Published in final edited form as:
Comput Stat Data Anal. 2015 June 1; 86: 42–51. doi:10.1016/j.csda.2015.01.001.

Multiple comparisons for survival data with propensity score adjustment

Hong Zhu[a,*] and **Bo Lu**[b]

[a]Division of Biostatistics, Department of Clinical Sciences, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX, 75390, USA

[b]Division of Biostatistics, College of Public Health, The Ohio State University, 1841 Neil Avenue, Columbus, OH, 43210, USA

Abstract

This article considers the practical problem in clinical and observational studies where multiple treatment or prognostic groups are compared and the observed survival data are subject to right censoring. Two possible formulations of multiple comparisons are suggested. Multiple Comparisons with a Control (MCC) compare every other group to a control group with respect to survival outcomes, for determining which groups are associated with lower risk than the control. Multiple Comparisons with the Best (MCB) compare each group to the truly minimum risk group and identify the groups that are either with the minimum risk or the practically minimum risk. To make a causal statement, potential confounding effects need to be adjusted in the comparisons. Propensity score based adjustment is popular in causal inference and can effectively reduce the confounding bias. Based on a propensity-score-stratified Cox proportional hazards model, the approaches of MCC test and MCB simultaneous confidence intervals for general linear models with normal error outcome are extended to survival outcome. This paper specifies the assumptions for causal inference on survival outcomes within a potential outcome framework, develops testing procedures for multiple comparisons and provides simultaneous confidence intervals. The proposed methods are applied to two real data sets from cancer studies for illustration, and a simulation study is also presented.

Keywords

Causal inference; Multiple comparisons; Propensity score stratification; Simultaneous confidence intervals

*Corresponding author: hong.zhu@utsouthwestern.edu (Hong Zhu).

1. Introduction

Many studies in health and social sciences have time to some event (such as cancer onset or marriage dissolution) as their endpoints, and often in such studies researchers are interested in comparing multiple treatment or prognostic groups in terms of their survival outcomes. In many situations, not all pairwise comparisons are necessary. For example, patients can be separated by polymorphism of a gene into different groups and if one genotype is considered as normal (control), then the primary focus is on identifying those genotypes whose risks exceed that of the control, which is referred to as Multiple Comparisons with a Control (MCC). Another case is when the comparisons of particular interest are comparisons with the unknown best treatment, named as Multiple Comparisons with the Best (MCB). Suppose among six treatments, two are much inferior than the other four. Then it is not of interest which of those two is worse, and the conclusion that neither is the best suffices. In addition, if the second best treatment is almost as good as the true best treatment, it is useful to identify both as practically the best, because there might be other consideration such as cost and safety impacting on the choice of treatment. Specifically, MCB compare each group with the best of the other groups.

Various multiple comparison procedures were proposed for several types of endpoints in the literature. In particular, Dunnett (1955) proposed one-sided normal means method for MCC, and MCB has been developed for and applied in general linear model settings to provide simultaneous confidence intervals for the difference between each group and the best of the others (Hsu, 1996; Hsu *et al.*, 2006). However, comparing several treatment or prognostic groups in terms of their survival outcomes has not yet received much attention. Chen (2000) proposed procedures of MCC for survival data based on the log-rank tests. Logan *et al.* (2005) considered the general setting of all pairwise comparisons of survival data, accounting for correlation among the log-rank tests. Coolen-Maturi *et al.* (2012) introduced nonparametric predictive inference (NPI) for comparison of multiple groups for right-censored data, which uses lower and upper probabilities for the event that a specific group will provide the largest next lifetime. For survival outcome, a common measure of covariate effect is hazard ratio comparing a certain treatment or prognostic group with a reference group, estimated from the Cox proportional hazards model. It is desirable to develop effective methods of MCC and MCB for survival data under right censoring based on the Cox model, taking multiplicity of comparison groups into account.

Also, non-randomized clinical trials and observational studies are quite common in health and social sciences, where randomized allocation of treatment is not feasible or ethical. In the presence of confounders, the Cox model incorporating them as covariates is frequently employed and typically allows adjustment for bias, but in some cases, the proportionality assumption may be invalid. Further, one needs to determine whether the model is linear in confounders and, if not, what transformations are suggested by the data and clinical consideration. With respect to the validity of the statistical inferences, the proportional hazards assumption is crucial. Alternatively, statistical adjustment based on potential outcome framework are popular for evaluating causal relationship (Rubin, 1974). When the treatment assignment is strongly ignorable, propensity score based methods were shown to yield unbiased results (Rosenbaum and Rubin, 1983). In survival analysis, propensity score

matching or stratification have been proposed (Cupples *et al.*, 1995; Nieto and Coresh, 1996), which are considerably more flexible and robust than the Cox model-based regression adjustment. But, little work has been done to conduct multiple comparisons to elucidate causal effect with survival outcomes, which is very frequently an objective of a clinical or observational study.

In this paper, we extend the multiple comparison procedures for normal error outcome in general linear models to survival data, and compare groups in terms of log hazard ratios obtained from a propensity-score-stratified Cox model, assuming proportional hazards for group variable. In Section 2, we introduce the notation and discuss the assumptions for causal inference with survival outcomes, in the presence of multiple treatment groups. In Section 3, we propose two methods of multiple comparisons under the Cox model. Real data examples to illustrate the implementation of our procedures are given in Section 4. Section 5 provides results from a simulation study. Section 6 concludes the paper with further discussions.

2. Notations and Assumptions for Causal Inference with Survival Data

In this section, we introduce the notations and the potential outcome framework for survival data, and discuss the assumptions for making causal inference via the propensity score adjustment.

2.1. Notations

Suppose we are interested in evaluating the intervention effects among K treatment groups and the outcome of interest is time-to-event data with right censoring. Let (T, δ, R, $\mathbf{Z}$) denote the observed vector of right-censored data with multiple groups, where T is the possibly right-censored event time, δ is the censoring indicator where $\delta = 1$ if T corresponds to an event and $\delta = 0$ if T is censored, R is the index for the group membership where $R = 1, \ldots, K$ for K different groups, and $\mathbf{Z}$ is a $p \times 1$ covariate vector.

