Published in final edited form as:
*Comput Stat Data Anal*. 2011 January 1; 55(1): 226–235.

# Cutpoint selection for discretizing a continuous covariate for generalized estimating equations

**Gisela Tunes-da-Silva**[a,*] and **John P. Klein**[b]

[a]Department of Statistics, University of São Paulo, São Paulo, São Paulo, Brazil

[b]Department of Population Health, Medical College of Wisconsin, Milwaukee, WI, USA

## Abstract

We consider the problem of dichotomizing a continuous covariate when performing a regression analysis based on a generalized estimation approach. The problem involves estimation of the cutpoint for the covariate and testing the hypothesis that the binary covariate constructed from the continuous covariate has a significant impact on the outcome. Due to the multiple testing used to find the optimal cutpoint, we need to make an adjustment to the usual significance test to preserve the type-I error rates. We illustrate the techniques on one data set of patients given unrelated hematopoietic stem cell transplantation. Here the question is whether the CD34 cell dose given to patient affects the outcome of the transplant and what is the smallest cell dose which is needed for good outcomes.

### Keywords

Dichotomized outcomes; Generalized estimating equations; Generalized linear model; Pseudo-values; Survival analysis

## 1. Introduction

In many medical studies it is of interest to investigate the relationship between explanatory variables, such as prognostic factors, treatment factors or patient characteristics, and the outcome. The outcome may be continuous, such as the time to some event or the level of some assayed enzyme, or it may be a discrete outcome, such as an indicator of relapse or death. We may also be interested in state occupation probabilities in multistate processes modeled via a pseudo-observation approach applied to censored data as discussed in, for example, Andersen et al. (2003). In any case, regression models based on generalized estimating equations (GEEs) (McCullagh and Nelder, 1989; Liang and Zeger, 1986) can be applied to examine the effects of covariates on the outcome.

In many cases the covariates of interest are continuous in nature. While they could be modeled as such, many clinicians find the interpretation of such covariates difficult and they prefer to model these effects as categorical or binary covariates reflecting different prognostic groups of patients based on the measured value of the continuous covariate. In other cases, the covariate may represent the dose of some drug or some other therapeutic



*Corresponding author. Tel.: +55 11 30916235. tunes@ime.usp.br (G. Tunes-da-Silva).

**Appendix. Supplementary data**

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.csda.2010.02.016.

agent, and the clinician is interested in identification of a therapeutic threshold. Even though evaluation of a variable's prognostic value is best done with the variable in its continuous form and there is a possible loss of information when categorizing a continuous covariate, the need for thresholds for clinical use and treatment decisions justifies the development of appropriate statistical methods for finding optimal cutpoints. Methodologies for finding optimal cutpoints are also needed and used in various tree building algorithms (Lausen et al., 2004).

When discretizing a continuous covariate and then testing for the covariate effect, a number of techniques can be used. One group of techniques relies on the investigator to provide cutpoints based on historical data or it uses cutpoints based on a split into groups at a predetermined percentile of the continuous covariate. Another approach is to let the data decide on the cutpoint and then perform the test of the covariate effect on the two resulting groups. In most cases the continuous covariate in this approach is split into groups based on either the largest value of the likelihood or the largest value of some two-sample test statistic after a search of possible cutpoints. This selection of the largest likelihood or test statistic leads to an inflated type-I error due to multiple testing, so some correction needs to be applied to obtain the correct type-I error.

A data-dependent cutpoint selection and test adjustment has been proposed for survival data by a number of authors. The approach is based on a Cox proportional hazards model with a single covariate defined as 1 or 0 depending on the value of the continuous covariate. Jespersen (1986) based an adjusted test on the maximum value of the score statistic from the Cox model which suitably standardized converges to a Brownian bridge under the null hypothesis. Contal and O'Quigley (1999) modify the log rank test statistic and show that the process consisting of the score statistic using cutpoints at the order statistics of the continuous covariates converges to a Brownian bridge process. Lausen and Schumacher (1992, 1996) showed that for any standardized test statistic $C(\delta)$ for the two-sample problem with groups defined by a threshold parameter, δ, the following convergence result holds:

$$\max\{|C(\delta)|, \delta \in [X_{(n\varepsilon)}, X_{(n[1-\varepsilon])}]\} \Rightarrow \sup_{u\in[\varepsilon,1-\varepsilon]} \frac{|W^0(u)|}{\sqrt{u(1-u)}}.$$

Here $X_{(m)}$ is the $m$th-order statistic of the continuous covariate and $W^0$ is a standard Brownian bridge process. Klein and Wu (2004) compare these estimators and extend the Contal and O'Quigley approach to the accelerated failure time model and the Cox model with additional covariates.

In this note we examine these data-driven methods in a generalized linear model framework. We are particularly interested in using these techniques in pseudo-observation regression problems. The pseudo-observation approach has been suggested as a method for direct censored data regression modeling for the survival function (Logan et al., 2008), for the cumulative incidence function (Klein and Andersen, 2005), for the mean survival time (Andersen et al., 2004), for multistate probabilities (Andersen et al., 2003) and for mean quality of life (Andrei and Murray, 2007). In this approach, pseudo-observations are formed as the difference between the full sample and leave-one-out estimator based on an approximately unbiased estimator of the parameter of interest. These pseudo-observations are then used in a GEE model.

The methods we developed will be illustrated using data from a study from the Center for Blood and Marrow Transplantation Registry. In this study 709 patients with Acute Myeloid

Leukemia ($n = 395$) or Acute Lymphocytic Leukemia ($n = 397$) in first ($n = 204$) or second ($n = 488$) complete remission were given a Bone Marrow Transplant (BMT) from an unrelated donor transplant. The cell source for all the transplants was bone marrow. The research question of interest was: "What level of CD34 cells in the graft is needed to lower the death in remission rates and is there a threshold dose of CD34 cells to affect the relapse rates?". The questions require an estimation of the threshold dose level of CD34 cells and a comparison of the "high" and "low" CD34 dose groups for death in remission and relapse cumulative incidence probabilities.

