A Two-Sample Test with Interval Censored Data via Multiple Imputation

(A98–124R)

Wei Pan
Division of Biostatistics
School of Public Health
University of Minnesota
A460 Mayo Building (Box 303)
Minneapolis, MN 55455-0378, USA
weip@biostat.umn.edu

Revised February 1998
A Two-Sample Test with Interval Censored Data via Multiple Imputation

SUMMARY

Interval censored data arise naturally in large scale panel studies where subjects can only be followed periodically and the event of interest can only be recorded as having occurred between two examination times. In this paper we consider the problem of comparing two interval-censored samples. We propose to impute exact failure times from interval-censored observations to obtain right censored data, then apply existing techniques, such as Harrington and Fleming’s $G^p$ tests to imputed right censored data. To appropriately account for variability, a multiple imputation algorithm based on the Approximate Bayesian Bootstrap (ABB) is discussed. Through simulation studies we find that it performs well. The advantage of our proposal is its simplicity to implement and adaptability to incorporate many existing two-sample comparison techniques for right censored data. The method is illustrated by reanalyzing the Breast Cosmesis Study dataset.
1. INTRODUCTION

Recently survival analysis with interval censored data has been attracting a lot of research attention. With interval censoring the event of interest (such as infection with a disease or other failure) is not observed exactly but only known to happen between two examination times (but may also happen before the first examination or after the last examination). Hence interval-censoring is a natural generalization of the commonly encountered right-censoring or double-censoring. Interval censored data arise often from large-scale panel studies, such as in AIDS studies, where the HIV infection time can only be detected at periodic follow-ups. Some of the subjects may have missed some scheduled follow-up examinations and come back later with a changed clinical status. Peto gives a Newton-Raphson algorithm to compute the nonparametric maximum likelihood estimator (NPMLE) of the survival function for interval censored data. Turnbull introduces a self-consistency algorithm, which is a special case of the EM algorithm, to compute the NPMLE of the survival function for arbitrarily censored and truncated data. Groeneboom and Wellner give a detailed treatment of the NPMLE of the survival function, including its computational algorithms and large-sample properties.

The problem we will consider here is two-sample comparison. This is particularly relevant in clinical trials where new medical treatments need to be evaluated. Throughout we are only interested in non- or semi-parametric approaches for interval censored data. Suppose two iid samples \( \{X_{i1}\}_{i=1}^{n_1} \) and \( \{X_{i2}\}_{i=1}^{n_2} \) are respectively from two (unknown) survival functions \( S_1(x) \) and \( S_2(x) \). The null hypothesis to be tested is \( H_0: S_1 = S_2 \) when \( \{X_{ij}\} \) are subject to (interval) censoring. This problem has been extensively studied for right censored data, where the most famous are probably the logrank test and the generalized Wilcoxon test (e.g. Kalbfleisch and Prentice). These two tests can be viewed as two members of a more general class of linear rank tests, the \( G^0 \) tests proposed by Harrington and Fleming. In addition to these linear rank tests, another important class of statistics is based on the integrated weighted difference in Kaplan-Meier estimators, which is more sensitive against the alternative of stochastic ordering.

Some related works have appeared in the literature for two-sample comparison with interval-censored data. Finkelstein derives a score test under the Cox proportional hazards model, which is also closely related with the logrank test. Sun gives a non-parametric test for discrete survival times. Fay proposes some rank invariant tests. Petroni and Wolfe generalize Pepe and Fleming's
weighted Kaplan-Meier (WKM) tests to interval censored data. Pan compares Mantel's test, Peto and Peto's rank invariant tests and Petroni and Wolfe's test. A common difficulty with interval censored data is how to estimate the (asymptotic) variance of a test statistic, though the observed information matrix from the nonparametric maximum likelihood function is often used. The observed information matrix often has a high dimension that is in the order of the sample size. Thus computationally it is not easy unless the sample size is small or the observed failure times are grouped. Furthermore, the use of the observed information matrix needs more theoretical justification. Notice that the asymptotic distribution theory of the NPMLE of the survival function has not completely been settled. Here the classic MLE theory (e.g., see Lehmann) cannot be applied directly since the number of the unknown parameters to be estimated is increasing with the sample size.