To identify the causal effect, we extend the potential outcome framework to survival data. The potential outcome framework was formally established by Rubin (1974) for dichotomous treatment comparison. With K groups, the potential event times are (S^1, S^2, ⋯, S^K) and the observed event time is represented as, if no censoring,

$$S = 1_{\{R=1\}} S^1 + 1_{\{R=2\}} S^2 + \cdots + 1_{\{R=K\}} S^K,$$

where $\sum_{k=1}^{K} 1_{\{R=k\}} = 1$ and $1_{\{R=k\}}$ is the indicator that patients belong to group k. Similarly, the potential censoring times are (C^1, C^2, ⋯, C^K) and the observed censoring time is

$$C = 1_{\{R=1\}} C^1 + 1_{\{R=2\}} C^2 + \cdots + 1_{\{R=K\}} C^K.$$

Therefore, the actually observed right-censored event time is $T = \min(S, C)$.

2.2. Assumptions for propensity score adjustment

Propensity score adjustment is widely used in making causal inference with observational data. Rosenbaum and Rubin (1983) provided the foundations of the propensity score theory for binary-valued groups ($K = 2$), where the propensity score is defined as the conditional probability of being in group $R = 1$ given a set of observed covariates,

$$e(\boldsymbol{Z})=P(R=1|\boldsymbol{Z}) \text{ where } 0<e(\boldsymbol{Z})<1.$$

Under strong ignorability assumption, they showed that $R \perp \mathbf{Z} \mid e(\mathbf{Z})$, so individuals from each group with the same propensity score are balanced in that the distributions of $\mathbf{Z}$ are the same regardless of group membership. Later, other researchers have extended the propensity score method to multilevel groups (Joffe and Rosenbaum, 1999; Imbens, 2000; Lu *et al.*, 2001; Imai and van Dyk, 2004).

To facilitate the causal comparison with multiple treatment groups in the presence of censoring, we identify the following assumptions based on the potential outcome framework.

1. Ignorable treatment assignment for event time:

   Condition on a set of observed pretreatment covariates, $\mathbf{Z}$, the treatment assignment is independent of the vector of potential event times.

   $$(S^1, S^2, \cdots, S^K) \perp R|\boldsymbol{Z}.$$

2. Conditional independent censoring given covariates:

   Condition on a set of observed pretreatment covariates, $\mathbf{Z}$, the potential censoring time is independent of the potential event time for any treatment group.

   $$S^1 \perp C^1, S^2 \perp C^2, \cdots, S^K \perp C^K|\boldsymbol{Z}.$$

3. Stable unit-treatment event time:

   There is a unique potential event time for each unit and treatment group. This is very similar to the original stable unit-treatment value assumption in Rubin (1980) and it ensures that different units do not interfere with each other's outcomes.

Given the first assumption above, it is easy to show the following proposition regarding the ignorability condition on the propensity score.

Proposition 1—If treatment assignment is strongly ignorable given $\boldsymbol{Z}$, then it is strongly ignorable given propensity score $\tilde{e}(\boldsymbol{Z})$; that is

$$(S^1, S^2, \cdots, S^K) \perp R|\boldsymbol{Z}$$

and

$$0<P(R=k|\boldsymbol{Z})<1,$$

for all k imply

$$(S^1, S^2, \cdots, S^K) \perp R|\tilde{e}(\boldsymbol{Z}),$$

where $\tilde{e}(\boldsymbol{Z})$ is a vector of the conditional probability for each treatment group.

The proof follows immediately from Theorem 3 in Rosenbaum and Rubin (1983) and is omitted here.

Propensity score adjustment can be implemented in several ways, including matching, stratification, weighting, or as regression covariates. Rosenbaum and Rubin (1983) advocated the use of propensity score quintiles for stratification with number of strata $J = 5$, which is a choice made in most published applications. This method can effectively reduce the covariate imbalance among different comparison groups within each stratum. The validity of the stratification procedure requires that the propensity model is correctly specified. In practice, the group variable is either ordinal or categorical. For a *K*-level ordinal group variable, a scalar propensity score is available using an ordinal logistic regression model

$$\log\left\{\frac{P(R\leq k)}{1-P(R\leq k)}\right\}=\gamma_k+\alpha'\boldsymbol{Z},$$

for $k = 1, \ldots, K - 1$. In this case, propensity score is determined by the scalar $\alpha'\mathbf{Z}$, and the major advantage is that we can balance the high-dimensional covariates by adjusting for a scalar propensity score. When the group variable is unordered, such as *K*-level categorical, we model the propensity function through a multinomial logistic regression model and obtain a set of $K - 1$ propensity scores. Since a set of propensity scores is estimated for each subject, there is no standard rule for stratification. If the number of levels is small, say $K = 3$, each subject have two estimated propensity scores and we could consider 2×2 or 3×3 stratification, depending on the sample size. An alternative approach for a categorical group variable with a large value of *K* is to use propensity score regression adjustment, where the estimate propensity score is included as a covariate in the regression model of outcome. Nevertheless, we illustrate the propensity score adjustment with stratification in the applications.

3. Methods of Multiple Comparisons under the Cox Model

To study the causal effect with survival outcome of interest, a propensity-score-stratified Cox proportional hazards model is then used. For convenience of discussion, we set group *K*

to be the reference (control) group, and create $K-1$ dummy variables ($X_1, \ldots, X_{K-1}$) to denote the group membership, where $X_k = 1$ if a subject is in group k, otherwise $X_k = 0$, for $k = 1, \ldots, K-1$. Note that $\{R = k\}$ is equivalent to $\{X_k = 1, X_{others} = 0\}$ for $k = 1, \ldots, K-1$, and $\{R = K\}$ is equivalent to $\{X_1 = 0, \ldots, X_{K-1} = 0\}$. We consider a stratified Cox model

$$h_j(t|X_1,\ldots,X_{K-1}) = h_{0j}(t)\exp(\beta_1 X_1 + \ldots + \beta_{K-1}X_{K-1}),\ j=1,\ldots,J, \quad (1)$$

where $h_{0j}(t)$ is the arbitrary baseline hazard function for stratum j, and β_k, $k = 1, \ldots, K-1$, is the logarithm of hazard ratio comparing group k to the reference group K in the stratified model, adjusting for confounders. The parameterization sets $\beta_K = 0$. In the following, we discuss multiple comparison procedures for survival outcome based on the estimates in model (1).

