180, and 240 days from the enrolment date.
2. Technically, any one of these follow-up visits could be the subject's last visit in the study because of early drop out. Therefore, we let c_v be the probability that the v th follow-up visit is the last visit, $v = 1, 2, 3, 4$. According to the discrete distribution of (c_1, c_2, c_3, c_4) , we select V_{iv} as the last visit time of the i th subject.
3. The true failure time for the i th subject, T_i , is then generated from Weibull(a, b), where a is the shape parameter and b is the scale parameter.
4. For the i th subject, we compare her infection time T_i with her last visit time V_{iv} . If $T_i \leq V_{iv}$, we let the observed survival time $Y_i = T_i$ and set the failure indicator $\Delta_i = 1$; if $T_i > V_{iv}$, we let $Y_i = V_{iv}$ and set $\Delta_i = 0$.
5. To generate the coital event times in an interval for an infected subject, we first generate the number of events N_i in the interval from the binomial distribution with the mean equal to 8 and the variance equal to 7 (parameters are chosen to mirror our experience with the target population). One of the coital times is the true infection time, while the rest of the coital episodes are generated within the interval $(V_{iv-1}, V_{iv}]$ under one of the following distributions: (a) $\text{Uniform}(V_{iv-1}, V_{iv}]$ or (b) a mixed episode distribution where approximately half of the events are from the aforementioned uniform distribution and the rest of the events are from the binomial distribution centred around the true event time.

Under this data simulation scheme, we generate 500 data sets with a 100 subjects, and 200 data sets with 500 subjects. For each simulated data set, we create 20 resampled data sets according to the proposed algorithm.

The parameter settings for survival distributions and censoring proportions are chosen to represent a variety of data situations. For example, the values of the Weibull shape parameter are selected to represent increasing ($a = 1.2$), decreasing ($a = 0.8$), and constant hazards ($a = 1$). For a given shape parameter, $c_1, \dots, c_4$ are chosen to control the proportion of right censoring. In the simulation, we considered three levels of right censoring: light (20 per cent), moderate (35 per cent) and heavy (50 per cent), implying the percentages of interval censoring at 80, 65, and 50 per cent, respectively.

Each simulation run corresponds to a unique combination of the following values:

sample size $\in \{ 100, 500 \}$

right censoring $\in \{ 20, 35, 50 \text{ per cent} \}$

shape parameter $\in \{ 1.2, 1, 0.8 \}$

The first nine rows of Table I corresponds to sample size $n = 100$. Among them, the first three rows correspond to right censoring of 20 per cent with the Weibull shape parameters $a = 1.2, 1$, and 0.8 , respectively. The last nine rows represent the data situations of $n = 200$ and they follow the same parameter sequence.

We compare the finite sample performance of several currently used imputation methods and self-consistent estimators for interval-censored data: (I) the Kaplan–Meier estimate when the interval censored infection times are *observed*, which represents the hypothetical situation of *no interval censoring*; (II) single imputation with the mid-point of the diagnostic interval; (III) single imputation with the end of the interval; (IV) single imputation with the first coital episode; (V) single imputation of the mid-point between the first and last coital episodes; (VI) single imputation of the last coital episode within the interval; (VII) self-consistent estimator using interview data; (VIII) self-consistent estimator using information about first and last coital episode in the infection interval; and (IX) resampling based estimator $\hat{S}(t)$. Since the true survival function is known in the simulations, we report the mean integrated squared error (MISE) for each of the tested methods, defined as $MISE = E[\int_t [S(t) - \hat{S}(t)]^2 dt]$.

In order to make meaningful comparisons among all the estimation methods, we restrict the calculation of MISE to the interval on which all the methods provide numerical estimates. For the imputation methods (I–VI, IX), the restrictions can be expressed as the minimum time at which the estimator is defined. Self-consistent estimators are flat with jumps on a discrete set of equivalence classes. If all equivalence classes contain single points we have a well-defined estimator for the whole time period under consideration. However, if an equivalence class is an interval, the self-consistent estimator changing by an appropriate amount within the equivalence class will not change the likelihood. In order to compare the self-consistent estimators with the imputation methods, we assumed that they change linearly in the interior of the equivalence classes.

