Published in final edited form as:

Stat Methods Med Res. 2016 February ; 25(1): 447–457. doi:10.1177/0962280212460434.

Comparing paired biomarkers in predicting quantitative health outcome subject to random censoring

Xinhua Liu,

Department of Biostatistics, Mailman School of Public Health, Columbia University, USA

Zhezhen Jin, and

Department of Biostatistics, Mailman School of Public Health, Columbia University, USA

Joseph H. Graziano

Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, USA

Abstract

This paper uses a non-parametric test, based on consistently estimated discrimination accuracy defined as concordance probability between quantitative predictor and outcome, to compare paired biomarkers in predicting a health outcome, possibly subject to random censoring. Comparing with the Wilcoxon test for paired predictors based on Harrell's C-index, we found that the proposed test is better in presence of random censoring, although the two unbiased tests are equivalent for outcome either uncensored or censored by a constant. A simulation study also demonstrates that the bias in estimated difference in concordance probability, due to ignoring random censoring, results in overestimated power, especially when random censoring is heavy. The method was applied in two studies, where the biomarkers measured from the same study subjects are correlated. The first study on 299 school children in Bangladesh found the associations that higher blood arsenic and manganese were related to lower intellectual test scores, while the differences between the biomarkers in predicting the intellectual test scores were not statistically significant. The second study on 418 patients with primary biliary cirrhosis found that the baseline serum bilirubin had greater discrimination accuracy than the baseline serum albumin in predicting survival time.

Keywords

C-index; Concordance probability; Discrimination accuracy; Paired predictors; Random censoring

1 Introduction

In biomedical studies, investigators are not only making effort to identify biomarkers that are associated with specific quantitative health outcome, but also seek to compare their

Address for correspondence: **Xinhua Liu**, Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 W 168th street, New York, NY 10032, USA. Phone: 212-342-1262, Fax: 212-305-9408, XL26@columbia.edu.

Conflict of Interest Statement

The authors declare that there is no conflict of interest.

discrimination accuracy in predicting the common outcome variable. Data on biomarkers and health outcomes are usually obtained from the same study subjects. In such cases, the measurements of the biomarkers are likely to be correlated. For example, in a study on the association between exposures to arsenic and manganese and children's intellectual function in Bangladesh, as described in Wasserman *et al.* (2011)¹, blood arsenic and manganese used as biomarkers of the exposures were measured using the blood samples from 299 school children who were individually administered an IQ test (The Wechsler Intelligence Scale for Children-Fourth Edition) to assess their intellectual function. The two biomarkers were correlated, with a Spearman correlation coefficient of 0.1256 ($p < 0.03$). It has been found that each of the biomarkers, blood arsenic and blood manganese, had higher levels related to the lower IQ test scores. The investigators then sought to determine whether one of the biomarkers has higher discrimination accuracy in predicting the intellectual function test scores. In studies concerning the health outcome of survival time, it is common that the outcome variable may be censored randomly due to dropout and limited follow up period.

In this paper, we compare paired biomarkers in discrimination accuracy of predicting a health outcome, where the biomarkers have continuous measurements and the quantitative outcome variable T may be subject to random censoring. To measure the discrimination accuracy of predictor X for T , Harrell's C-index^{2, 3}, defined as a concordance probability on independent pairs of observations (X_j, T_j) , $j = 1, 2$;

$$C = \frac{P(X_1 < X_2, T_1 < T_2)}{P(X_1 < X_2, T_1 < T_2) + P(X_1 > X_2, T_1 < T_2)}$$