3.1. Multiple comparisons with a control for the Cox model

If one group is considered as the normal control group, group K say, and of primary interest is which groups are associated with lower risk than the control group K, then an analogue of one-sided Dunnett's normal means method for MCC (Dunnett, 1955) under our setting is

$$\beta_k < \hat{\beta}_k + d\hat{\sigma}_k,\ \text{for}\ k=1,\ldots,K-1,$$

where the log hazard ratio β_k measures the difference between group k and the control group K, $\hat{\sigma}_k^2$ is the estimated variance of $\hat{\beta_k}$ for model (1), and the multiplicity-adjusted critical value d is the upper α quantile of the maximum of $K-1$ random variables from a multivariate normal distribution with means zero and correlation matrix Σ equals the correlation matrix of ($\hat{\beta_1}, \ldots, \hat{\beta_{K-1}}$). The maximum partial likelihood estimate ($\hat{\beta_1}, \ldots, \hat{\beta_{K-1}}$) for coefficient vector asymptotically follows a multivariate normal distribution with a variance-covariance matrix which can be consistently estimated. Therefore, following the discussion in Hsu (1992), if the correlation matrix σ has a one-factor structure, that is, there exists constants $\lambda_1, \ldots, \lambda_{K-1}$ with all $|\lambda_i| < 1$ such that the correlation between $\hat{\beta_i}$ and $\hat{\beta_j}$, $\rho_{ij} = \lambda_i\lambda_j$ for all $i \neq j$, or equivalently,

$$\sum = diag(1-\lambda_1^2,\ldots,1-\lambda_{K-1}^2) + (\lambda_1,,\ldots,\lambda_{K-1})'(\lambda_1,,\ldots,\lambda_{K-1}),$$

the critical value d can be computed exactly. Having a one-factor structure implies $\hat{\beta_1}, \ldots, \hat{\beta_{K-1}}$ are conditionally independent, and this conditional independence facilitates critical value computation. When the correlation matrix Σ does not have a one-factor structure, the factor-analytic approximation of Hsu (1992) can be used to deterministically approximate the critical value d. The idea is to use factor analysis algorithms in multivariate analysis to find the correlation matrix Σ_{fa} with a one-factor structure that most closely approximates the correlation matrix Σ, and use the approximate correlation matrix Σ_{fa} to compute d. The variance reduction technique of Hsu and Nelson (1998) can be used to efficiently approximate d by simulation. It has been implemented in statistical software R by the

method of Genz and Bretz (1999), which uses a variance-reduced Monte Carlo algorithm to compute multivariate normal probabilities for arbitrary Σ.

In estimation of causal treatment effect for continuous outcome, Lunce-ford and Davidian (2004) developed asymptotic variance estimation of treatment effect estimator based on propensity score stratification, and it has a rather complicated form. While they mentioned that such theory is not used in practice, and it's routine to approximate the sampling variance by ignoring the additional variance from the estimated propensity score. We implement a similar strategy for our setting with survival outcome. In our analysis, $\hat{\sigma}_k^2$ as well as the estimated asymptotic variance-covariance matrix are approximated by estimates from "estimated-propensity-score-stratified" Cox model. In simulations, we compare empirical variance estimate with bootstrap variance estimate to assess the performance of the inference procedure.

3.2. Multiple comparisons with the best for the Cox model

We then consider MCB problem of comparing with the unknown best groups and determining whether each individual group has a lower risk than the best of the others. The purpose is to generalize the procedure of MCB for general linear model (Hsu, 1996) and to derive an asymptotical valid MCB method for comparing survival among groups in terms of log hazard ratio. The parameters of primary interest are

$$\{\beta_k - \min_{l\neq k}\beta_l, k=1,\dots K\},\ \text{or equivalently},\ \{\max_{l\neq k}(\beta_k-\beta_l), k=1,\dots K\},$$

where $\beta_K = 0$. If $\beta_k - \min_{l\neq k}\beta_l < 0$, then group k is the best, for it is better than every other group in terms of a lower hazard. If $\beta_k - \min_{l\neq k}\beta_l > 0$, group k is not the best. Further, suppose $\beta_k - \min_{l\neq k}\beta_l < \varepsilon$ where ε is a small positive number, then even if we cannot say group k is the best, we are sure that it is at least close to the best and has practically minimum risk. On the other hand, if the lower confidence bound for $\beta_k - \min_{l\neq k}\beta_l$ equals zero then we can infer group k is not the best. We provide simultaneous confidence intervals for the set of parameters of interest, by which the familywise error rate is strongly controlled.

Consider K tests for

$$H_{0k}:\max_{l\neq k}(\beta_k-\beta_l)<0\ v.s.\ H_{ak}:\max_{l\neq k}(\beta_k-\beta_l)\geq 0,\ k=1,\dots,K.$$

MCB simultaneous confidence intervals for $\{\max_{l\neq k}(\beta_k - \beta_l), k = 1, \ldots, K\}$ are derived as follows. For each k, $k = 1, \ldots, K$, suppose d^k is a constant value such that

$$P(\hat{\beta}_k-\hat{\beta}_l<\beta_k-\beta_l+d^k\hat{\sigma}_{kl},\ \text{for all}\, l,\ l\neq k)=1-\alpha, \quad (2)$$

where $\hat{\sigma}_{kl}^2$ is the estimated variance of $\hat{\beta_k} - \hat{\beta_l}$, and this implies the constant d^k is the one-sided MCC asymptotic critical value with group k as the control. Then for that k,

$$\hat{\beta}_k - \hat{\beta}_l - d^k \hat{\sigma}_{kl}, \text{ for all } l, l \neq k,$$

form simultaneous 100(1 − α)% lower confidence bounds for $\beta_k - \beta_l$ for all l, $l \neq k$. Thus, $\max_{l \neq k}(\hat{\beta_k} - \hat{\beta_l} - d^k \hat{\sigma_{kl}})$ is a 100(1 − α)% lower confidence bound for $\max_{l \neq k}(\beta_k - \beta_l)$. The parameter $\max_{l \neq k}(\beta_k - \beta_l)$, $k = 1, \ldots, K$, is negative if k is the unknown index of the best group, otherwise, it is positive. Define $D_k^- = -\{\max_{l \neq k}(\hat{\beta}_k - \hat{\beta}_l - d^k \hat{\sigma}_{kl})\}^-$ where $-x^- = \min\{x, 0\}$, then D_k^-, $k = 1, \ldots, K$, are simultaneous 100(1 − α)% lower confidence bounds for $\{\max_{l \neq k}(\beta_k - \beta_l), k = 1, \ldots, K\}$.