## 2. Techniques for the generalized linear model and generalized estimating equations with a single covariate

In this section we show two approaches to discretizing a continuous covariate and testing for its significance based on the generalized estimating equation approach. Here we assume that the response for individual $i$, $i = 1, \ldots, n$, is possibly multivariate and is denoted by $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})^t$. We allow for the $\mathbf{Y}$ to be continuous or discrete outcomes. Let $\boldsymbol{\mu}_i = \mathrm{E}[\mathbf{Y}_i] = (\mu_{i1}, \ldots, \mu_{iJ})^t$ be the mean vector for individual $i$. We assume a generalized linear model (GLM) framework, so that the distribution of the $Y_{ij}$ belongs to the same family with mean $\mu_{ij}$ and common dispersion parameter $\phi$ (Liang and Zeger, 1986). Also, let $X_i$ be the continuous covariate associated with individual $i$. We assume that the mean and the dichotomized covariate, $Z^{\delta}$, are related by

$$g(\mu_{ij}) = \alpha_j + \gamma Z_i^{\delta},$$

where $g(\cdot)$ is a known link function, $\alpha_j$ for $j = 1, \ldots, J$, $\delta$ and $\gamma$ are unknown parameters and $Z_i^{\delta}$ is the dichotomized covariate obtained from $X$, namely

$$Z_i^{\delta} = \begin{cases} 1, & \text{if } X_i \le \delta, \\ 0, & \text{if } X_i > \delta. \end{cases} \quad (1)$$

We are interested in estimating the threshold parameter $\delta$ as well as the inference on the parameter $\gamma$. For a given $\delta$, estimates of parameters can be based on the generalized estimating equations (GEEs) suggested by Liang and Zeger (1986):

$$\sum_{i=1}^{n} \left( \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} \right)^T \mathbf{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \sum_{i=1}^{n} \mathbf{U}_i(\boldsymbol{\beta}) = \mathbf{0}, \quad (2)$$

where $\boldsymbol{\beta} = (\gamma, \alpha_1, \ldots, \alpha_J)^T$ and $\mathbf{V}_i$ is a working covariance matrix. GEE estimates of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$ are found by solving Eq. (2). A sandwich estimator can be used to estimate the covariance matrix of $\boldsymbol{\beta}$:

$$\widehat{\mathrm{Var}(\hat{\beta})} \approx \mathbf{I}(\hat{\beta})^{-1} \widehat{\mathrm{Var}(\mathbf{U}(\hat{\beta}))} \mathbf{I}(\hat{\beta})^{-1},$$

where

$$\mathbf{I}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left( \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} \right)^T \mathbf{V}_i^{-1} \left( \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} \right)$$

and

$$\widehat{\mathrm{Var}(\mathbf{U}(\boldsymbol{\beta}))}=\sum_{i=1}^{n}\mathbf{U}_i(\boldsymbol{\beta})^T\mathbf{U}_i(\boldsymbol{\beta}).$$

Standard packages such as GEESE in R or PROC GENMOD in SAS can be used to obtain estimates and their variances for a given δ.

In our setup, we are interested on the hypothesis $H_0 : \gamma = 0$. One test that can be used is the Wald test given by $\chi_W^2=\hat{\gamma}/(\mathrm{var}(\hat{\gamma}))^{1/2}$, which has a standard normal distribution when $H_0$ is true and δ is known *a priori*. A second test is the generalized score test (Boos, 1992). For the two-sample problem with no covariates, Liu et al. (2008) show that the score statistic for the test $H_0 : \gamma = 0$ is given by

$$\chi_S^2=\left(\frac{n}{n_0n_1}\right)^2\frac{\left(\sum_{i=1}^{n}Z_i\sum_{j=1}^{J}(\dot{g}_j^0)^2(Y_{ij}-\bar{Y}_j)\right)^2}{\left(\frac{\sum_i(1-Z_i)\left[\sum_{j=1}^{J}(\dot{g}_j^0)^2(Y_{ij}-\bar{Y}_j)\right]^2}{n_0^2}+\frac{\sum_i Z_i\left[\sum_{j=1}^{J}(\dot{g}_j^0)^2(Y_{ij}-\bar{Y}_j)\right]^2}{n_1^2}\right)},$$

where $n_0$ is the number of observations with $Z_i^\delta=0$, $n_1$ is the number of observations with $Z_i^\delta=1$, $\bar{Y}_j=\sum_{i=1}^{n}Y_{ij}/n$ is the estimator of $\alpha_j$ under the null hypothesis, $\dot{g}(x)=\frac{\partial g^{-1}(x)}{\partial x}$ and $\dot{g}_j^0=\dot{g}(\bar{Y}_j)$. Under the null hypothesis, $\chi_S^2$ has a chi-square distribution with one degree of freedom when δ is known *a priori*.

In order to estimate δ, one common approach is to compute either $\chi_S^2$ or $\chi_W^2$ for values of δ in some range and pick up the value that maximizes $\chi_S^2$ (or $\chi_W^2$). This procedure is appropriate for estimating δ, but an adjustment must be made in order to make an inference for γ and preserve the type-I error rate.

The first adjustment that can be made is a GEE version of the approach in Lausen and Schumacher (1992, 1996). Both the generalized score statistic and the Wald statistic can be used. In order to apply this approach, the possible range of threshold values must be restricted to $[X_{(n\varepsilon)}; X_{(n(1-\varepsilon))}]$, where $X_{(k)}$ is the $k$th-order statistic of the continuous variable and $0 < \varepsilon < 0.5$. Let $C(\delta)$ be the test statistic for the two-sample problem with groups defined by δ. It can be shown that

$$\max\{|C(\delta)|, \delta \in [X_{(n\varepsilon)};X_{(n(1-\varepsilon))}]\} \Rightarrow \sup_{u\in[\varepsilon,1-\varepsilon]}\frac{|W^0(u)|}{\sqrt{u(1-u)}}$$

as $n \to \infty$, where $W^0$ is a Brownian bridge. Details of the proof can be found in Lausen and Schumacher (1992) and Billingsley (1968, chap. 4). Miller and Siegmund (1982) show that, for $b > 0$,

$$P\left(\sup_{u\in[\varepsilon,1-\varepsilon]}\frac{|W^0(u)|}{\sqrt{u(1-u)}}>b\right)\approx\varphi(b)\left(b-\frac{1}{b}\right)\log\left(\frac{(1-\varepsilon)^2}{\varepsilon^2}\right)+4\frac{\varphi(b)}{b}+o\left(\frac{\varphi(b)}{b}\right),$$

where $\varphi(\cdot)$ is the standard normal density function. This result motivates the following approximation for a corrected $p$-value:

$$p_{cor}=\varphi(z)\left(z-\frac{1}{z}\right)\log\left(\frac{(1-\varepsilon)^2}{\varepsilon^2}\right)+4\frac{\varphi(z)}{z},$$

where $z=\Phi^{-1}(1-p/2)$, $\Phi(\cdot)$ is the standard normal cumulative distribution function and $p$ is the unadjusted $p$-value computed at the maximum of the test statistic. Notice that the approximation is valid for $p<0.5$. Inferences for $\boldsymbol{\gamma}$ can be made by computing the corrected $p$-value for the maximum value of the test statistic.