is widely used. The index takes values between zero and one with 0.5 for X and T being independent and it is invariant to rank-preserving transformation on either X or T . Using counts of concordant and discordant pairs of observations to estimate C , Nam and D'Agostino⁴ developed a method to estimate the variance of the estimator. Pencina and D'Agostino⁵ used the relationship between the C-index and the modified Kendall's τ for bivariate correlation to derive alternative formulas for variance estimation. Based on the linear relationship between the C-index and Somers D rank correlation that $D = 2C - 1$, Softwares STATA and R have functions to estimate the index with bootstrap estimate of variance⁶. Note that when X is continuous, we have $C = C_X = P(X_1 < X_2 | T_1 < T_2)$, where the concordance probability C_X may also be used as an alternative measure of discrimination accuracy. To compare discrimination accuracy in paired biomarkers, a Wilcoxon test for the difference in C is available when a quantitative outcome is completely observed or subject to constant censoring⁷. In presence of random censoring, however, the estimator converges to a quantity depending on the censoring distribution, and is no longer consistent to C ^{8, 9}. We adapt the method that Liu and Jin (2009)⁹ proposed for consistent estimator of C_X and testing difference in C_X for item reduction, which selects items from a uni-dimensional scale (X) to form a subscale with similar or improved discrimination accuracy C_X in predicting the quantitative outcome variable subject to random censoring. The selection procedure is based on evaluation of the change in discrimination accuracy C_X resulting from excluding an item from, or adding an item to, a sub-scale.

In the next section, we describe the statistical test for the difference between paired biomarkers in discrimination accuracy for predicting a quantitative health outcome variable subject to random censoring. We present a simulation study to demonstrate finite sample performance of the consistent estimator of C_X and the statistical test for the difference between paired predictors in discrimination accuracy with and without random censoring on the response variable. The results are compared with that from using the function `rcorr.cens` in Hmisc package of R software⁷ which works well with response either uncensored or censored by a constant. Finally, we apply the method to the two studies. The first application compares biomarkers of blood arsenic and blood manganese in discrimination accuracy for predicting child's intelligence test scores. The second application uses the data in Fleming and Harrington¹⁰ to examine the difference between important prognostic factors of baseline serum albumin and bilirubin in discrimination accuracy for predicting survival time among the patients with primary biliary cirrhosis.

2 The procedure

Suppose that variables X and Z are paired continuous predictors of a quantitative response variable T and all these variables have been measured in n independent subjects with data (X_i, Z_i, T_i) , $i = 1, \dots, n$. Then the discrimination accuracy measure C_X and C_Z can be consistently estimated respectively by

$$\hat{C}_X = \frac{\sum_{i=1}^n \sum_{j=1}^n I(X_i < X_j) I(T_i < T_j)}{\sum_{i=1}^n \sum_{j=1}^n I(T_i < T_j)},$$

where $I(\cdot)$ is an indicator function taking values of 0 or 1; so for C_Z . To examine whether X and Z have the same discrimination accuracy for predicting T , we may examine the difference between their concordance probabilities $C = C_X - C_Z$ and test null hypothesis $H_0: C = 0$, or equivalently, $H_0: P = P(X_1 < X_2, T_1 < T_2) - P(Z_1 < Z_2, T_1 < T_2) = 0$. With uncensored response T , a commonly used non-parametric estimator of P is in the form of a U -statistic,

$$\Delta_{X-Z} = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1}^n \{I(X_i < X_j) - I(Z_i < Z_j)\} I(T_i < T_j) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1}^n U_{ij}. \quad (1)$$

As U_{ij} takes values of $-1, 0$, or 1 , the Wald type test statistic $TS_1 = \frac{\Delta_{X-Z}}{se(\Delta_{X-Z})}$ is then equivalent to Wilcoxon test, where

$$se^2(\Delta_{X-Z}) = \frac{2}{n(n-1)} \{ \text{Var}(U_{12}) + 2(n-1) \text{Cov}(U_{12}, U_{13}) \}.$$

Under $H_0: P = 0$, the test statistic has asymptotic normal distribution with zero mean and unit variance. To estimate $se(\Delta_{X-Z})$, we may replace the moment estimators for $\text{Var}(U_{12})$

and $\text{Cov}(U_{12}, U_{13})$ in the formula. In R software the function `rcorr.cens` provides estimated C_X and C_Z and a test statistic for their difference, while using bootstrap method for their standard errors^{6, 7}.