In this paper, we propose a general method based on multiple imputation for interval censored data. The basic idea is simple. If we can impute from each interval censored (but not right censored) observation, we can take advantage of existing techniques for right censored data, such as the $G^p$ tests and WKM tests. Multiple imputation also has an advantage of potentially being able to take account of the between imputation variability. We consider one multiple imputation technique, the Approximate Bayesian Bootstrap (ABB). We concentrate on using the $G^p$ test statistics, leaving it open to investigate applying other two-sample comparison techniques for right-censored data.

Multiple imputation has proved useful in many statistical applications. However, to our knowledge, there are only a few applications of multiple imputation to censored failure time data. The first seems to be Wei and Tanner's work in linear regression for right censored data. Recently Pan and Kooperberg applied multiple imputation to bivariate linear regression for right censored data. We are currently investigating its application to regression problems with interval censored data.

This paper is organized as follows. We first briefly introduce some background on interval-censoring and multiple imputation in Section 2. In Section 3 we propose a multiple imputation based method to compare interval-censored samples. In Section 4 simulation studies were conducted to assess its performance under different configurations. An example is taken from the Breast Cosmesis Study for illustration in Section 5, followed by a short discussion.
2. BACKGROUND

2.1 Interval Censoring

Because of interval-censoring, we cannot observe the event time of interest, say survival time \( \{X_i\} \), directly. Instead we only have observations \( \{(U_i, V_i)\} \), which constitute the endpoints of censoring intervals. It is only known that the survival time \( X_i \) satisfies \( 0 \leq U_i < X_i \leq V_i \leq \infty \) for \( i = 1, \ldots, n \). Throughout this paper as usual we assume that the censoring mechanism is independent of the distribution of the underlying survival times.

Suppose that the survival function (i.e., one minus the cumulative distribution function) of \( X_i \) is \( S(x) \). Given the observations \( \{(U_i, V_i)\} \), the likelihood can be written down as:

\[
L(S|\{(U_i, V_i)\}) = \prod_{i=1}^{n}(S(U_i) - S(V_i)).
\]

The nonparametric maximum likelihood estimator (NPMLE) \( \hat{S} \) of \( S \) is the maximizer of the above likelihood function with the constraint that \( \hat{S} \) is nondecreasing and bounded between 0 and 1. Turnbull\(^3\) shows that \( \hat{S} \) can have probability mass only in some intervals with end points all from \( \{U_i, V_i\} \). We follow Groeneboom and Wellner\(^5\) to further restrict the NPMLE \( \hat{S} \) as being piecewise constant with possible discontinuities only at the end points of those intervals, and as being right-continuous.

A conceptually attractive EM algorithm was proposed by Turnbull\(^3\). But computationally it is more efficient to use the iterative convex minorant algorithm\(^5\), which we will use throughout the paper. Groeneboom and Wellner have shown that the NPMLE is consistent under suitable conditions.

The NPMLE from interval censored data tends to have a smaller number of jumps and hence larger jump sizes than the usual square-root-\( n \) estimators, such as the empirical distribution function for complete data or the Kaplan-Meier estimator for right censored data. In imputing for an interval-censored observation, we use a linear smooth of the NPMLE, which is an analog of Link's estimator for right censored data (see Miller,\(^{20}\) p.134). The linear smooth of the NPMLE is obtained by connecting neighboring two jumps of the NPMLE by a straight line; in other words, it has a piecewise uniform distribution between any two neighboring jumps of the NPMLE, and it is a continuous piece-wise linear function. Note that like the NPMLE, the linear smooth of the NPMLE is also
consistent under suitable conditions since the jump size of the NPMLE will shrink to zero (such
that the NPMLE tends to the underlying continuous true survival function) as the sample size
increases. More will be given in the next section on why we smooth the NPMLE.