In this research, we have examined the performance of the proposed method under various coital episodes and visit time distributions. Because the method performs consistently under alternative distributional assumptions, we only report the simulation results for one setup with mixed episode distribution and uniformly perturbed visit times in Table I. Simulation results show that the resampling based multiple imputation method (IX) consistently outperforms its competitors and has MISE that is only a fraction of those of the single imputation methods. The ratios of estimated MISEs between single imputation methods and Method IX are between 2.1 and 9.4 when the sample size is 100. With a larger sample size of 500 the ratios are between 3.3 and 57.6. Among the single imputation methods, Method V is consistently better than other methods, suggesting the potential contributions of the diary information. The self-consistent estimator compared favourably with the single imputation methods, but our multiple imputation method had a much smaller estimated MISE.

We noted that in the simulated data with uniformly perturbed visit times, the MISE for the self-consistent estimator based on interview data was at times smaller than that based on encounter data (Table I). Such seemingly counterintuitive results could be explained by the smaller number of equivalence classes and hence larger jump sizes for the self-consistent estimator based on interview data. For example, we had to use linear interpolation for approximately 50

per cent of the time points for Method VII as compared to approximately 20 per cent of the time points for Method VIII.

It is also interesting to note that the proposed method is comparable with Kaplan–Meier estimates *when the interval censored infection times are directly observable*. Such empirical evidence highlights the advantages of the proposed method.

To visually assess the relative performance of the considered estimation methods, we present a sample of estimated curves in Figure 2.

5. Analysis of STD Diary Data

As mentioned in Section 2, infection status and behavioural diaries were obtained from 155 adolescent women, 77 per cent of the whom were African American. During the follow-up period 78 women experienced at least one STI. Subjects enrolled in the study had on average 1.3 coital episodes per week, approximately 60 per cent of which were unprotected. In this example, we consider all the coital episodes in the infection interval as possible infection times because protection failures are very common in the target population [16–18]. The main objective of the analysis is to estimate the survival function from the treatment of the initial STI to the first subsequent re-infection.

Estimates of the survival function are obtained using the resampling algorithm described in Section 3, self-consistent estimator and various *ad hoc* single imputation methods considered in Section 4. For the proposed method, we use $B = 50$ multiply imputed data sets. The top left panel in Figure 3 shows a comparison between the multiple imputation method and the self-consistent estimator. The remaining panels contain comparisons of the survival function estimates under other imputation methods considered in the paper.

Comparing the survival curve estimates, we note that: (1) the resampling based method produces a smoother survival curve due to the multiply imputed infection times; (2) the proposed method is able to produce a confidence band (shaded bands in Figure 3) while the directly obtained confidence bands from single imputation methods tend to underestimate the variability associated with the estimation [20]; (3) single imputation methods might produce biased estimates. The over-estimation of the survival function is particularly salient when the end of the diagnostic interval is used (Method III). The biases in the survival function estimates that are associated with the first or last coital episodes (Methods IV and VI, respectively) are relatively small (but still visible in Figure 3); (4) the mid-interval and mid-coital episode methods (Methods II and V) are essentially unbiased; (5) the self-consistent estimator exhibits big jumps, especially around the scheduled visit times of 30 and 210 days.

The estimated survival function shows a steady decline over time in the probability that a subject stays uninfected. The decline is slightly steeper in the first 3 months. If re-tests are recommended 90 days after the treatment of the previous infection and are duly implemented, roughly 40 per cent of the subjects will test positive and thus will require repeated treatment.

6. Discussion

Multiple imputation procedures have recently been used to analyse interval censored survival data [5,6]. However, existing methods often require the specification of baseline survival (or hazard) function. While the mis-specification of a baseline survival function may not be consequential in Cox regression analysis because of the proportionality assumption of the baseline hazard, using an incorrectly specified baseline hazard in survival function estimation can be difficult to justify and may lead to questionable results. In this manuscript, we propose a procedure that takes advantage of the auxiliary information provided by daily behavioural

diaries. Using a resampling based approach to estimate the survival function of an infection process in interval censored situations is an appealing idea for several reasons: (1) it is able to incorporate the important behavioural information provided by STD diaries; (2) simulation study suggests that among the methods that utilize the diary information (such as Methods IV, V, and VI), the resampling based methods are unbiased and tend to have drastically reduced MISE, suggesting a substantial gain in efficiency; (3) the resampling based methods are able to produce valid confidence intervals for the survival function; (4) the proposed approach is also more biologically defensible because the resampled event times represent the potential true infection times, as opposed to arbitrarily imputed time points at which sexual contacts may or may not have occurred.