Further, for each k, we conclude $\max_{l \neq k}(\beta_k - \beta_l) < 0$, or equivalently, accept the null hypothesis H_{0k} when $D_k^- < 0$. Thus, $G = \{k : D_k^- < 0\}$ is a 100(1 − α)% confidence set for the unknown index of the best group with the minimum risk. Lastly, if we define

$$D_k^+ = \begin{cases} 0 & \text{if } G = \{k\}, \\ \max_{l \in G, l \neq k}(\hat{\beta}_k - \hat{\beta}_l + d^l \hat{\sigma}_{kl}) & \text{otherwise.} \end{cases}$$

then, D_k^+, $k = 1, \ldots, K$, are simultaneous 100(1 − α)% upper confidence bounds for $\{\max_{l \neq k}(\beta_k - \beta_l), k = 1, \ldots, K\}$. The confidence bounds D_k^- and D_k^+, $k = 1, \ldots, K$, are derived from the same 100(1 − α)% probability event as in (2). Therefore, we have the following result, which can be proved rigorously along the lines of Theorem 7.3.1 of Hsu (1996).

Proposition 2—For all β's, as $n_k \to \infty$ for $k = 1, \ldots, K$,

$$P_\beta\{(\beta_k - \min_{l \neq k} \beta_l) \in [D_k^-, D_k^+], \ \textit{for} \ \ k = 1, \ldots, K\} \geq 1 - \alpha,$$

where n_k is the number of subjects in group k.

Proof: Let b denote the unknown index such that $\beta_b = \min_{1 \leq k \leq K} \beta_k$. Define the event E as follow:

$$E = \{\hat{\beta}_b - \beta_b < \hat{\beta}_k - \beta_k + d^b \hat{\sigma}_{kb} \text{ for all } k, k \neq b\}.$$

By the definition of the critical value d^b, $P(E) = 1 - \alpha$. First of all, we derive the lower confidence bounds for $\{\beta_k - \min_{l \neq k} \beta_l, k = 1, \ldots, K\}$.

$$
\begin{aligned}
E&=\{\hat{\beta}_b-\beta_b<\hat{\beta}_k-\beta_k+d^b\hat{\sigma}_{kb}\text{ for all }k,k\neq b\}\\
&=\{\hat{\beta}_b-\hat{\beta}_k-d^b\hat{\sigma}_{kb}<\beta_b-\beta_k\text{ for all }k,k\neq b\}\\
&\subseteq\{\hat{\beta}_b-\hat{\beta}_k-d^b\hat{\sigma}_{kb}<\beta_b-\min_{l\neq b}\beta_l\text{ for all }k,k\neq b\}\\
&=\{\max_{l\neq b}(\hat{\beta}_b-\hat{\beta}_l-d^b\hat{\sigma}_{lb})<\beta_b-\min_{l\neq b}\beta_l\}\\
&=\{\max_{l\neq b}(\hat{\beta}_b-\hat{\beta}_l-d^b\hat{\sigma}_{lb})<\beta_b-\min_{l\neq b}\beta_l\text{ and }\beta_k-\min_{l\neq b}\beta_l\geq 0\text{ for all }k,k\neq b\}\\
&\subseteq\{-\{\max_{l\neq k}(\hat{\beta}_k-\hat{\beta}_l-d^k\hat{\sigma}_{kl})\}^-\leq\beta_k-\min_{l\neq k}\beta_l\text{ for all }k\}\\
&=\{D_k^-\leq\beta_k-\min_{l\neq k}\beta_l\text{ for all }k\}=E_1.
\end{aligned}
$$

We then derive the upper confidence bounds for $\{\beta_k-\min_{l\neq k}\beta_l,\ k=1,\ldots,K\}$.

$$
\begin{aligned}
E&=\{b\in G\text{ and }\hat{\beta}_b-\beta_b<\hat{\beta}_k-\beta_k+d^b\hat{\sigma}_{kb}\text{ for all }k,k\neq b\}\\
&=\{b\in G\text{ and }\beta_k-\beta_b<\hat{\beta}_k-\hat{\beta}_b+d^b\hat{\sigma}_{kb}\text{ for all }k,k\neq b\}\\
&=\{b\in G\text{ and }\beta_k-\min_{l\neq k}\beta_l<\hat{\beta}_k-\hat{\beta}_b+d^b\hat{\sigma}_{kb}\text{ for all }k,k\neq b\}\\
&\subseteq\{b\in G\text{ and }\beta_k-\min_{l\neq k}\beta_l<\max_{l\in G,l\neq k}(\hat{\beta}_k-\hat{\beta}_l+d^l\hat{\sigma}_{kl}\text{ for all }k,k\neq b\}\\
&=\{b\in G\text{ and }\beta_k-\min_{l\neq k}\beta_l<\max_{l\in G,l\neq k}(\hat{\beta}_k-\hat{\beta}_l+d^l\hat{\sigma}_{kl})\text{ for all }k,k\neq b\text{ and }\beta_k-\min_{l\neq k}\beta_l\leq 0\text{ for }k=b\}\\
&\subseteq\{\beta_k-\min_{l\neq k}\beta_l\leq D_k^+\text{ for all }k\}=E_2.
\end{aligned}
$$

We have shown $E\subseteq E_1\cap E_2$. Therefore,

$$
P_\beta\{(\beta_k-\min_{l\neq k}\beta_l)\in[D_k^-,D_k^+]\text{ for all }k\}=P(E_1\cap E_2)\geq P(E)=1-\alpha.
$$

To implement the MCB procedure, it is then required to compute critical values $\{d_k,\ k=1,\ldots,K\}$. The techniques for computing MCC critical value discussed in Section 3.1 is used. Since $\hat{\beta_k},\ k=1,\ldots,K-1$ are the maximum partial likelihood estimates for model (1), for each k, we can use the fact that $\{\hat{\beta_k}-\hat{\beta_l},\ l\neq k\}$ has an asymptotic multivariate normal distribution, with a variance-covariance matrix that can be consistently estimated, to calculate critical value d^k. Similarly, d^k can be computed exactly when the correlation matrix of $\{\hat{\beta_k}-\hat{\beta_l},\ l\neq k\}$ has a one-factor structure. Otherwise, factor-analytic approximation can be used to deterministically approximate d^k.