The second approach considered here is a modification of the method in Contal and O'Quigley (1999). To construct the test statistic, recall that, for the two-sample problem without covariates, assuming that the working covariance matrix is the identity matrix, the score function can be written as

$$\mathbf{U}(\boldsymbol{\beta})=\sum_{i=1}^{n}\mathbf{U}_i(\boldsymbol{\beta}),$$

$$\mathbf{U}_i(\boldsymbol{\beta})=\begin{pmatrix} Z_i^{\delta}\left(\sum_{j=1}^{J}\dot{g}(\eta_{ij})\,(Y_{ij}-\mu_{ij})\right)\\ \dot{g}(\eta_{i1})\,(Y_{i1}-\mu_{i1})\\ \dot{g}(\eta_{i2})\,(Y_{i2}-\mu_{i2})\\ \vdots\\ \dot{g}(\eta_{iJ})\,(Y_{iJ}-\mu_{iJ})\end{pmatrix},$$

where $\eta_{ij}=\alpha_j+Z_i^{\delta}\gamma$ is the linear predictor and $\mu_{ij}=g^{-1}(\eta_{ij})$. Let

$$U_{\gamma}^{\delta}=\sum_{i=1}^{n}Z_i^{\delta}\left(\sum_{j=1}^{J}\dot{g}\,(\eta_{ij})\,(Y_{ij}-\mu_{ij})\right)=\sum_{i=1}^{n}Z_i^{\delta}\xi_i, \quad (3)$$

$\xi_i=\sum_{j=1}^{J}\dot{g}\,(\eta_{ij})\,(Y_{ij}-\mu_{ij})$ and let $U_0^{\delta}$ be the score evaluated under the null hypothesis.

Billingsley (1999, pg. 196) proved the following theorem.

**Theorem 1.** *Let* $\{\xi_1, \xi_2, \ldots\}$ *be a stationary and ergodic process for which its conditional expected value* $\mathrm{E}(\xi_i|\xi_1, \ldots, \xi_{i-1})$ *is equal to zero with probability* 1 *and for which the expected value* $\mathrm{E}(\xi_i^2)=\upsilon^2$ *is positive and finite. Define*

$$R_n=\xi_1+\cdots+\xi_n.$$

*If*

$$S_n(p)=\frac{R_{[np]}}{\upsilon\sqrt{n}}$$

*for* $p \in [0, 1]$; *then* $S_n(p) \Rightarrow W(p)$ *as* $n \to \infty$, *where* $W(p)$ *is a Brownian motion process.*

In our setup, in order to apply this result, we must first put the data in increasing order with respect to the continuous covariate $X$. Now, with the ordered data, define

$$\xi_i=\sum_{j=1}^{J}\dot{g}\,(\eta_{ij})\,(Y_{ij}-\mu_{ij})$$

and assume that $\mathbf{Y}_i \perp \mathbf{Y}_{i'}$. With this assumption, we can then see that the condition $E(\xi_i|\xi_1, \ldots, \xi_{i-1}) = 0$ is verified. It is important to notice that we are assuming that individuals are independent, but observations within the same individual may be correlated. Also, under the null hypothesis $H_0 : \gamma = 0$, recalling that we are assuming a common dispersion parameter for distribution of the responses, we have that the $\mathbf{Y}_i$ are independent and identically distributed (i.i.d.), so the $\xi_i$ are also independent and identically distributed. Therefore, the $\xi_i$ are stationary and ergodic. Finally, under the null hypothesis, we have

$$\upsilon^2=\mathrm{E}(\xi_i^2)=\dot{\mathbf{g}}^T\mathbf{V}_Y\dot{\mathbf{g}},$$

where $\mathbf{V}_Y$ is the covariance matrix of $\mathbf{Y}_i$ and $\dot{\mathbf{g}}$ is the vector of elements $\dot{g}\,(\eta_{ij}^0)$, which is finite. With the ordered data, it is clear that there is a value $\delta_p$ for which $U_\gamma^{\delta_p}=\sum_{i=1}^{[np]}\xi_i$. In fact, any value of $\delta_p$ between $X_{([np])}$ and $X_{([np]+1)}$ satisfies $U_\gamma^{\delta_p}=\sum_{i=1}^{[np]}\xi_i$, so the midpoint can be used. If we define

$$S_n(p)=\frac{1}{\upsilon\sqrt{n}}U_\gamma^{\delta_p}=\frac{1}{\upsilon\sqrt{n}}\sum_{i=1}^{[np]}\xi_i,$$

we have $S_n(p) \Rightarrow W(p)$ as $n \to \infty$. Also, the process defined by $S_n(p) - pS_n(1)$ converges to a Brownian bridge (O'Quigley, 2008, pg. 245).