As the first attempt to use diary data in the estimation of a survival function with interval censored data, our method has a number of limitations. Despite the introduction of the auxiliary behavioural information into the estimation procedure, the proposed approach gives a simplified presentation of a rather complicated process. For example, the continuous process of sexual encounters is treated as discrete by the resampling scheme. Diary data are also accepted as complete and accurate recordings of sexual behaviours without consideration of the potential measurement errors. In addition to these general limitations, the proposed method also relies on the implicit assumption that *given an infection has been observed*, all sexual encounters within the interval are equally likely to be the first infection point. This assumption allows us to resample the potential infection times from the coital episodes recorded in set $\mathcal{G}_i$ with equal resampling probabilities. But the resampling scheme also has some clear limitations: (1) It does not accommodate the effect of condom use, nor does it account for potentially differential STI risks posed by different partners. (2) The method uses discrete uniform distributions to approximate the truncated geometric distributions for the description of the conditional probability that the first disease transmission occurs at a specific encounter within the censoring interval, given that an STI has occurred. The uniform approximation works well as long as the per-encounter STI risk is reasonably small. In our data application, the per-encounter STI risk is very small because of the relatively low transmissibility of the bacterial pathogens and the low prevalence rates of the diseases among the partners of the study subjects. But as a general methodology, modifications of the resampling procedure are likely needed in different data situations. Currently, we are expanding the investigation to examine the procedure's sensitivity to non-uniformly distributed STI risks within the interval and to test procedures using unequal resampling weights for coital episodes with or without condom protection. In summary, our paper has introduced a general methodological framework for the incorporation of behavioural diary information into survival analysis of interval censored data. The current procedure is carried out under a simplified resampling plan with equal weights. We have presented empirical evidence that the method performs well in our data application. Further modifications and extensions of the procedure within the methodological framework are certainly possible and are likely to improve the performance of the proposed survival function estimator.

Finally, we note that although our method is proposed in the context of bacterial STD infections, this method can be used in other clinical or epidemiological investigations where interval censoring arises. For example, it is conceivable that the same technique can be used to study the HIV/AIDS infections among intravenous (IV) drug users where the true infection times are among the IV injection times.

Acknowledgements

This research is supported by grant RO1 HD042404 from the National Institute of Child Health and Human Development (NIH/NICHD). The authors are grateful to an anonymous reviewer for his or her many constructive and insightful comments.

Contract/grant sponsor: National Institute of Child Health and Human Development (NIH/NICHHD); contract/grant number: RO1 HD042404

References

1. Peto R. Experimental survival curves for interval-censored data. *Applied Statistics* 1973;22:86–91.
2. 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.
3. Finkelstein DM. A proportional hazards model for interval-censored failure time data. *Biometrics* 1986;42:845–854. [PubMed: 3814726]
4. Kooperberg C, Stone CJ. Log-spline density estimation for censored data. *Journal of Computational and Graphical Statistics* 1992;1:301–328.
5. Pan W. A multiple imputation approach to Cox regression with interval-censored data. *Biometrics* 2000;56:199–203. [PubMed: 10783796]
6. Pan W, Chappell R. Estimation in the Cox proportional hazard model with left-truncated and interval censored data. *Biometrics* 2002;58(1):64–70. [PubMed: 11890328]
7. Sinha D, Chen MH, Ghosh SK. Bayesian analysis and model selection for interval-censored survival data. *Biometrics* 1999;55(2):585–590. [PubMed: 11318218]
8. Calle ML, Gomez G. Nonparametric Bayesian estimation from interval-censored data using Monte Carlo methods. *Journal of Statistical Planning and Inference* 2001;98(1–2):73–87.
9. Lindsey JC, Ryan LM. Tutorial in biostatistics: methods for interval-censored data. *Statistics in Medicine* 1998;12:219–238. [PubMed: 9483730]
10. Taylor JMG, Murray S, Hsu CH. Survival estimation and testing via multiple imputation. *Statistics and Probability Letters* 2002;58:221–232.
11. Fortenberry JD, Tu W, Harezlak J, Katz BP, Orr DP. Condom use as a function of time in new and established adolescent sexual relationships. *American Journal of Public Health* 2002;92(2):211–213. [PubMed: 11818293]
12. Odell PM, Anderson KM, D'Agostino RB. Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics* 1992;48:951–959. [PubMed: 1420849]
13. Dorey FJ, Little RJ, Schenker N. Multiple imputation for threshold-crossing data with interval censoring. *Statistics in Medicine* 1993;12:1589–1603. [PubMed: 7694349]
14. Katz BP, Fortenberry JD, Tu W, Harezlak J, Orr DP. Condom use patterns of adolescent women as reported in personal diaries. *Sexually Transmitted Diseases* 2001;28(5):247–251. [PubMed: 11354261]
15. Fortenberry, JD.; Cecil, H.; Zimet, GD.; Orr, DP. Concordance between diary and self-report questionnaires of sexual behaviors of adolescent women. In: Bancroft, J., editor. *Researching Sexual Behavior*. Indiana University Press; Bloomington, IN: 1997. p. 237–249.
16. Macaluso M, Kelaghan J, Artz L, Austin H, Fleenor M, Hook EW, Valappil T. Mechanical failure of the latex condom in a cohort of women at high STD risk. *Sexually Transmitted Diseases* 1999;26(8):450–458. [PubMed: 10494936]
17. Rosenberg MJ, Waugh MS. Latex condom breakage and slippage in a controlled clinical trial. *Contraception* 1997;56(1):17–21. [PubMed: 9306027]
18. Lindberg LD, Sonenstein FL, Ku L, Levine G. Young men's experience with condom breakage. *Family Planning Perspectives* 1997;29(3):128–133. [PubMed: 9179582]
19. Rubin DB, Schenker N. Multiple imputations in health-care database: an overview and some applications. *Statistics in Medicine* 1992;10:585–598. [PubMed: 2057657]
20. Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. 2nd. Wiley Interscience; New York: 2002.