2.2 Multiple Imputation

Now we briefly review the multiple imputation in a general missing data context. Its connection
to our current setting will be explored later.

A common method of handling missing data is to impute a value of each missing datum.
This single imputation technique allows standard methods applied to the resulting complete data
set. The major drawback of the single imputation is that the application of standard methods
to the imputed data set treats the missing values as if they were known as imputed, hence may
under-estimate the true variability due to not knowing the missing values. Multiple imputation
was thus proposed by Rubin to retain the advantage of the single imputation but overcomes its
shortcoming by replacing each missing value with two or more values representing a distribution
of likely values. Therefore, Multiple imputation leads to multiple complete data sets. Analyses of
these multiple complete data sets have the potential to take proper account of the variability due
to the missingness in the original data set.

More formally, suppose that a parameter of interest $\theta$ is estimated as $\hat{\theta}$ from a complete data
set. The inference of $\theta$ can then be based on the distribution of $\hat{\theta}$. Further we assume that $\hat{\theta}$ is
(asymptotically) normal, $N(\theta, \Sigma)$. Now from our given incomplete data, we use multiple imputation
to obtain multiple, say $M$, sets of complete data, and hence $M$ estimates, $(\hat{\theta}_m, \hat{\Sigma}_m)$, $m = 1, ..., M$.
The normal-based inference is based on

$$\hat{\theta}_s \sim N(\theta, \hat{\Sigma}_s),$$

where

$$\hat{\theta}_s = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m / M$$

is the estimate of $\theta$ based on the $M$ imputations, and

$$\hat{\Sigma}_s = \frac{1}{M} \sum_{m=1}^{M} \hat{\Sigma}_m + (1 + \frac{1}{M}) Var(\hat{\theta}_1, ..., \hat{\theta}_M)$$

is the estimate of the variance, where the first term is the average of the within-imputation variances
and the second term is the between-imputation variance. The between-imputation variance is
inflated by a factor of $1/M$ to reflect extra variability of using a finite $M$ instead of an infinite number of imputations.

For small $M$, it may be more accurate to use $t$-based inference. Specifically, we use

$$
\frac{\hat{\theta} - \theta}{\sqrt{\sum_i}} \sim t_\nu,
$$

where the degree of freedom $\nu$ is given by

$$
\nu = \left[1 + \left(\frac{M}{M+1}\right)\frac{\hat{W}}{\hat{B}}\right]^2 (M - 1),
$$

$$
\hat{W} = \frac{1}{M} \sum_{m=1}^{M} \hat{\sigma}_m, \quad \hat{B} = \text{Var}(\hat{\theta}_1, \ldots, \hat{\theta}_M).
$$

In our implementation, we used moderate $M = 5$ or $M = 10$ and found that the $t$-based inference is almost the same as the normal-based one. Hence, in the sequel, we will only use the normal-based procedure.

3. NEW METHOD WITH INTERVAL CENSORED DATA

3.1 General Idea

Our proposal is to impute from finite-interval-censored but not right-censored observations. Specifically, for each interval censored observation $(U_i, V_i)$, we replace it with an event time $T_i$ consonant with $(U_i, V_i)$ (i.e. $U_i < T_i \leq V_i < \infty$), and a non-censoring indicator $\delta_i = 1$ (and we will discuss how to do it later); for each right censored data $(U_i, \infty)$, keep it as $(U_i, \delta_i = 0)$. Thus we obtain a set of right censored data $\{(T_i, \delta_i)\}$, to which many existing techniques, such as the $G$ tests can be applied. Apparently, the true variability of a test statistic $G$ obtained from the original observations should be larger than that from using the imputed observations. It is well-known that such single imputation usually under-estimates the variability due to incomplete data. To remedy this, multiple imputation has since been proposed.\textsuperscript{17}

3.2 A Method Based on Approximate Bayesian Bootstrap

In our current context, the observed data are censoring intervals, and the underlying “complete” data are imputed survival times and observed right-censoring times. Here we do not need to impute right-censored survival times. For simplicity we still use $T$ to represent a resulting right censored
data set (and ignore the censoring indicators). From now on, when we say imputing from interval censored data, we only mean those censored in a finite interval, hence we do not impute from right censored data.