When the response variable is the time to an event, where the event could be death or initial diagnosis of a disease, then the time variable is likely to be right-censored due to dropout or limitation of follow up period. Let T_i be the length of time between the baseline assessment and event occurrence for subject i during follow-up. When subject i does not have the event at the last follow-up time Q_i , then censoring occurs and the observed time $Y_i = T_i d_i + Q_i (1-d_i)$ with $d_i = I(T_i < Q_i)$. Suppose the censoring variable Q is independent of T . If Q is constant, then the estimator of the C-index using usable pairs of observations,

$$\hat{C}_x = \frac{\sum_{i=1}^n \sum_{j=1}^n I(X_i < X_j) I(T_i < T_j) d_i}{\sum_{i=1}^n \sum_{j=1}^n I(T_i < T_j) d_i}$$

is consistent to C_X . Using it to estimate C_X and C_Z , and form the test to detect difference in discrimination accuracy, R function `rcorr.cens` will give valid result with constant Q .

However, when the censoring time Q is random, the estimator $\hat{C}_x$ converges to a quantity depending on the distribution of censoring time, as pointed out by Koziol and Ji⁸ and Liu and Jin⁹. Assuming that random censoring time Q is independent of predictors, a consistent estimator of C_X proposed by Liu and Jin⁹ has the form

$$\hat{C}_x^* = \frac{\sum_{i=1}^n \sum_{j=1}^n I(X_i < X_j) I(Y_i < Y_j) w_i}{\sum_{i=1}^n \sum_{j=1}^n I(Y_i < Y_j) w_i}$$

where $w_i = \frac{d_i}{G^2(Y_i)}$ with $G(t) = P(t < Q)$ for $t > 0$. It is easy to see that $\hat{C}_x^*$ becomes $\hat{C}_x$ when censoring variable Q is constant. To estimate P consistently, we may modify (1) to be

$$\Delta_{x-z}^* = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1}^n \{I(X_i < X_j) - I(Z_i < Z_j)\} I(Y_i < Y_j) w_i \quad (2)$$

If $G(t)$ is unknown, a consistent estimator $\hat{G}(t)$, constructed by the Kaplan-Meier product limit method, may be used. For $Y_{(n)} = \max_{1 \leq i \leq n} Y_i$, if $d_{(n)} = 0$ and $\hat{G}(Y_{(n)}) = 0$ then $w_{(n)} = 0$ is required. Thus (2) becomes

$$\hat{\Delta}_{x-z}^* = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1}^n \{I(X_i < X_j) - I(Z_i < Z_j)\} I(Y_i < Y_j) \hat{w}_i \quad (3)$$

Because $\hat{\Delta}_{x-z}^*$ is in the same form of the statistic Liu and Jin⁹ used to evaluate the changes in discrimination accuracy for item selection, under some regularity conditions it retains asymptotic normality that $\sqrt{n}(\hat{\Delta}_{x-z}^* - \mu) \rightarrow N(0, V)$ as $n \rightarrow \infty$, with $\mu = E(\Delta_{x-z}^*)$ and asymptotic variance V . The detail was given in the Appendix. The test statistic is then

$$TS_2 = \frac{\hat{\Delta}_{x-z}^*}{se(\hat{\Delta}_{x-z}^*)}, \text{ where } se(\hat{\Delta}_{x-z}^*) \text{ can be estimated empirically.}$$

