The International Journal of Biostatistics

Volume 5, *Issue* 1 2009 *Article* 7

A Non-Parametric Approach to Scale Reduction for Uni-Dimensional Screening Scales

Xinhua Liu, *Columbia University*
Zhezhen Jin, *Columbia University*

Recommended Citation:
Liu, Xinhua and Jin, Zhezhen (2009) "A Non-Parametric Approach to Scale Reduction for Uni-Dimensional Screening Scales," *The International Journal of Biostatistics*: Vol. 5: Iss. 1, Article 7.
DOI: 10.2202/1557-4679.1094

A Non-Parametric Approach to Scale Reduction for Uni-Dimensional Screening Scales

Xinhua Liu and Zhezhen Jin

Abstract

To select items from a uni-dimensional scale to create a reduced scale for disease screening, Liu and Jin (2007) developed a non-parametric method based on binary risk classification. When the measure for the risk of a disease is ordinal or quantitative, and possibly subject to random censoring, this method is inefficient because it requires dichotomizing the risk measure, which may cause information loss and sample size reduction. In this paper, we modify Harrell's C-index (1984) such that the concordance probability, used as a measure of the discrimination accuracy of a scale with integer valued scores, can be estimated consistently when data are subject to random censoring. By evaluating changes in discrimination accuracy with the addition or deletion of items, we can select risk-related items without specifying parametric models. The procedure first removes the least useful items from the full scale, then, applies forward stepwise selection to the remaining items to obtain a reduced scale whose discrimination accuracy matches or exceeds that of the full scale. A simulation study shows the procedure to have good finite sample performance. We illustrate the method using a data set of patients at risk of developing Alzheimer's disease, who were administered a 40-item test of olfactory function before their semi-annual follow-up assessment.

KEYWORDS: discrimination accuracy, item selection, reduced scale, risk, test score

Author Notes: Research for this paper was partially supported by the NSF Career award DMS-0134431 to Zhezhen Jin. We thank Drs. Devanand and Tabert for providing us data from their prospective study conducted at the Memory Disorders Center of the New York State Psychiatric Institute and Columbia University, funded by grant R01-AG17761 from the National Institute on Aging. We also thank Dr. Wenbin Lu for providing a program of adaptive LASSO with Cox proportional hazards model. We are grateful to the reviewers for their valuable comments and suggestions.

Introduction

In biomedical studies, uni-dimensional scales composed of a set of test items are often used to assess a latent trait or function that correlates to the item responses (Lord and Novick, 1968). Usually, the scale weighs all the items equally and a sum of item responses is used as a scale score to measure the latent trait or function. For example, the standardized University of Pennsylvania Smell Identification Test (UPSIT), used to measure olfactory function (Doty et al., 1984), contains 40 odor items, and is a uni-dimensional scale with binary item responses. The score of the self-administered test equals the total number of odors correctly identified.

When a latent variable is predictive of the development of a disease, a scale measuring the latent variable may be used for screening. However, when some items on the scale are redundant or not relevant to predicting disease development, reduction of the size of the scale (item reduction) is warranted. To reduce the screening costs, clinicians wish to use a reduced scale that screens patients more efficiently than the full scale. To illustrate this point, we consider the following example. It has been reported that an increased risk of developing Alzheimer's disease (AD) is associated with olfactory deficit (Doty et al., 1987; Devanand et al., 2000). The 40-item test (UPSIT) measuring olfactory function takes approximate 30 minutes to complete. To improve the clinical utility of UPSIT, researchers attempted to select items with high predictive ability from the full scale (Tabert et al., 2005). For screening purposes, the number of correctly recognized odors is more meaningful, as a scale score, than any weighted sum of the odor item responses. This implies that a reduced scale for screening should also be uni-dimensional, that is, that its items should be weighted equally. Consequently, the range of scores of a reduced scale will be narrower than that of the full scale.