The multiple imputation scheme we use is the Approximate Bayesian Bootstrap (ABB), which was originally proposed by Rubin\textsuperscript{21} as a simple imputation technique to approximate the Bayesian Bootstrap for categorical data.\textsuperscript{22} Efron\textsuperscript{23} gives a more general formulation (see his expression (3.7)), which we follow here. The algorithm works as follows:

1. For $m = 1, \ldots, M$ do
   a. Draw a bootstrap sample $\{(U_{ij}^{(m)}, V_{ij}^{(m)})\}$ from each of the two given samples $\{(U_{ij}, V_{ij})\}$, $i = 1, \ldots, n_j$ and $j = 1, 2$.
   b. Estimate the survival functions from two bootstrap samples as $\hat{S}_1^{(m)}$ and $\hat{S}_2^{(m)}$ respectively.
   c. For each of the two original samples, generate a set of possibly right-censored observations $T_{ij}^{(m)}$ as follows: for $j = 1, 2$ and $i = 1, \ldots, n_j$, 1) if $(U_{ij}, V_{ij})$ is a finite interval, sample $X_{ij}$ from the distribution $\hat{S}_j^{(m)}$, conditional on that $\{U_{ij} < X_{ij} \leq V_{ij}\}$ and let $T_{ij}^{(m)} = X_{ij}$ (and $\delta_{ij}^{(m)} = 1$); 2) otherwise, keep right-censored observation and let $T_{ij}^{(m)} = U_{ij}$ (and $\delta_{ij}^{(m)} = 0$).

2. Using each $\{T_1^{(m)}, T_2^{(m)}\}$ to calculate a test statistic $\hat{G}^{(m)}$ and its variance estimate $\hat{\Sigma}^{(m)}$.

3. Let
   $$\hat{G} = \frac{1}{M} \sum_{m=1}^{M} \hat{G}^{(m)}, \quad \hat{\Sigma} = \frac{1}{M} \sum_{m=1}^{M} \hat{\Sigma}^{(m)} + \left(1 + \frac{1}{M}\right) Var(\hat{G}^{(1)}, \ldots, \hat{G}^{(M)}).$$

4. Calculate the p-value using $\hat{G}$ and its null distribution $N(0, \hat{\Sigma})$.

An interesting feature of the above algorithm is that it is non-iterative. In principle, we can use any reasonable estimator of the survival function in Step 1(b). In our implementation, we used a linear smooth of the NPMLE. Intuitively, the larger the number of imputations $M$ in Step 1, the more accurate the resulting estimate. But we find that the number of imputations $M$ does not need to be large; usually $M = 5$ or $M = 10$ suffices. If for each set of right-censored data, $\hat{G}^{(m)}$ is approximately (e.g. asymptotically) $N(0, \hat{\Sigma}^{(m)})$ under the null hypothesis, then our final statistic $\hat{G}$ is also approximately $N(0, \hat{\Sigma})$ under the null hypothesis, from which we can calculate the p-value.
in Step 4. Here, we do not need to restrict the use of the test statistic $G$. In principle, it is valid to use any two-sample statistic $G$ for right censored data if it is asymptotically normal. In this paper, we concentrate on using the class of $G^p$ test statistics for its generality. In particular, $\rho = 0$ or 1 corresponds to the logrank test or Prentice’s generalized Wilcoxon test respectively.