3 Simulation study

We conducted a simulation study to examine the finite sample performance of the estimator for discrimination accuracy and the statistical test for detection of difference in discrimination accuracy using the procedures with and without the weight taking into account random censoring, where the unweighted procedure is the R function `rcorr.cens`. Sample sizes of $n = 150$ and 300 were used in the two scenarios, one without censoring and one with 30% and 60% randomly censored responses. For each case, we generated 1,000 data sets. For the data set with uncensored responses, we generated n independent triplets (X_i, Z_i, T_i) , $i = 1, \dots, n$. Note that when (X, T) follows a bivariate normal distribution, C_X is a monotonic function of the bivariate correlation coefficient r only, with $C_X = 0.5$ corresponding to $r = 0$. We generate (X_i, T_i) and (Z_i, T_i) from bivariate normal distributions with correlation coefficient for $C_Z = 0.5$ and 0.7 and various C_X such that the difference $\delta = C_X - C_Z = 0, 0.02, 0.04, 0.06, 0.08$, and 0.10 . For the data set with censored responses, we first generated n independent triplets as described above. Then we generated n independent random numbers from uniform distribution $U(0, \theta)$ for censoring variable Q with θ specified according to the preset censoring proportion. Afterwards, we calculated the observed values of the outcome variable $Y_i = \min(\exp(T_i), Q_i)$ and $d_i = I(T_i < Q_i)$.

Tables 1 and 2 summarize the results. With uncensored response the two procedures give almost identical results. As expected, in presence of random censoring on the response variable, the mean estimates of C_X using procedure with weight adjusting for random censoring are very close to the true values. In contrast, the unweighted estimator ignoring random censoring has bias that increases with proportion of censored responses. It is interesting to note that the unweighted procedure produced negligible biases in estimating δ when percent of random censoring is low (30%), while the biases become somewhat larger when censoring is heavy (60%). Nevertheless, both estimators have standard deviations decrease with increasing discrimination accuracy, sample size or proportion of response uncensored.

The percent of rejections of the null hypothesis of $\delta = 0$ estimates the power of a statistical test. Because the two tests are unbiased, the estimated power in the simulation study is close to the nominal level of 5% when $\delta = 0$, independent of random censoring. As expected, both tests have estimated power increases with increasing δ for a fixed sample size, or increases with sample size for a given value of δ , with and without independent random censoring on the response variable. With $\delta > 0$, however, for given sample size and δ , the power estimates of the two tests decrease with increasing percent of censoring. When percent of censoring is 30% or less, the two tests have comparable power estimates; while with heavy censoring

(60%), the unweighted test has larger power estimates than the weighted test, as a result of overestimating δ .

4 Applications

4.1 Comparing blood arsenic and manganese in predicting children's IQ test scores

In the children's study investigating the association between exposures to arsenic and manganese and children's intellectual function in Bangladesh, as described in Wasserman *et al.*¹, a culturally adapted version of the Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV: Wechsler 2003) was administered to children individually to assess their intellectual function. With excellent psychometrics, WISC-IV provides measures of general intellectual ability (Full Scale IQ) and specific cognitive domains (Perceptual Reasoning, Processing Speed Indices, Verbal Comprehension, and Working Memory). The WISC-IV materials were translated into Bengali and back-translated into English, with the incorporation of culturally appropriate adaptations. Raw scores for Verbal Comprehension, Perceptual Reasoning, Working Memory and Processing Speed were added to generate a measure of Full Scale intelligence.

Following a survey of the well characteristics of the villages enrolled in the "Health Effects of Arsenic Longitudinal Study" in Araihaazar, Bangladesh¹¹, all household wells within commuting distance of the field clinic were designated into one of four groups: i) High arsenic (> 10 ug/L) and high manganese (> 500 ug/L), ii) High arsenic (> 10 ug/L) and low manganese (< 500 ug/L), iii) Low arsenic (< 10 ug/L) and high manganese (> 500 ug/L), and iv) Low arsenic (< 10 ug/L) and low manganese (< 500 ug/L). From the villages, a random sample of children estimated to be between 8 and 11 years old were recruited with approximately 75 each well category group.

Blood arsenic and manganese as biomarkers of the exposures were measured using the blood samples from 299 school children who were administered individually the WISC-IV test to assess their intellectual function. The two biomarkers were correlated (Spearman correlation coefficient $r = 0.1256$, $p = 0.0299$).