4. Applications

4.1. Bone marrow transplantation for leukemia

Bone marrow transplantation is a standard treatment for acute leukemia, and recovery following bone marrow transplant is a complex process. Prognosis for recovery may depends on risk factors known at the time of transplantation, such as patient and/or donor age and gender, the stage of initial disease, the time from diagnosis to transplantation, etc. A study was conducted to illustrate this process and characterize disease-free survival for bone marrow transplantation patients of acute leukemia. Details of the study are found in Copelan

et al. (1991). In this multi-center clinical trial, a total of 137 patients were treated with bone marrow transplants at one of four hospitals. They were followed for relapse or death, with the maximum follow-up of 7 years. There were 42 patients who relapsed and 41 who died while in remission. The censoring percentage was around 40%. In addition to time to relapse or death, several potential risk factors were measured at the time of transplantation. Patients were grouped into risk categories based on their initial disease status at the time of transplantation as follows: 38 acute lymphoblastic leukemia (ALL) patients, 54 acute myelotic leukemia (AML) low-risk patients, and 45 acute myelotic leukemia (AML) high-risk patients. Other factors measured include recipient and donor gender and age, waiting time from diagnosis to transplantation, the French-American-British (FAB) classification and graft-versus-host prophylactic combining methotrexate (MTX) status. The primary interest is in comparing the disease-free survival of AML low-risk patients with that of ALL patients (control), and comparing AML high-risk patients with ALL patients. Since it was not a randomized clinical trial, we shall adjust these comparisons to reduce the possible confounding bias. We check the proportional hazards assumption for each factor using scaled Schoenfeld residual plot and find the nonproportionality of MTX status as shown in Figure 1, which suggests a simple Cox model-based regression adjustment is inappropriate. Therefore, we execute the proposed MCC procedure based on a propensity-score-stratified Cox model.

To calculate the propensity scores, the probability of being in AML low-risk group and that of being in AML high-risk group conditioning on gender, age, waiting time, FAB and MTX are estimated by a multinomial logistic regression model with ALL patients as the reference group. For each patient, we obtain the estimated propensity score for AML low-risk (*pscore*1), and the estimated propensity score for AML high-risk (*pscore*2). The patients are then categorized into 2 × 2 strata based on a vector of estimated propensity scores, (*pscore*1, *pscore*2). The first stratum contains subjects with both *pscore*1 and *pscore*2 higher than their sample medians, the second stratum contains subjects with *pscore*1 higher than its median and *pscore*2 lower than its median, the third stratum contains subjects with *pscore*1 lower than its median and *pscore*2 higher than its median, and the last stratum contains subjects with both *pscore*1 and *pscore*2 lower than their medians. The distributions of the potential confounders are approximately balanced among the three risk groups within each stratum. In the corresponding propensity-score-stratified Cox model, covariate X_1 denotes the indicator of AML low-risk and X_2 denotes the indicator of AML high-risk. A patient has ALL if $X_1 = 0$ and $X_2 = 0$. The point estimates of log hazard ratios are $\hat{\beta_1} = -0.997$ for comparing AML low-risk with ALL, and $\hat{\beta_2} = -0.186$ for comparing AML high-risk with ALL. For $\alpha = 0.05$, with estimated variance-covariance matrix of $(\hat{\beta_1}, \hat{\beta_2})$,

$$\hat{cov} = \begin{pmatrix} 0.132 & 0.100 \\ 0.100 & 0.146 \end{pmatrix},$$

the one-sided multiplicity-adjusted critical value $d = 1.872$, based on the factor-analytic approximation. We have $\hat{\beta_1} + d\hat{\sigma_1} = -0.750 < 0$ and $\hat{\beta_2} + d\hat{\sigma_2} = 0.087 > 0$. Therefore, adjusting for potential confounders, at the 95% confidence level, we conclude that AML

low-risk patients have a significant lower risk of leukemia relapse or death compared with ALL patients, while AML high-risk patients have a non-significant lower risk of leukemia relapse or death compared with ALL patients.

4.2. Death times of male laryngeal cancer patients

The proposed MCB procedure is applied to survival data of male laryngeal cancer patients. A study of 90 males diagnosed with cancer of the larynx during the period 1970–1978 at a Dutch hospital is reported (Kardaun, 1983). Times (in years) between first treatment and either death or the end of the study (January 1, 1983) are recorded. Also recorded are patient's age at diagnosis, the year of diagnosis, and the stage of the patient's cancer. The four stages of disease in the study were based on the T.N.M. classification used by the American Joint Committee for Cancer Staging. The stages have 4 levels, and are ordered from least serious (Stage I) to most serious (Stage IV). We focus on identifying stages with the minimum risk of death (the best), and the multiple comparisons need to be adjusted for potential confounders. The proportionality assumption for cancer stage is not violated. However, since the effect of age at diagnosis tends not to be linear based on a martingale residual plot in Figure 2, the confounding adjustment is implemented by propensity score stratification. Cancer stage is an ordinal variable, so a scalar propensity score can be obtained by an ordinal logistic regression model with age at diagnosis and the year of diagnosis as covariates, and is used for stratification. We then categorize the patients into 5 strata by the quintiles of estimated propensity scores. The log hazard ratios comparing survival among different disease stages are estimated from a propensity-score-stratified Cox model, where covariate X_1 denotes the indicator of Stage I, X_2 denotes the indicator of Stage II, X_3 denotes the indicator of Stage III, and Stage IV is the reference group with $X_1 = X_2 = X_3 = 0$. Our parameterization sets $\beta_4 = 0$, and $\hat{\beta_k}$ estimates $\beta_k - \beta_4$ for $k = 1, 2, 3$.