In practice, we replace the unknown quantities in the $\xi_i$ by consistent estimators under the null, so we work with $\hat{S}_n(p)$. Following O'Quigley (2008, pg. 236), also based on Slutsky's theorem, the large sample properties of the resulting test statistic will not be affected. Also, in general, the values of $\alpha_j$ and $\upsilon^2$ are not known and we substitute by their estimators. The parameters $\alpha_j$ can be estimated using the estimating equations under the null hypothesis, and for $\upsilon^2$ we can use the sample variance $\frac{\sum_{i=1}^{n}(\hat{\xi}_i-\bar{\xi})^2}{n}$. Hence

$$\hat{S}_n(p)=\frac{1}{\hat{v}\sqrt{n}}\hat{U}_{\gamma}^{\delta_p}=\frac{\sum_{i=1}^{[np]}\hat{\xi}_i}{\sqrt{\sum_{i=1}^{n}(\hat{\xi}_i-\bar{\xi})^2}}. \quad (4)$$

Therefore, because $\hat{S}_n(1) = 0$, we have that $\hat{S}_n(p) = \hat{S}_n(p) - p\hat{S}_n(1)$ (Klein and Wu, 2004) and, finally, $\hat{S}_n(p) \Rightarrow W^0(p)$ as $n \to \infty$, where $W^0(p)$ is a Brownian bridge. It has been shown that the distribution of the extreme value of the Brownian bridge (Billingsley, 1999, pg. 103) is given by

$$P\left(\sup_{p\in[0,1]}|W^0(p)|<b\right)=1+2\sum_{k=1}^{\infty}(-1)^k\exp\{-2k^2b^2\}. \quad (5)$$

If we order the $\xi$ by the continuous covariate for which we want to find a cutpoint, under the null hypothesis this sequence is formed by i.i.d. observations and the sequence of $\hat{S}_n(p)$ constructed with the $\xi$ ordered converges to a Brownian bridge. We can then use as test statistic the maximum of $\hat{S}_n(p)$, $p \in [0, 1]$. If the null hypothesis is rejected, we estimate as cutpoint any value between $X_{(m)}$ and $X_{(m+1)}$, where $X_{(m)}$ is the $m$th value of the ordered $Z$ and $m = np^*$, where $p^*$ is the value of $p$ for which $|\hat{S}_n(p)|$ is maximum. An important remark that should be made here is regarding the use of pseudo-observations. Although the pseudo-observations are computed based on a jackknife technique (and, therefore, they are computed based on $(n - 1)$ observations), it can be shown that under some conditions they can be approximated by independent and identically distributed variables (Graw et al., 2009), so the results for the $\xi$ remain valid.

## 3. The multiple covariate case

In the previous section, we considered the situation with only a single covariate in the model. However, in most practical applications there are other covariates, and we now discuss the extension to allow for many covariates in the model. The test statistic remains the same, so we discuss it here briefly.

As before, assume that $Y_{ij}$ is the outcome for the $j$th observation on individual $i$, $j = 1, \ldots, J$ and $i = 1, \ldots, n$. Also, let $X_i$ be the continuous covariate associated with individual $i$ and $\boldsymbol{Z}_i^*$ be a vector of other covariates. We assume that the mean and the covariates are related by

$$g(\mu_{ij})=\alpha_j+\gamma Z_i^{\delta}+\boldsymbol{\beta}^*\boldsymbol{Z}_i^*,$$

where $g(\cdot)$ is a known link function, $\alpha_j$ for $j = 1, \ldots, J$, $\gamma$ and $\boldsymbol{\beta}^*$ are the unknown parameters and $Z_i^{\delta}$ is the categorized covariate given by (1).

As before, we are interested in finding the cutpoint $\delta$ and testing the hypothesis about $\gamma$. For a given $\delta$, denote by $\boldsymbol{\beta} = (\gamma, \alpha_1, \alpha_2, \ldots, \alpha_J, \boldsymbol{\beta}^{*T})^T$ the vector of all unknown parameters. The estimating equations are given by (2) and the variance of estimates can be estimated by the sandwich estimator.

In order to test the hypothesis $H_0 : \gamma = 0$, the usual Wald test or the generalized score test (Boos, 1992) can be used. The Wald test can be obtained easily using any standard package

for GEEs. In order to derive an expression for the generalized score test, using an independence working covariance matrix $\mathbf{V}_i = \mathbf{I}$, we rewrite the score Eq. (2) as

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{U}_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{Z}_i^T \dot{\mathbf{G}}_i \, (\boldsymbol{Y}_i - \boldsymbol{\mu}_i),$$

where $\mathbf{Z}_i$ is the $(J \times p)$ matrix of covariates (design matrix) of the $i$th observation, i.e., it is the augmented matrix given by $\boldsymbol{Z}_i = \left(Z_i^{\delta} \mathbf{1}_J | \mathbf{I}_{J\times J} | \mathbf{1}_J \boldsymbol{Z}_i^{*T}\right)$, $p$ is the number of rows of the parameter vector $\boldsymbol{\beta}$, $\mathbf{1}_J$ is the vector with $J$ rows of 1s, $\mathbf{I}_{J\times J}$ is the identity matrix and $\dot{\mathbf{G}}_i = \mathrm{diag}(\dot{g}(\eta_{ij}))$ is the diagonal matrix of derivatives $\dot{g}(\eta_{ij}) = \frac{\mathrm{d}g^{-1}(x)}{\mathrm{d}x}$. The score test for $H_0 : \boldsymbol{\gamma} = 0$ is based on the first element of $\mathbf{U}(\boldsymbol{\beta})$ and the expression is exactly the same as that obtained for the single covariate case. As before, we must order the data increasing on the continuous covariate $X$ and let $U_{\gamma}^{\delta}$ be defined as in (3). We assume that the $\mathbf{Y}_i$ are independent random variables with mean $\boldsymbol{\mu}_i$ and the same covariance matrix $\boldsymbol{\Sigma}$, so the $(\mathbf{Y}_i - \boldsymbol{\mu}_i)$ are also independent with mean zero and common covariance matrix $\boldsymbol{\Sigma}$. Now suppose we have covariates $\boldsymbol{Z}_1^*, \boldsymbol{Z}_2^*, \ldots, \boldsymbol{Z}_n^*$ which constitute an i.i.d. sample from a distribution $P(\cdot)$. The random vectors $\dot{\mathbf{G}}_i(\mathbf{Y}_i - \boldsymbol{\mu}_i)$ are, therefore, i.i.d. with zero mean and variance $\mathrm{E}_P(\dot{\mathbf{G}}_i \boldsymbol{\Sigma} \dot{\mathbf{G}}_i)$. This allows us to apply the same results as in the situation without any covariates. Therefore, we can compute the sequence of $\hat{S}_n(p)$ given by (4) and the test statistic is given by $\sup_{p\in[0,\,1]} |\hat{S}_n(p)|$, with distribution given by (5).