In this study, evidence was found that higher level of blood arsenic was significantly related to lower test scores of Full Scale IQ ($r = -0.1507$, $p = 0.0091$) and three cognitive domains of: Perceptual Reasoning ($r = -0.1409$, $p = 0.0148$); Verbal Comprehension ($r = -0.1450$, $p = 0.0121$); and Working Memory ($r = -0.1428$, $p = 0.0135$). Similarly, higher levels of blood manganese were significantly related to lower scores of the same set of the tests including Full Scale IQ ($r = -0.1324$, $p = 0.0220$), Perceptual Reasoning ($r = -0.1727$, $p = 0.0027$), Verbal Comprehension ($r = -0.1306$, $p = 0.00239$) and Working Memory ($r = -0.1767$, $p = 0.0022$).

Because the correlations were not large in magnitude, the estimated discrimination accuracy of blood arsenic C_X or blood manganese C_Z (with opposite sign to account for negative association) were not high. Blood arsenic had slightly higher discrimination accuracy than blood manganese in predicting Full score IQ scores ($\hat{C}_X = 0.5504$ vs. $\hat{C}_Z = 0.5444$) and Verbal Comprehension test scores ($\hat{C}_X = 0.5493$ vs. $\hat{C}_Z = 0.5456$), while blood manganese

had slightly higher discrimination accuracy than blood arsenic for the subscales of Perceptual Reasoning ($\hat{C}_z=0.5595$ vs. $\hat{C}_x=0.5494$) and Working Memory ($\hat{C}_z=0.5616$ vs. $\hat{C}_x=0.5496$). These differences between blood arsenic and blood manganese, however, were not statistically significant in the discrimination accuracy predicting children's intellectual function (test statistic $TS_1 < 0.193$, p-values > 0.66).

4.2 Comparing prognostic factors of patients with primary biliary cirrhosis

Serum albumin and bilirubin are the two important continuous predictors of prognosis in primary biliary cirrhosis (PBC)¹². They are associated with survival time among PBC patients. The data set of 418 PBC patients presented in Fleming and Harrington¹⁰ provides us with an opportunity to test which of the two measures has better discrimination accuracy in predicting the survival time of PBC patients. In the sample, the PBC patients ranged in age between 26 and 76 years, with a mean age of 51; 89.5% were female. At baseline, all patients' serum albumin and bilirubin levels were measured. Their serum albumin levels ranged between 1.96 and 4.64 (mg/dl) with a median of 3.53 (mg/dl), and serum bilirubin levels ranged between 0.3 and 28 (mg/dl) with a median of 1.40 (mg/dl). The two serum measures were inversely correlated, with a Spearman correlation coefficient of -0.3367 ($p < 0.0001$). During the follow up period, 38.5% of the patients ($n = 161$) died, with a mean survival time of 3.77 years among those who died. The follow up time for the 257 survivors had a mean length of 6.18 years. Lower serum bilirubin was related to longer survival time; the discrimination accuracy for survival with serum bilirubin (with opposite sign to account for negative association) was $\hat{C}_x^*=0.7507$. Serum albumin, on the other hand, was positively related to survival time, and its discrimination accuracy was $\hat{C}_z^*=0.6430$. The test statistic for the difference in discrimination accuracy was $TS_2 = 3.84$ ($p = 0.0001$), suggesting that serum bilirubin should be a better predictor of survival time than serum albumin.

5 Discussion

In biomedical research, it is useful to compare paired biomarkers in discrimination accuracy of predicting a quantitative health outcome. When survival time is the outcome variable, it could be subject to independent random censoring. Using Harrell's C-index to measure discrimination accuracy of biomarkers with continuous measures, we adapt a consistent estimator using weight to take into account random censoring. Based on the non-parametric estimator of discrimination accuracy for a predictor, we apply a statistical test to detect difference between paired biomarkers in predicting a health outcome possibly subject to random censoring. The test statistic is an extension of the well known unbiased Wilcoxon test for paired predictors of a health outcome uncensored or censored by a constant.