As mentioned earlier, we used a linear smooth of the NPMLE in Step 1(b). One purpose is to increase the variability of imputed event times. Another consideration is due to the ABB scheme. Since we use a bootstrap sample in Step 1(a), there is no guarantee that any observation from the original sample will also be in the bootstrap sample. Therefore, for an observation $(U_{ij}, V_{ij})$, it is not necessary that the NPMLE $\hat{S}_j^{(m)}$ will put any probability mass in the interval $(U_{ij}, V_{ij}]$; thus there is no way to impute for $(U_{ij}, V_{ij})$ in Step 1(c) if $\hat{S}_j^{(m)}(U_{ij}) = \hat{S}_j^{(m)}(V_{ij})$. A linear smooth of the NPMLE overcomes this difficulty. Of course, other more sophisticated smoothes may also be applicable. But for its simplicity we choose the linear smooth.

4. SIMULATION

4.1 Simulation procedure

First we need to generate some interval censored samples. To mimic the pattern of many panel surveys the time interval between two examinations is taken to be constant, $\ell en = 0.25$. After the baseline examination, there are four follow-up examinations. Suppose $\tau_1$ is the (random) baseline examination time, the four follow-up times are $\tau_k = \tau_1 + (k - 1) * \ell en$, $k = 2, 3, 4, 5$. But a subject may miss the four follow-ups with probabilities 0.3, 0.3, 0.4 and 0.4 respectively. Denote $\tau_0 = 0$ and $\tau_6 = \infty$. To facilitate an easy description, we stipulate that any subject will not miss any “examinations” at $\tau_0$, $\tau_1$ and $\tau_6$. A random sample is generated as follows:

Step 0. For $i = 1$ to $n$ repeat Step 1 and Step 2:

Step 1. The survival time $X_i$ and the baseline examination time $C_i$ are generated independently from some specified distributions;

Step 2. Let $\tau_1 = C_i$, and calculate other $\tau_k$ as described above. We obtain an interval-censored observation $(U_i, V_i)$, where $U_i = \tau_j$ and $V_i = \tau_k$ for some $0 \leq j < k \leq 6$ and $(\tau_j, \tau_k)$ is the shortest interval covering $X_i$ such that the subject did not miss the examinations at $\tau_j$ and $\tau_k$.

The distribution of $C_i$ is always taken as $\text{Uniform}(0, \alpha)$, where $\alpha$ is chosen such that about 25% observations are right-censored.
Our simulation was conducted under the proportional hazards model (PHM) and non-PHM, as in Fleming et al.\(^{24}\) In the PHM, the distribution of the first group is taken to be \(W(a)\) (i.e. Weibull with scale and shape parameters 1 and \(a\) respectively); specifically, \(S_1(x) = \exp(-x^a)\), and \(S_2(x) = S_1(x)^\beta\). In addition to \(\beta = 1\), three non-unit values, 1.73, 2 and 2.25 are examined. For non-PHM, piece-wise exponential distributions are applied, where the relation between two survival curves are characterized by early difference, late difference and crossing hazards difference. Specifically, suppose that \(\lambda_j(t)\) is the corresponding hazard function for \(S_j(t)\), \(j = 1, 2\). Then the non-PHM configurations are:

- **Early difference:**
  \[
  \lambda_1(t) = .25, \lambda_2(t) = .75 \text{ if } t \in (0, .75]; \lambda_1(t) = \lambda_2(t) = 1 \text{ if } t \in (.75, \infty).
  \]

- **Late difference:**
  \[
  \lambda_1(t) = \lambda_2(t) = 1 \text{ if } t \in (0, .5]; \lambda_1(t) = 1, \lambda_2(t) = 2 \text{ if } t \in (.5, \infty).
  \]

- **Crossing hazard a):**
  \[
  \lambda_1(t) = .5, \lambda_2(t) = 1.5 \text{ if } t \in (0, .75]; \lambda_1(t) = 1.5, \lambda_2(t) = .5 \text{ if } t \in (.75, 1.5]; \lambda_1(t) = \lambda_2(t) = 1 \text{ if } t \in (1.5, \infty).
  \]

- **Crossing hazard b):**
  \[
  \lambda_1(t) = .5, \lambda_2(t) = 1.5 \text{ if } t \in (0, .5]; \lambda_1(t) = \lambda_2(t) = .1 \text{ if } t \in (.5, 1.25]; \lambda_1(t) = 1.5, \lambda_2(t) = .5 \text{ if } t \in (1.25, 1.75]; \lambda_1(t) = \lambda_2(t) = 1 \text{ if } t \in (1.75, \infty).
  \]

The configurations are also shown in Figure 1.