## 4. Monte Carlo study

A simulation study was designed to compare the performance of the different test procedures in terms of their type-I error rate and power. We constructed the simulation for correlated observations, and $J$-variate normal random variables were generated for each subject, $J = 2$ or 4. We took the means to be

$$\mu_{i,j} = \alpha_j + \beta z_i, \quad (6)$$

for $i = 1, \ldots, n$, $j = 1, \ldots, J$. Here the $J$ variables had a common correlation of $\rho = 0$, 0.25, or 0.5. The $z$-values used in generating the $T$-values were obtained by dichotomizing a continuous covariate $X$ taken to be uniform $[-1, 1]$. Under the alternative hypothesis we used cutpoints of zero and 0.2. Total samples of size $n = 50$, 100, 200 and 400 were examined. Ten thousand samples were generated for each combination of $\gamma$, $\delta$, $\rho$, $J$ and $n$.

We computed the Wald and score test statistics along with both uncorrected and corrected $P$-values, and the sup score test statistic. We used $\varepsilon = 0$, 0.1, 0.15, 0.2 and 0.25 to restrict the range of possible values for the cutpoint. Although this restriction is not required for the new test statistic, we also computed the new test statistic using this range restriction.

Since we have a large number of scenarios, we applied analysis of variance (ANOVA) techniques to summarize the results. For the type-I error rate, we defined the outcome, $R$, as the percentage rejection rate minus the nominal rate of 5%, so good performance is indicated by values of the expectation of $R$ near 0. Negative values indicate a conservative test procedure and positive vales indicate that the test procedure inflates the type-I error. The first model fit to our results has the following effects:

$$(\text{Statistic} \times \text{Epsilon}) + \text{Sample size} + \text{Correlation} + \text{Number of points}. \quad (7)$$

When fitting the model without an intercept and normalizing the effects of the other factors to have a sum of zero, the estimates for the interaction terms have the interpretation as average deviations from the nominal level of 5% adjusted for the effects of the other factors. Table 1 compares the size of the tests by values of $\varepsilon$. The results on Table 1 show clearly that an adjustment must be made because there is an enormous inflation of type-I error rate for both the uncorrected Wald and score test statistics. The adjusted Wald statistic also inflates the type-I error rate a little bit, but the inflation gets smaller when we increase the value of $\varepsilon$. The score test statistic is conservative, and we can also see that the type-I error gets closer to 5% when we move $\varepsilon$ towards 0.5. The sup score test statistic is still conservative, but has a type-I error closer to the nominal one when compared to the other test statistic.

Table 2 examines, using appropriate ANOVA models, the effect of sample size, number of points and correlation on the size of the three adjusted tests. In Table 2, in the first model fitted we see that the Wald test is anti-conservative for small samples while the other two tests are conservative. For moderate samples the test based on the sup scores seems to perform the best. The models for the number of time points and the correlation $\rho$ in Table 2 suggest that the results hold regardless of the dimension of $\mathbf{Y}$ or the correlation.

To study the power of the tests we simulated as described above with $\gamma = 0.25$ or 0.5 and $\delta = 0$ or 0.5. ANOVA methods on the percent rejections, $R$, are used to summarize the data. Table 3, based on the ANOVA model

$$(\text{Statistic} \times \text{Epsilon}) + \text{Sample size} + \text{Correlation} + \text{Number of points} + \text{True cutpoint} \quad (8)$$

shows that the power of the tests is not affected by the choice of $\varepsilon$. We see in Table 4 that the power increases as the sample size increases and as the dimensionality of $\mathbf{Y}$ increases. Also in Table 4, we see no difference in power for the two true cutpoints. In all tables the adjusted Wald statistic has the highest power. However, since it was anti-conservative this does not suggest that it is the best statistic to use. The new sup score statistic performs quite well and we suggest its use.

Finally, we show some results for the cutpoint estimates. Fig. 1 shows the mean bias for different sample sizes and different cutpoints. The results suggest small bias for all three procedures, with a smaller standard error for the sup score method. We also conducted a simulation study with data generated with a different link function and different values for the $\beta$ in the linear predictor of the mean function and different cutpoints. The conclusions are similar to the ones shown here, so we do not include the results here.

## 5. An application to analysis of survival data based on pseudo-values

We illustrate the use of these techniques using data on 708 patients given an unrelated donor bone marrow transplant for acute leukemia. The question of interest is the effect of the number of CD34 cells in the donor marrow on the incidence of relapse and death in remission. Clinicians would like to know if there is a minimum cell dose beyond which patients have either lower relapse or less death in remission.

To examine the threshold CD34 count question for these two competing risks we use the pseudo-observation method of Klein and Andersen (2005). In this approach for relapse we compute at a grid of time points $\tau_j$ and the pseudo-observations, $Y_{ji}$, are based on the difference between the full sample estimate of the cumulative incidence of relapse and the leave-one-out estimator of the cumulative incidence. This difference or pseudo-observation is used in a generalized estimating equation to study the covariate effects on the relapse cumulative incidence function. In the example we use $j = 0.5$, 1, 2 and 3 years post

transplant and a complementary log–log link function. Note that once the pseudo-observations are computed the censored data competing risks problem has been reduced to a GEE problem. Of course, we have similar pseudo-observations for the death in remission probability.

Fig. 2 shows the value of the three statistics. All statistics are adjusted for the patient's age, Karnofsky performance score and the degree of HLA matching. Here we see that the optimal cutpoint for both relapse and death in remission is at $2.93 \times 10^6$ CD34 cells. Table 5 shows the unadjusted and adjusted tests for the effect of CD34 count. We see that if no adjustment is made for the estimation of the cutpoint that all tests are significant at the 5% level, while the more appropriate adjusted tests suggest that the CD34 cell count is only important for death in remission. In Table 6 we present the parameter estimates which shows that patients with low CD34 counts have significantly more death in remission.

## 6. Discussion

We have proposed a new test statistic that can be used to estimate a threshold value and also to make inference about the regression coefficient associated with the categorized covariate for generalized linear models and generalized estimating equations. We compared our test statistic with the Wald and score tests statistics as well as their corrected versions. Our simulations show clearly that adjustments must be made when making inference for the regression parameter after a cutpoint is selected because the type-I error rates are extremely high when no correction is used. The adjusted Wald test still inflates the type-I error a little while the adjusted score test statistic is more conservative. Our proposed test statistic is conservative, but with type-I error probability closer to the desired one. Also, the method provides good estimates of the true cutpoint when a threshold model is correct.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

Professor Tunes-da-Silva's work was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo – FAPESP, Brazil, grant number 2007/02823-3. Professor Klein's work was supported by a grant from the National Cancer Institute (R01 CA54706-14).