In a simulation study, we examined finite sample performance of the two procedures with and without a weight to take into account random censoring in estimating discrimination accuracy and in testing difference between paired biomarkers in predicting a common outcome. With completely observed responses, the two procedures performed equally well. In presence of random censoring, however, the result was in favor of the weighted method in estimation of discrimination accuracy. To estimate or test for the difference in

discrimination accuracy between paired predictors, although the weighted test worked better in the cases with heavy random censoring, the two procedures produced similar result when percent of random censoring was low (30%).

As all other non-parametric methods, the weighted test needs a large sample size to detect a small difference in discrimination accuracy between continuous paired predictors, especially when response variable is subject to random censoring. In this paper, we considered random censoring that is independent of predictors. A generalized method allowing for dependence on predictors could be developed by modeling the censoring time and worth further investigation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was partially supported by the National Institute of Environment Health Sciences (grants NOs: P42 ES10349 and P30 ES09089). The authors gratefully thank the reviewers for valuable comments and suggestions.

References

1. Wasserman GA, Liu X, Parvez F, Factor-Litvak P, Ahsan H, Levy D, Kline J, Van Geen A, Mey J, Slavkovich V, Siddique A, Islam T, Graziano J. Arsenic and manganese exposure and children's intellectual function. *NeuroToxicology*. 2011. in press.
2. Harrell FE, Lee KL, Calife RM, Pryor DB, Rosati RA. Regression modeling strategies for improved prognostic prediction. *Statistics in Medicine*. 1984; 3(2):143–152. [PubMed: 6463451]
3. Harrell FE, Lee KL, Mark DB. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*. 1996; 15(4):361–387. [PubMed: 8668867]
4. Nam, BH.; D'Agostino, RB. Goodness-of-Fit tests and Model Validity. Boston: Birkhauser; 2002. Discrimination Index, the area under the ROC curve; p. 267-279.
5. Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine*. 2004; 23:2109–2123. [PubMed: 15211606]
6. Newson R. Confidence intervals for rank statistics: Somers' D and extensions. *Stata Journal*. 2006; 6:309–334.
7. Harrell, FE.; Williams, S. Rank Correlation for Paired Predictors with a Possibly Censored Response, and Integrated Discrimination Index; Hmisc version 3.9-3. 2012. p. 208-211. <http://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf>
8. Koziol JA, Ji Z. The concordance index C and the Mann–Whitney parameter $\Pr(X>Y)$ with randomly censored Data. *Biometrical Journal*. 2009; 51(3):467–474.10.1002/bimj.200800228 [PubMed: 19588452]
9. Liu X, Jin Z. A non-parametric approach to scale reduction for uni-dimensional screening scales. *International Journal of Biostatistics*. 2009; 5(1) Article 7.
10. Fleming, TR.; Harrington, DP. Counting Processes and Survival Analysis. New York: Wiley; 1991.
11. Ahsan H, Chen Y, Parvez F, Argos M, Hussain AI, Momotaj H, Levy D, Van Geen A, Howe G, Graziano J. Health effects of arsenic longitudinal study (HEALS): Description of a multidisciplinary epidemiologic investigation. *Journal of Exposure Science and Environmental Epidemiology*. 2006; 16:191–205. [PubMed: 16160703]

12. Llorenc P, Caballería A, Rodés J. Excellent long-term survival in patients with primary biliary cirrhosis and biochemical response to ursodeoxycholic acid. *Gastroenterology*. 2006; 130(3):715–720. [PubMed: 16530513]
13. Lee, AJ. *U-Statistics: Theory and Practice*. Marcel Dekker Inc; New York and Basel: 1990.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Result from 1000 simulated data sets with sample size of $N = 150$.