The point estimates of log hazard ratios are $\hat{\beta_1} = -1.656$ for Stage I against Stage IV, $\hat{\beta_2} = -1.759$ for Stage II against Stage IV, and $\hat{\beta_3} = -1.024$ for Stage III against Stage IV. The estimated variance-covariance matrix of $(\hat{\beta_1}, \hat{\beta_2}, \hat{\beta_3})$ is

$$\hat{cov} = \begin{pmatrix} 0.212 & 0.138 & 0.139 \\ 0.138 & 0.282 & 0.132 \\ 0.139 & 0.132 & 0.195 \end{pmatrix}.$$

For $\alpha = 0.05$, based on the factor-analytic approximation, the critical values d^1, d^2, d^3 and d^4 are 2.365, 2.317, 2.371 and 2.324, respectively. To perform the proposed MCB procedure, we first calculate the lower confidence bounds D_1^-, D_2^-, D_3^- and D_4^- using the critical values. The lower confidence bounds turn out to be −1.002, −1.186, −0.217 and 0. Therefore, the confidence set G for the unknown minimum risk groups is {1, 2, 3}, which implies that Stage IV does not have the minimum risk of death and there exists other group with lower risk. Then we compute the upper confidence bounds D_k^+, $k = 1, \ldots, 4$, which are

$$D_k^+ = \max_{l\in\{1,2,3\},l\neq k}(\hat{\beta}_k - \hat{\beta}_l + d^l\hat{\sigma}_{kl}),$$

in this case. They turn out to be 1.186, 1.002, 1.802 and 2.991. Thus, at the 95% confidence level, the MCB simultaneous confidence intervals for Stage I, II III and IV minus the best of the others are:

$$\begin{gathered}\beta_1 - \min\{\beta_2,\beta_3,\beta_4\} \in [-1.002, 1.186],\\ \beta_2 - \min\{\beta_1,\beta_3,\beta_4\} \in [-1.186, 1.002],\\ \beta_3 - \min\{\beta_1,\beta_2,\beta_4\} \in [-0.217, 1.802] \text{ and}\\ \beta_4 - \min\{\beta_1,\beta_2,\beta_3\} \in [0, 2.991].\end{gathered}$$

So at the 95% confidence level, one can say Stage IV is not the stage with the minimum risk. Stage I is within 1.186 of the best, Stage II is within 1.002 of the best and Stage III is within 1.802 of the best, but it is insufficient to decide which of them gets the minimum risk. If a log hazard ratio no bigger than 1.200 compared to the theoretical lowest log hazard ratio is considered practically minimum risk, then Stage I and Stage II can be inferred as practically minimum risk groups.

5. Simulation

A simulation study is conducted to assess performance of the proposed MCB method of comparing survival among multiple groups adjusting for confounding effect, based on multiplicity-adjusted simultaneous confidence intervals. We first consider a case of $K = 4$ groups, under independent and covariate-dependent censoring, respectively. A sample is generated as follows: (i) the covariate Z (confounder) is generated from a Bernoulli distribution with $Z \sim$ Bernoulli(0.5); (ii) the treatment group variable R is generated given Z with probability $P\{R = (1, 2, 3, 4)\} = (1/4, 1/4-z/4, 1/4, 1/4+z/4)$; (iii) based on the simulated R, we create variables X_1, X_2 and X_3 which are indicators of group 1, group 2 and group 3; (iv) the survival times are generated from a Cox proportional hazards function

$$h(T=t|X_1=x_1, X_2=x_2, X_3=x_3, Z=z) = \exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + z),$$

where $(\beta_1, \beta_2, \beta_3) = (-1.5, -1, -0.5)$ for increasing risk of failure with higher group level and group 4 is the reference group; (v) In Scenario 1, the censoring times are independently generated from a uniform distribution with a censoring proportion of around 30%, and in Scenario 2, the censoring times depend on the confounder Z and are generated from an exponential distribution with a censoring proportion of around 35%. With $(\beta_1, \beta_2, \beta_3, \beta_4) = (-1.5, -1, -0.5, 0)$, the true values of the set of parameters for MCB $\{\beta_k - \min_{l\neq k}\beta_l, k = 1, \ldots 4\}$ are $(-0.5, 0.5, 1, 1.5)$. The simulation is repeated 10000 times with sample size $n =$ 250, 500 and 1000. Next, we consider $K = 6$ groups. The simulation settings keep the same except: (i) the group variable R is generated given Z with probability $P\{R = (1, 2, 3, 4, 5, 6)\} = (1/6, 1/6 - z/6, 1/6, 1/6 + z/6, 1/6, 1/6)$; (ii) based on the simulated R, we create variables X_1, X_2, X_3, X_4 and X_5 which are indicators of groups 1–5; (iii) the survival times are

generated from a Cox proportional hazards function $h(T = t \mid X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4, X_5 = x_5, Z = z) = \exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + z)$, where $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = (-1.25, -1, -0.75, -0.5, -0.25)$ and group 6 is the reference group.

For each simulated data set, we fit an ordinal logistic regression model of R on Z to estimate the propensity score, and use the propensity score stratification to adjust for the confounding effect from Z. The proposed MCB method is applied to derive simultaneous confidence intervals, using estimates from a propensity-score-stratified Cox model with group indicators as covariates. Table 1 summarizes the simulation results on estimation of regression coefficients $(\beta_1, \beta_2, \beta_3)$ for $K = 4$, including the empirical bias, empirical standard error and average bootstrap standard error. Under both independent and covariate-dependent censoring, the empirical biases are small and decrease as sample size increases. The bootstrap standard errors agree with the empirical standard errors, implying that the inference procedure performs reasonably well. Table 2 presents the simulation results of MCB method, which includes the estimated simultaneous coverage probabilities of 90% and 95% MCB confidence intervals, for different values of K and n, under independent and covariate-dependent censoring. For both censoring schemes, the estimated simultaneous coverage probabilities of MCB confidence intervals based on normal approximation are slightly higher than the nominal level. The coverage probabilities get closer to the nominal level for smaller K or larger n. The reason for the over-coverage might be related to normal approximation. For example, to compare group k and group l based on the asymptotic normality of $\hat{\beta}_k$ and $\hat{\beta}_l$, one refers a statistic

$$\frac{\hat{\beta}_k - \hat{\beta}_l - (\beta_k - \beta_l)}{\hat{\sigma}_{\hat{\beta}_k - \hat{\beta}_l}},$$

to the standard normal distribution or t distribution for all β's. However, there is a correlation between $\hat{\beta}_k - \hat{\beta}_l$ and $\hat{\sigma}_{\hat{\beta}_k - \hat{\beta}_l}$, which depends on β's, therefore, the statistic is not as pivotal as one would expect.