## References

Andersen PK, Hansen MG, Klein JP. Regression analysis of restricted mean survival time based on pseudo-observations. Lifetime Data Analysis. 2004; 10:335–350. [PubMed: 15690989]

Andersen PK, Klein JP, Rosthoj S. Generalized linear models for correlated pseudo-observations, with applications to multi-state models. Biometrika. 2003; 90:15–27.

Andrei A-C, Murray S. Regression models for the mean of the quality-of-life-adjusted restricted survival time using pseudo-observations. Biometrics. 2007; 63:398–404. [PubMed: 17688492]

Billingsley, P. Convergence of Probability Measures. New York: Wiley and Sons; 1968.

Billingsley, P. Convergence of Probability Measures. 2nd ed.. New York: Wiley and Sons; 1999.

Boos D. On generalized score tests. The American Statistician. 1992; 46(4):327–333.

Contal C, O'Quigley J. An application of changepoint methods in studying the effect of age on survival in breast cancer. Computational Statistics and Data Analysis. 1999; 30:253–270.

Graw F, Gerds TA, Schumacher M. On pseudo-values for regression analysis in competing risks models. Lifetime Data Analysis. 2009; 15:241–255. [PubMed: 19051013]

Jespersen NCB. Dichotomizing a continuous covariate in the Cox regression model. Statistical Research Unit of University of Copenhagen, Research Report. 1986; 86(2)

Klein JP, Andersen PK. Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. Biometrics. 2005; 61:223–229. [PubMed: 15737097]

Klein, JP.; Wu, J-T. Discretizing a continuous covariate in survival studies. In: Balakrishnan, N.; Rao, CR., editors. Handbook of Statistics 23: Advances in Survival Analysis. New York: Elsevier; 2004. p. 27-42.

Lausen B, Hothorn T, Bretz F, Schumacher M. Assessment of optimal selected prognostic factors. Biometrical Journal. 2004; 46(3):364–374.

Lausen B, Schumacher M. Maximally selected rank statistics. Biometrics. 1992; 48:73–85.

Lausen B, Schumacher M. Evaluating the effect of optimized cutoff values in the assessment of prognostic factors. Computational Statistics and Data Analysis. 1996; 21:307–326.

Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 78:13–22.

Liu L, Logan B, Klein JP. Inference for current leukemia free survival. Lifetime Data Analysis. 2008; 14:432–446. [PubMed: 18663574]

Logan BR, Klein JP, Zhang MJ. Comparing treatments in the presence of crossing survival curves: an application to bone marrow transplantation. Biometrics. 2008; 64(3):733–740. [PubMed: 18190619]

McCullagh, P.; Nelder, J. Generalized Linear Models. London: Chapman and Hall; 1989.

Miller R, Siegmund D. Maximally selected chi square statistics. Biometrics. 1982; 38:1011–1016.

O'Quigley, J. Proportional Hazards Regression. New York: Springer; 2008.