The NPMLE of the survival function is computed by the iterated convex minorant algorithm,\(^5\) which is much faster than the EM algorithm. All the programs were implemented in SPLUS except that the iterated convex minorant algorithm in computing the NPMLE was written in C and callable from SPLUS functions. In particular, we used SPLUS function `survdiff()` to calculate the \(G^0\) statistics.

### 4.2 Simulation results

For ABB data augmentation, there is not much difference between using \(M = 5\) and \(M = 10\) imputations (Table 1). They all yield reasonable size. For power properties, as for right-censored data, the logrank test (i.e. \(G^0\)) is the most powerful for PHMs and in detecting the late difference;
however, the Wilcoxon test (i.e. $G^1$) is the most powerful with crossing hazards. Note that the difference between two survival curves appears at early times for each of two crossing hazards configurations (Figure 1). For configuration “early difference”, the logrank test has slightly higher power than the Wilcoxon test does. This seems to contradict the well-known result of high power of the Wilcoxon test for an early difference. However, in Figure 1d), it appears that the difference between two survival curves persists from the beginning to the end. Hence, the conclusion is consistent with that for right censored data. The logrank test is powerful in detecting the late difference between two survival curves while the Wilcoxon test is good at detecting the early difference. The $G^{1/2}$ test always lies between the logrank and the Wilcoxon tests. When we used the larger sample sizes $100+100$, similar results (not shown) were obtained and the power of each test increased.

5. EXAMPLE

Now we apply the new method to reanalyze the Breast Cosmesis Study data set.9,25 Two medical treatments were given to 94 early breast cancer patients after tumorectomy: one was a mix of primary radiotherapy and adjuvant chemotherapy, and another was radiotherapy only. The interest of this clinical trial was to investigate which treatment has better long-term cosmetic effects. Interval-censoring results since the patients could only be visited every 4 to 6 months and those living farther away from the clinic had even longer follow-up intervals. Applying the ABB data augmentation with $M = 10$ imputations, the logrank, the Wilcoxon and the $G^{1/2}$ tests respectively have p-values 0.0014, 0.013 and 0.004. It only took several seconds of CPU time. These results are consistent with our simulation studies by considering that the main difference between two survival curves happens at later times (Figure 2) and hence the logrank test is the most sensitive in detecting it. In addition, our p-values from the logrank and Wilcoxon tests are a bit smaller than those from previous studies.9,11 Note that our extension of the logrank test or Wilcoxon test may be different from Finkelstein’s or Fay’s. Further studies are needed for a comparison. In addition, some differences in implementation, such as the algorithm used in calculating the NPMLE or its convergence criteria, may also contribute to the different p-values. However, the p-value difference seems minor and the conclusion is the same: there is a significant long-term benefit with radiation-alone therapy.
6. DISCUSSION

We have proposed a general method of comparing two interval-censored samples based on multiple imputation. The basic idea is simple: first to impute from interval-censored but not right-censored observations and then apply an existing two-sample comparison procedure to the imputed right-censored data. We have examined the application of the ABB multiple imputation scheme to interval censored data with a class of linear rank tests, the $G^b$ tests. In summary, our simulation results are similar to those of Fleming et al.$^{24}$ for right-censored data. In particular, the logrank test is more powerful in detecting the PHM and late differences. As pointed out by Pepe and Fleming,$^8$ since long-term differences in two survival functions, or long-term benefits of some clinical treatments, are often of more interest, in this sense the logrank test is more important in biomedical studies.