6. Discussion

This paper discusses multiplicity-adjusted inference for survival data that are subject to random right-censorship, in presence of confounders. Assumptions for making causal comparison among multiple groups are discussed within a potential outcome framework. A MCC testing procedure is described to determine which groups have lower risk than the control, and MCB simultaneous confidence intervals are provided to identify the groups that deliver the minimum risk or the maximum benefit. The testing formulation for multiple comparisons has been more popular, but confidence intervals are more informative because they not only infer the existence of the difference, but also bound the magnitude of the difference. The existing methods for multiple comparisons with survival outcomes are either based on log-rank tests (Chen, 2000; Logan *et al.*, 2005) or some nonparametric inference (Coolen-Maturi *et al.*, 2012), which may be inconvenient for covariate adjustment in many applications. On the contrary, the proposed methods use estimates from the Cox

proportional hazards model, which has the advantage of adjusting for confounders and/or covariates in a straightforward way.

Specifically, we use propensity score stratification to adjust for the confounding bias and log hazard ratios are obtained based on a stratified Cox model for multiple comparisons. The correlation structure of log hazard ratio estimates can then be derived, allowing to obtain accurate multiplicity-adjusted critical value. The appropriate choice of a model for estimating propensity score depends on the nature of the group variable, categorical or ordinal. The balanced confounder distributions are expected among the comparison groups after stratification, and we suggest some practical guidelines for implementing stratification. The multiple comparison procedures with confounding bias adjustment are illustrated with two real data sets, where we compare the disease-free survivals of AML low-risk, AML high risk groups with that of ALL group for bone marrow transplantation patients, and for male laryngeal cancer patients, we identify the disease stage that has the minimum risk or practically minimum risk of death. A simulation study is carried out to assess the finite sample performance of the proposed MCB method with multiplicity adjustment based on the normal approximation. The over-coverage probability of the MCB method remains an interesting research problem and will be explored in future work.

There are several limitations to the current work. First, the causal comparison is based on the strong ignorability assumption, which requires all confounding variables are observed. In practice, this might not be true for many observational studies and unmeasured confounders are often of concern. Some sensitivity analysis procedure may be developed to address how robust the conclusions are in the presence of unobserved confounding, following the idea of Rosenbaum (2002). Second, the Cox model with the proportional hazards assumption for group variable is employed after propensity score stratification. However, this assumption may also not be appropriate, and in such cases, the subsequent multiple comparison procedures are not reliable. It is interesting to investigate performances of testing and multiple comparison procedures based on the Cox model when the proportional hazards assumption is invalid, and to provide some alternative approach that does not rely on the proportional hazards assumption for multiple comparisons. Lastly, we consider multiple comparison problems for survival data under right censoring in this paper. In clinical and observational studies, survival data may be subject to various types of censoring and truncation, such as informative censoring due to competing risk. Therefore, further research is required to extend the multiple comparison procedures to these types of data under appropriate survival models.

Acknowledgments

The authors thank Dr. Jason Hsu in the Department of Statistics at the Ohio State University for insightful discussions. The authors are also grateful for constructive comments from the editor, the associate editor and three referees. The research is partially supported by R24HD058484 from the Eunice Kennedy Shriver National Institute for Child Health and Human Development awarded to The Ohio State University Initiative in Population Research. Hong Zhu is supported in part by the Cancer Center Support Grant from the National Cancer Institute (5P30CA142543) awarded to the Harold C. Simmons Cancer Center at the University of Texas Southwestern Medical Center. Bo Lu is partially supported by a grant from the National Institute on Drug Abuse (R03DA030662).

References

Chen YI. Multiple comparisons in carcinogenesis study with right-censored survival data. Statistics in Medicine. 2000; 19:353–367. [PubMed: 10649301]

Coolen-Maturi T, Coolen-Schrijner P, Coolen F. Nonparametric predictive multiple comparison of lifetime data. Communications in Statistics-Theory and Method. 2012; 41(22):4164–4181.

Copelan EA, Biggs JC, Thompson JM, Crilley P, Szer J, Klein JP, Kapoor N, Avalos BR, Cunningham I, Atkinson K, Downs K, Harmon GS, Daly MB, Brodsky I, Bulova SI, Tutschka PJ. Treatment for acute myelocytic leukemia with allogeneic bone marrow transplantation following preparation with Bu/Cy. Blood. 1991; 78:838–843. [PubMed: 1859895]

Cupples LA, Gagnon DR, Ramaswamy R, D'Agostino RB. Age-adjusted survival curves with application in the Framingham study. Statistics in Medicine. 1995; 14:1731–1744. [PubMed: 7481206]

Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association. 1955; 50:1096–1121.

Genz A, Bretz F. Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts. Journal of Statistical Computation and Simulation. 1999; 63:361–378.

Hsu JC. The factor analytic approach to simultaneous inference in the general linear model. Journal of Graphical and Computational Statistics. 1992; 1:151–168.

Hsu, JC. Multiple Comparisons: Theory and Methods. Chapman & Hall; London: 1996.

Hsu JC, Nelson BL. Multiple comparisons in the general linear model. Journal of Computational and Graphical Statistics. 1998; 7:23–41.

Hsu JC, Chang JY, Wang T. Simultaneous confidence intervals for differential gene expressions. Journal of Statistical Planning and Inference. 2006; 136:2182–2196.

Imai K, van Dyk DA. Causal inference with general treatment regimes: generalizing the propensity score. Journal of the American Statistical Association. 2004; 99:854–866.

Imbens GW. The role of the propensity score in estimating dose-response functions. Biometrika. 2000; 87:706–710.

Joffe MM, Rosenbaum PR. Propensity scores. American Journal of Epidemiology. 1999; 150:327–333. [PubMed: 10453808]

Kardaun O. Statistical analysis of male larynx-cancer patients - a case study. Statistica Neerlandica. 1983; 37:103–125.

Logan BR, Wang H, Zhang MJ. Pairwise multiple comparison adjustment in survival analysis. Statistics in Medicine. 2005; 24:2509–2523. [PubMed: 15977296]

Lu B, Zanutto E, Hornik R, Rosenbaum PR. Matching with doses in an observational study of a media campaign against drug abuse. Journal of the American Statistical Association. 2001; 96:1245–1253. [PubMed: 25525284]

Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in Medicine. 2004; 23:2937–2960. [PubMed: 15351954]

Nieto FJ, Coresh J. Adjusting survival curves for confounders: a review and a new method. American Journal of Epidemiology. 1996; 143:1068–1069.

Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70:41–55.

Rosenbaum, PR. Observational Studies. Springer; New York: 2002.

Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology. 1974; 66:688–701.

Rubin DB. Discussion of "Randomization analysis of experimental data in the Fisher randomization test," by D. Basu. Journal of the American Statistical Association. 1980; 75:591–593.

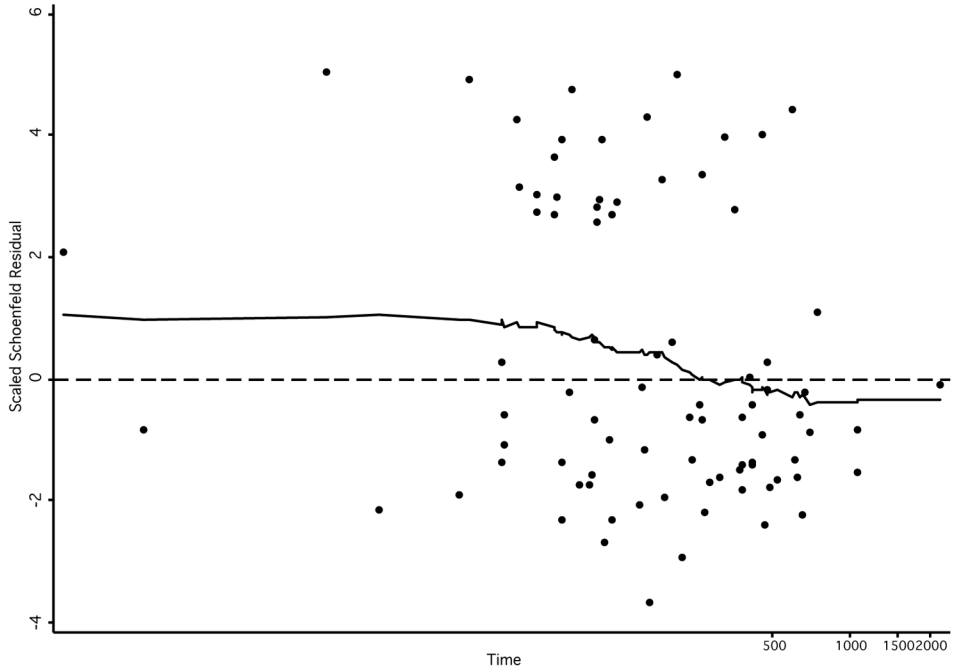

Figure 1.
Scaled Schoenfeld residual plot of MTX status

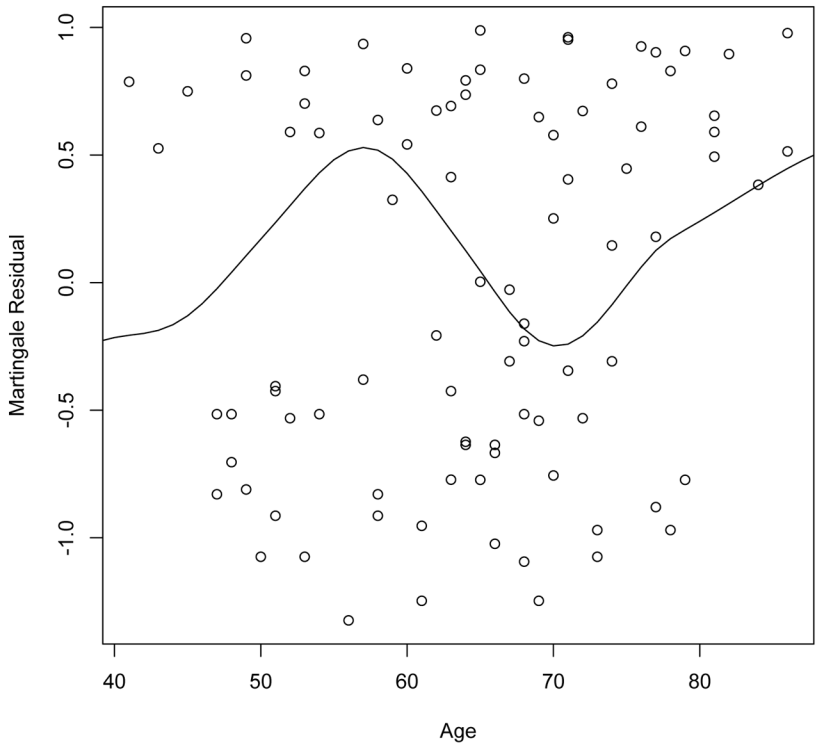

Figure 2.
Martingale residual plot of age at diagnosis

Table 1

Simulation results of estimation of (β_1, β_2, β_3) for $K = 4$.

n	b_1	SEE_1	SEB_1	b_2	SEE_2	SEB_2	b_3	SEE_3	SEB_3
Ind. Cen.									
250	0.026	0.168	0.171	−0.096	0.196	0.198	−0.021	0.170	0.174
500	0.004	0.081	0.085	−0.053	0.094	0.097	−0.015	0.081	0.084
1000	0.005	0.036	0.038	−0.047	0.042	0.045	−0.013	0.038	0.043
Dep. Cen.									
250	−0.073	0.148	0.152	0.022	0.138	0.143	−0.012	0.134	0.137
500	−0.054	0.070	0.076	0.013	0.066	0.073	0.001	0.065	0.071
1000	−0.046	0.034	0.040	0.009	0.033	0.038	0.041	0.032	0.037

For β_i ($i = 1, 2, 3$), b_i, empirical bias; SEE_i, empirical standard error; SEB_i, average bootstrap standard error. Ind. Cen., independent censoring; Dep. Cen., dependent censoring.

Table 2

Simulation results of MCB approach.

K	*n*	90%CP	95%CP
Ind. Cen.			
4	250	92.52	96.24
	500	92.15	96.17
	1000	91.93	96.06
6	250	92.79	96.48
	500	92.55	96.27
	1000	92.41	96.13
Dep. Cen.			
4	250	93.79	97.19
	500	93.49	97.17
	1000	93.25	97.08
6	250	93.95	97.58
	500	93.78	97.45
	1000	93.64	97.33

CP, estimated simultaneous coverage probabilities of MCB confidence intervals (×100); Ind. Cen., independent censoring; Dep. Cen., dependent censoring.