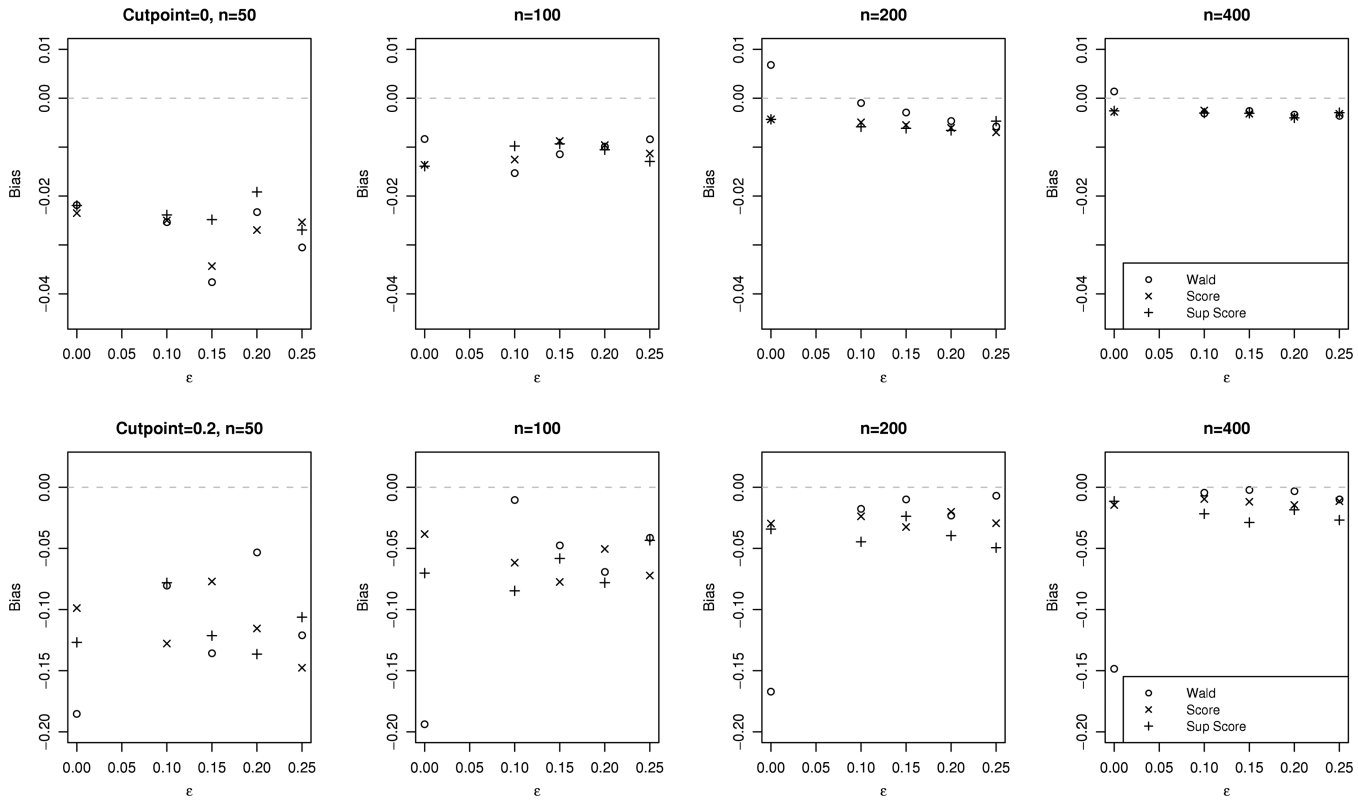


**Fig. 1.**
Mean bias for cutpoint estimates.

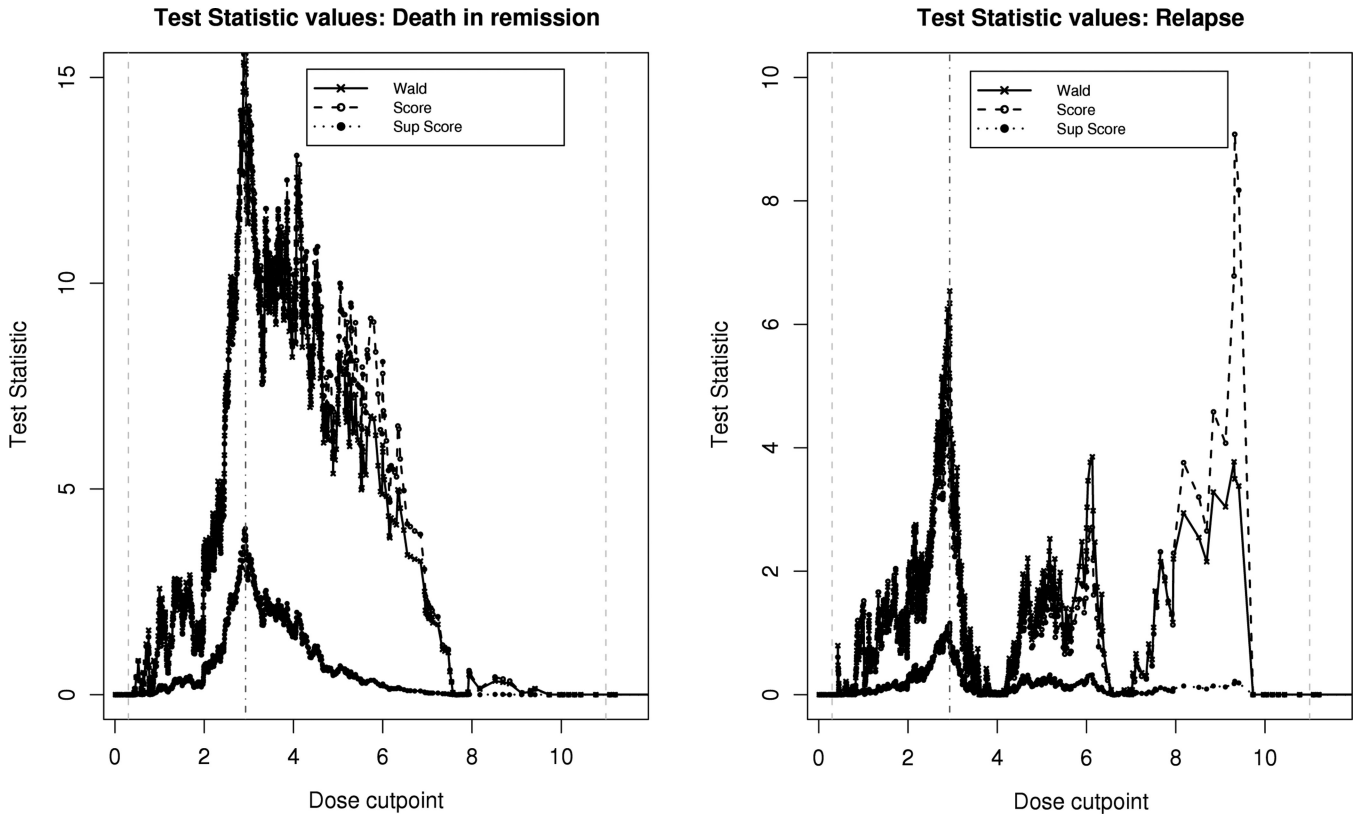


**Fig. 2.**
Test statistic values for the bone marrow data example.

**Table 1**

Average deviations from nominal 5% level for four tests adjusted using the ANOVA model (7).

| Test statistic | Value of **e** | | | | |
|---|---|---|---|---|---|
| | **0 (entire range)** | **0.1** | **0.15** | **0.2** | **0.25** |
| Wald | 94.48 | 44.18 | 36.10 | 29.92 | 24.80 |
| Adjusted Wald | – | 3.17 | 1.46 | 0.64 | 0.28 |
| Score | 44.36 | 35.71 | 30.76 | 26.23 | 22.11 |
| Adjusted score | – | −2.46 | −2.06 | −1.72 | −1.49 |
| Sup score | −1.39 | −1.39 | −1.39 | −1.42 | −1.52 |

*Comput Stat Data Anal*. Author manuscript; available in PMC 2012 May 23.

**Table 2**

Average deviations from nominal 5% level for three tests adjusted using ANOVA models.

| Factor | | Test statistic | | |
|---|---|---|---|---|
| | | **Adjusted Wald** | **Adjusted score** | **Sup score** |
| Sample size | 50 | 4.01 | −3.07 | −2.14 |
| | 100 | 1.36 | −2.17 | −1.64 |
| | 200 | 0.31 | −1.46 | −1.12 |
| | 400 | −0.13 | −1.04 | −0.82 |
| Number of time points | 2 | 1.36 | −1.96 | −1.44 |
| | 4 | 1.41 | −1.91 | −1.42 |
| Correlation ρ | 0.00 | 1.42 | −1.87 | −1.37 |
| | 0.25 | 1.29 | −2.01 | −1.52 |
| | 0.50 | 1.49 | −1.94 | −1.41 |
| | 0.75 | 1.34 | −1.92 | −1.43 |

**Table 3**

Power results for three tests adjusted using the ANOVA model (8).

| True β | Test statistic | Value of ε | | | | |
|---|---|---|---|---|---|---|
| | | 0 (entire range) | 0.1 | 0.15 | 0.2 | 0.25 |
| 0.25 | Adjusted Wald | – | 43.26 | 43.29 | 44.01 | 44.85 |
| | Adjusted score | – | 34.42 | 36.66 | 38.67 | 40.47 |
| | Sup score | 41.98 | 41.98 | 41.97 | 41.94 | 41.79 |
| 0.50 | Adjusted Wald | – | 80.51 | 80.72 | 81.24 | 81.81 |
| | Adjusted score | – | 72.37 | 74.27 | 75.91 | 77.40 |
| | Sup score | 78.77 | 78.77 | 78.77 | 78.76 | 78.69 |

*Comput Stat Data Anal*. Author manuscript; available in PMC 2012 May 23.