We proposed to impute from only interval-censored observations. There are two reasons why we did not impute from right-censored observations. First, after imputing from interval-censored observations, we have the imputed right-censored data and can thus use two-sample comparison techniques for right-censored data. There is no practical need to impute from right-censored observations (to obtain imputed complete data). The second reason is that often there is no enough information to impute from right-censored observations. For instance, due to the end of study, many observations may be right-censored at the last follow-up time, and essentially we have no information to impute exact survival times from them.

Our proposed method is attractive for its simplicity and generality. With the available SPLUS software for computing the NPMLE and comparing two right-censored samples, it only needs less than 100 lines of SPLUS code to implement our proposed method. (All sample programs used in this paper are available from the author upon request or from http://www.biostat.umn.edu/~weip.) In addition, there is no reason why we can not apply other existing two-sample comparison procedures for right censored data, instead of the $G^b$ tests, to implement our method. Particularly, it will be interesting to investigate the performance of the method combining the ABB multiple imputation and the WKM tests when applied to interval censored data. It is also possible to extend our method to compare more than two samples.$^{21}$
ACKNOWLEDGEMENTS

The author would like to thank Zhengqing Li for helpful discussions. The author is very grateful to two referees for insightful comments and suggestions that dramatically improved this paper.

REFERENCES


Table 1: Size and power of two-sided tests at nominal level 0.05 (0.01) for $H_0 : S_1 = S_2$ vs $H_1 : S_1 \neq S_2$ using ABB data augmentation. Sample size is 50 + 50 with 1000 replications.

<table>
<thead>
<tr>
<th>$S_1, S_2$</th>
<th>$M = 5$</th>
<th>$M = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\rho = 0$</td>
<td>1</td>
</tr>
<tr>
<td>$S_1 = S_2 = W(0.5)$</td>
<td>.053</td>
<td>.063</td>
</tr>
<tr>
<td></td>
<td>(.009)</td>
<td>(.017)</td>
</tr>
<tr>
<td>$S_1 = S_2 = W(1)$</td>
<td>.048</td>
<td>.050</td>
</tr>
<tr>
<td></td>
<td>(.011)</td>
<td>(.010)</td>
</tr>
<tr>
<td>$S_1 = S_2 = W(2)$</td>
<td>.057</td>
<td>.041</td>
</tr>
<tr>
<td></td>
<td>(.014)</td>
<td>(.006)</td>
</tr>
<tr>
<td>$S_1 = W(0.5), \lambda_2 = 1.73\lambda_1$</td>
<td>.398</td>
<td>.246</td>
</tr>
<tr>
<td></td>
<td>(.190)</td>
<td>(.108)</td>
</tr>
<tr>
<td>$S_1 = W(1), \lambda_2 = 2\lambda_1$</td>
<td>.714</td>
<td>.454</td>
</tr>
<tr>
<td></td>
<td>(.458)</td>
<td>(.262)</td>
</tr>
<tr>
<td>$S_1 = W(2), \lambda_2 = 2.25\lambda_1$</td>
<td>.830</td>
<td>.606</td>
</tr>
<tr>
<td></td>
<td>(.635)</td>
<td>(.359)</td>
</tr>
<tr>
<td></td>
<td>(.093)</td>
<td>(.084)</td>
</tr>
<tr>
<td>Late difference:</td>
<td>.330</td>
<td>.128</td>
</tr>
<tr>
<td></td>
<td>(.133)</td>
<td>(.051)</td>
</tr>
<tr>
<td>Crossing hazard a):</td>
<td>.164</td>
<td>.276</td>
</tr>
<tr>
<td></td>
<td>(.066)</td>
<td>(.112)</td>
</tr>
<tr>
<td>Crossing hazard b):</td>
<td>.109</td>
<td>.137</td>
</tr>
<tr>
<td></td>
<td>(.037)</td>
<td>(.069)</td>
</tr>
</tbody>
</table>
Figure 1: Configurations of the two survival curves to be tested in the simulation studies.
Figure 2: NPMLEs of the survival (defined as being cosmetically positive) probabilities for each treatment group from the Breast Cosmesis Study.