Figure 1.

Coital events and transmission of STDs for a sample of five subjects from an STD study of adolescent females. We represent by ■—negative STD test; ◆—positive STD test; capital letters (A, B, aso)—condom protected coital events with Partners A, B, aso; small letters (a, b, aso)—unprotected coital event with Partners A, B, aso.

Figure 2.

Survival function estimates using simulated data: (a) sample size of 100, Weibull shape parameter 1.2, 35 per cent right-censoring rate, uniformly perturbed visit distribution; and (b) sample size of 100, Weibull shape parameter 0.8, 20 per cent right-censoring rate, triangularly perturbed visit distribution. Method II refers to the single imputation with the mid-point of the diagnostic interval, Method V to the single imputation of the mid-point between the first and last coital episodes, and Method IX to the new resampling based estimate.

Figure 3.

Diary data analysis. Top left panel presents the comparison of the multiple imputation method (IX) with the self-consistent estimator (VII). The remaining panels present the comparisons of the multiple imputation method (IX) with single imputation methods (I—hypothetical situation with no interval censoring, II–VI single imputation methods, IX—multiple imputation method). Shaded region depicts the estimated confidence band for the survival curve estimated using Method IX.

Table 1 Mean integrated squared error (MISE) for the simulation with mixed episode distribution and uniformly perturbed visit times.

Setting	Method								
	I	II	III	IV	V	VI	VII	VIII	IX
1	38.9	129.6	433.0	214.8	82.6	229.2	100.7	109.8	34.0
2	42.1	137.0	504.7	188.8	88.8	269.7	101.1	112.3	37.8
3	42.1	162.2	597.1	168.5	107.2	326.3	106.5	126.4	41.3
4	43.4	131.4	386.7	214.5	85.4	217.8	103.6	110.3	37.3
5	44.8	137.6	449.8	184.8	90.7	255.6	106.1	113.9	39.5
6	44.1	161.9	559.3	160.1	107.1	316.7	111.2	125.5	42.1
7	77.0	159.5	310.5	233.1	116.0	204.5	145.3	134.9	63.7
8	82.8	171.4	395.7	220.1	125.9	239.3	155.4	142.3	68.9
9	104.7	212.4	580.2	227.4	159.4	311.4	177.4	166.4	88.6
10	8.4	101.0	412.9	179.5	53.1	207.0	21.1	60.9	7.8
11	8.9	107.1	472.4	161.4	57.9	236.7	26.0	64.6	9.0
12	9.1	132.5	566.1	139.8	75.7	294.1	36.8	77.2	12.1
13	9.8	101.8	369.3	171.0	54.6	196.9	27.0	63.3	9.5
14	9.2	107.6	416.7	160.1	58.2	219.5	26.7	65.6	8.9
15	10.7	133.7	531.4	133.8	77.5	286.7	42.2	80.0	13.6
16	17.8	107.1	280.4	157.6	61.8	166.4	46.0	74.6	15.5
17	19.2	113.5	365.1	138.8	65.8	195.5	47.8	76.4	16.9
18	24.8	145.1	584.0	128.8	90.5	264.6	69.9	93.8	25.0