0% censored			Weighted			Unweighted		
Cz	δ	$\hat{\delta}$ Mean	$\hat{C}_x$ Mean (SD)	% reject H_0	$\hat{\delta}$ Mean	$\hat{C}_x$ Mean (SD)	% reject H_0	
.50	.00	-.0007	.5007 (.0272)	5.1	-.0007	.5007 (.0272)	5.1	
	.02	.0199	.5192 (.0265)	6.8	.0199	.5192 (.0265)	7.1	
	.04	.0390	.5394 (.0267)	18.4	.0390	.5394 (.0267)	18.4	
	.06	.0589	.5588 (.0276)	32.7	.0589	.5588 (.0276)	32.9	
	.08	.0796	.5802 (.0263)	53.5	.0796	.5802 (.0263)	53.7	
.70	.10	.1006	.6002 (.0264)	72.6	.1006	.6002 (.0264)	72.7	
	.00	.0008	.7002 (.0222)	4.9	.0008	.7002 (.0222)	4.9	
	.02	.0209	.7207 (.0223)	11.0	.0209	.7207 (.0223)	11.1	
	.04	.0400	.7398 (.0203)	29.0	.0400	.7398 (.0203)	29.3	
	.06	.0606	.7601 (.0192)	61.2	.0606	.7601 (.0192)	61.5	
30% censored	.08	.0780	.7783 (.0183)	83.8	.0780	.7783 (.0183)	84.0	
	.10	.0998	.8000 (.0161)	97.7	.0998	.8000 (.0161)	97.7	
.50	.00	-.0021	.4989 (.0289)	5.1	-.0014	.4993 (.0317)	5.0	
	.02	.0237	.5201 (.0292)	7.7	.0255	.5218 (.0320)	9.0	
	.04	.0410	.5400 (.0290)	16.2	.0444	.5434 (.0313)	16.7	
	.06	.0597	.5599 (.0296)	28.1	.0657	.5662 (.0320)	31.5	
	.08	.0809	.5798 (.0294)	49.2	.0888	.5873 (.0317)	51.5	
.70	.10	.0986	.5991 (.0293)	65.7	.1078	.6083 (.0309)	67.8	
	.00	-.0015	.6979 (.0242)	5.0	-.0015	.7148 (.0255)	5.7	
	.02	.0191	.7188 (.0232)	8.8	.0198	.7363 (.0243)	8.4	
	.04	.0400	.7402 (.0221)	24.0	.0415	.7582 (.0230)	25.4	
	.06	.0600	.7588 (.0209)	49.8	.0624	.7782 (.0215)	52.9	
	.08	.0826	.7792 (.0216)	82.0	.0850	.7984 (.0214)	82.8	

0% censored			Weighted			Unweighted		
Cz	δ	$\hat{\delta}$ Mean	$\hat{C}_x$ Mean (SD)	% reject H_0	$\hat{\delta}$ Mean	$\hat{C}_x$ Mean (SD)	% reject H_0	
60% censored								
.50	.00	.0017	.5012 (.0394)	4.9	.0016	.5003 (.0428)		6.8
	.02	.0200	.5205 (.0395)	5.6	.0255	.5247 (.0415)		7.4
	.04	.0443	.5425 (.0381)	12.1	.0514	.5501 (.0404)		13.8
	.06	.0626	.5626 (.0379)	21.3	.0736	.5744 (.0402)		25.9
	.08	.0841	.5857 (.0395)	31.2	.1015	.6018 (.0401)		42.2
	.10	.1023	.6027 (.0393)	45.3	.1220	.6226 (.0398)		56.9
.70	.00	.0040	.7107 (.0354)	4.9	.0033	.7421 (.0331)		5.3
	.02	.0257	.7335 (.0339)	8.6	.0258	.7667 (.0314)		9.7
	.04	.0407	.7510 (.0317)	15.0	.0450	.7860 (.0283)		20.4
	.06	.0636	.7708 (.0317)	31.0	.0659	.8059 (.0278)		41.3
	.08	.0838	.7912 (.0302)	53.4	.0866	.8267 (.0255)		65.8
	.10	.1040	.8132 (.0280)	73.4	.1065	.8478 (.0225)		86.0

Table 2

Result from 1000 simulated data sets with sample size of $N = 300$.