**Table 4**

Power results for three tests adjusted using the ANOVA models.

| Effect | | True β = 0.25 | | | True β = 0.50 | | |
|---|---|---|---|---|---|---|---|
| | | Adjusted Wald | Adjusted score | Sup score | Adjusted Wald | Adjusted score | Sup score |
| Sample size | 50 | 20.45 | 8.57 | 12.06 | 51.87 | 34.88 | 42.89 |
| | 100 | 29.05 | 21.39 | 26.04 | 76.70 | 70.27 | 75.76 |
| | 200 | 49.21 | 45.04 | 50.29 | 95.80 | 94.91 | 96.40 |
| | 400 | 76.70 | 75.23 | 79.29 | 99.90 | 82.91 | 90.92 |
| Number of time points | 2 | 39.10 | 32.95 | 37.25 | 78.00 | 71.41 | 75.47 |
| | 4 | 48.61 | 42.16 | 46.59 | 84.14 | 78.57 | 82.03 |
| True cutpoint | 0.0 | 44.47 | 38.31 | 43.08 | 81.54 | 75.72 | 79.63 |
| | 0.2 | 43.23 | 36.80 | 40.75 | 80.60 | 74.26 | 77.86 |
| Correlation ρ | 0 | 57.49 | 50.82 | 55.27 | 90.41 | 85.65 | 88.57 |
| | 0.25 | 46.59 | 40.00 | 44.60 | 83.79 | 77.68 | 81.42 |
| | 0.5 | 38.34 | 32.25 | 36.62 | 77.72 | 71.08 | 75.17 |
| | 0.75 | 32.98 | 27.15 | 31.19 | 72.36 | 65.55 | 69.83 |

**Table 5**

Cutpoints for the bone marrow data example, model for death in remission and relapse with covariates age, group and KPS.

| Outcome | Test statistic | Cutpoint selected | Value of test statistic | Unadjusted *P*-value | Adjusted *P*-value |
|---|---|---|---|---|---|
| | Wald | 2.93 | 16.40 | < 0.0001 | 0.0019 |
| Death in remission | Score | 2.93 | 16.75 | < 0.0001 | 0.0016 |
| | Sup score | 2.93 | 3.88 | – | 0.0006 |
| | Wald | 2.93 | 6.54 | 0.0105 | 0.1680 |
| Relapse | Score | 2.93 | 5.11 | 0.0237 | 0.3017 |
| | Sup score | 2.93 | 1.16 | – | 0.1962 |

**Table 6**

Parameters estimates for the bone marrow data example, death in remission and relapse.

| Parameter | | *N* | Death in remission | | | | Relapse | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Estimative | Standard error | Test statistic | *P*-value | Estimative | Standard error | Test statistic | *P*-value |
| Intercept | | | −2.18 | 0.27 | 66.26[a] | <0.0001 | −0.63 | 0.22 | 8.35[a] | 0.0039 |
| | | | | | 9.69[b] | 0.0079 | | | 1.81[b] | 0.4037 |
| Match | 8/8 | 442 | – | – | – | – | – | – | – | – |
| | 7/8 | 234 | 0.13 | 0.14 | 0.86[a] | 0.3502 | −0.04 | 0.17 | 0.06[a] | 0.8046 |
| | 6/8 | 113 | 0.51 | 0.17 | 9.67[a] | 0.0019 | −0.33 | 0.25 | 1.82[a] | 0.1786 |
| | | | | | 41.85[c] | <0.0001 | | | 13.35[c] | 0.0203 |
| Age | <10 years | 170 | – | – | – | – | – | – | – | – |
| | 10–19 years | 161 | 1.37 | 0.26 | 26.83[a] | <0.0001 | −0.36 | 0.23 | 2.53[a] | 0.1123 |
| | 20–29 years | 154 | 1.18 | 0.27 | 19.18[a] | <0.0001 | −0.29 | 0.23 | 1.64[a] | 0.2004 |
| | 30–39 years | 102 | 1.45 | 0.28 | 27.67[a] | <0.0001 | −0.01 | 0.25 | 0.00[a] | 0.9568 |
| | 40–49 years | 146 | 1.64 | 0.26 | 39.44[a] | <0.0001 | −0.68 | 0.27 | 6.25[a] | 0.0123 |
| | ≥ 50 years | 56 | 1.39 | 0.31 | 20.25[a] | <0.0001 | −1.07 | 0.40 | 6.97[a] | 0.0082 |
| | | | | | 3.01[b] | 0.2220 | | | 2.22[b] | 0.3928 |
| Karnofsky score | 0–90 | 162 | – | – | – | – | – | – | – | – |
| | > 90 | 579 | −0.17 | 0.15 | 1.35[a] | 0.2449 | −0.26 | 0.18 | 2.04[a] | 0.1524 |
| | Missing | 48 | −0.46 | 0.28 | 2.69[a] | 0.1009 | −0.07 | 0.34 | 0.04[a] | 0.8472 |
| Dose | > 2.93 | 407 | – | – | – | – | – | – | – | – |
| | ≤ 2.93 | 382 | 0.47 | – | 3.88[d] | 0.0006 | −0.43 | – | 1.16[d] | 0.1962 |

[a] $\chi^2$ distributed, 1 degree of freedom.

[b] $\chi^2$ distributed, 2 degrees of freedom.

[c] $\chi^2$ distributed, 5 degrees of freedom.

[d] Sup score test.