0% censored		Weighted			Unweighted		
Cz	δ	$\hat{\delta}$ Mean	$\hat{C}_x$ Mean (SD)	% reject H_0	$\hat{\delta}$ Mean	$\hat{C}_x$ Mean (SD)	% reject H_0
.50	.00	.0004	.5008 (.0193)	4.9	.0004	.5008 (.0193)	4.9
	.02	.0191	.5194 (.0192)	10.5	.0191	.5194 (.0192)	10.5
	.04	.0398	.5402 (.0193)	31.9	.0398	.5402 (.0193)	32.2
	.06	.0596	.5596 (.0185)	59.5	.0596	.5596 (.0185)	59.9
	.08	.0790	.5799 (.0188)	84.1	.0790	.5799 (.0188)	84.1
	.10	.0992	.5999 (.0186)	95.8	.0992	.5999 (.0186)	95.8
.70	.00	.0001	.6997 (.0163)	5.1	.0001	.6997 (.0163)	5.2
	.02	.0204	.7196 (.0152)	14.8	.0204	.7196 (.0152)	14.8
	.04	.0404	.7405 (.0148)	54.4	.0404	.7405 (.0148)	54.6
	.06	.0600	.7595 (.0130)	89.6	.0600	.7595 (.0130)	89.7
	.08	.0797	.7798 (.0128)	99.3	.0797	.7798 (.0128)	99.3
	.10	.1004	.8004 (.0115)	100	.1004	.8004 (.0115)	100
30% censored							
.50	.00	.0002	.5007 (.0215)	5.0	.0002	.5007 (.0232)	4.7
	.02	.0195	.5189 (.0209)	11.1	.0217	.5209 (.0225)	11.6
	.04	.0416	.5402 (.0207)	31.3	.0456	.5443 (.0231)	32.6
	.06	.0591	.5599 (.0208)	53.0	.0648	.5658 (.0225)	54.5
	.08	.0792	.5792 (.0206)	75.8	.0868	.5867 (.0224)	77.5
	.10	.0993	.5990 (.0198)	93.1	.1091	.6087 (.0213)	93.2
.70	.00	.0000	.6979 (.0178)	5.0	.0002	.7145 (.0187)	5.0
	.02	.0210	.7190 (.0174)	14.7	.0221	.7369 (.0182)	14.7
	.04	.0402	.7383 (.0156)	45.5	.0420	.7568 (.0162)	46.9
	.06	.0590	.7579 (.0153)	81.2	.0611	.7768 (.0158)	81.6
	.08	.0795	.7786 (.0147)	97.4	.0819	.7976 (.0151)	97.7

0% censored			Weighted			Unweighted		
Cz	δ	$\hat{\delta}$ Mean	$\hat{C}_x$ Mean (SD)	% reject H_0	$\hat{\delta}$ Mean	$\hat{C}_x$ Mean (SD)	% reject H_0	
60% censored								
.50	.00	.0004	.5000 (.0285)	5.1	.0018	.5010 (.0294)	6.5	
	.02	.0197	.5192 (.0289)	8.9	.0235	.5226 (.0305)	9.2	
	.04	.0390	.5394 (.0268)	16.9	.0462	.5471 (.0288)	20.5	
	.06	.0633	.5623 (.0269)	35.5	.0749	.5742 (.0277)	44.7	
	.08	.0840	.5837 (.0267)	58.3	.0996	.5993 (.0277)	66.6	
.70	.10	.1041	.6047 (.0270)	77.1	.1240	.6235 (.0280)	87.2	
	.00	-.0007	.7079 (.0240)	5.1	.0004	.7413 (.0224)	5.7	
	.02	.0221	.7304 (.0233)	12.5	.0244	.7652 (.0222)	15.4	
	.04	.0418	.7498 (.0226)	29.1	.0443	.7855 (.0202)	35.8	
	.06	.0627	.7707 (.0221)	56.8	.0652	.8062 (.0191)	67.3	
	.08	.0830	.7912 (.0207)	84.4	.0854	.8263 (.0183)	91.5	
	.10	.1033	.8108 (.0197)	96.9	.1060	.8460 (.0161)	99.7	