Selection of risk-related items from a uni-dimensional scale is challenging, because the scale items are all positively related to the same latent variable. The available variable selection procedures based on regression models for a risk measure outcome treat each item as an independent variable. Also, without constraints on the model parameters, the estimated coefficients for selected items may have opposite signs, violating the necessary condition for creating a reduced uni-dimensional scale. Liu and Jin (2007) offers a detailed discussion on selection of binary (high vs. low) risk-related items from a uni-dimensional scale to create a reduced scale.

To quantify a scale's ability to discriminate between levels of risk, there are several measures. For binary measure of risk, if one specifies a

parametric model for the probability of being in the high risk class with the item responses used as predictors, then the available classification accuracy measures, such as classification error rate (Hastie et al., 2001) or Brier score (Brier, 1950), can be defined as functions of the discrepancies between risk observations and model based estimates. The classification accuracy measure, CA, used by Liu and Jin (2007) is the probability that a subject randomly selected from the high risk class has a higher (or lower) scale score than a subject randomly selected from the low risk class. This measure is similar to Hanley and McNeil's (1982) interpretation of the area under the receiver operating characteristic (ROC) curve for a continuous variable, where the ROC curve is defined through sensitivity and specificity at each of the scale's possible cutoff points (Zhou et al., 2002; Pepe, 2003). Since CA can be estimated non-parametrically, Liu and Jin (2007) proposed an item selection method which, without specification of a parametric regression model, evaluates the change in CA when deleting or adding an item to a reduced scale, and thus monitors the process of selecting items for a reduced uni-dimensional scale.

In a longitudinal study of patients at risk for a disease, the time from baseline assessment to first diagnosis can be used as a measure of risk. Often, the observed time is censored due to study termination or subject dropout. In a prospective study of patients with mild cognitive impairment (MCI) who are at risk of developing AD (Tabert et al., 2006), for example, the observed time to AD conversion in many of the 128 patients who were administered the olfactory test UPSIT at baseline was censored. Only 38 patients were found to have converted to AD at follow up. If conversion to AD within two years from the baseline assessment is the criterion used to define the high and low risk groups, then the patients who did not meet diagnosis criteria, or who dropped out from the study before completing the follow-up assessment at two years, cannot be classified into either class. Thus, data on these patients cannot be used by the item reduction method for binary risk classes. In summary, for quantitative risk measures, the binary classification-based item reduction method has limitations due to dichotomization of the risk measure: it overlooks quantitative information on the measure of risk, and reduces the sample size by excluding censored data.

For risk discrimination in the context of survival analysis, Harrell et al. (1984) proposed the index C to estimate the probability of concordance between the predicted risk of event occurrence and the observed time to either the event or the end of the study. The concordance probability defined for a pair of bivariate observations is often used to assess the discriminatory power of a statistical model (Harrell et al., 1996). Related to Somers' d rank

DOI: 10.2202/1557-4679.1094

correlation (Somers, 1962), the concordance probability is also an extension of the area under the ROC curve for continuous variables used for binary classifications. In this light, Pencina and D'Agostino (2004) discussed the relationship between the C-index and the modified Kendall's τ for bivariate correlation (Kendall, 1970). Along with the interpretation of Harrell's C-index, Antolini et al. (2005) derived a time-dependent discrimination index for survival data. Other work related to time-dependent ROC includes the papers by Heagerty et al. (2000); Heagerty and Zheng (2005); Chambless and Diao (2006); and Zheng et al. (2006).

After deriving an analytical expression for the concordance probability in the Cox proportional hazards model, Gönen and Heller (2005) proposed an asymptotically unbiased estimator of the concordance probability as a function of the regression parameters and the covariate distribution. However, the variable selection procedures based on Cox proportional hazards models, including LASSO and adaptive LASSO, cannot select items for a reduced uni-dimensional scale that has equally weighted risk-related items.

In this paper, we modify Harrell's C-index to obtain a consistent estimator of concordance probability, which can be used to assess the discrimination accuracy of a uni-dimensional scale. The proposed estimator takes into account possible random censoring when the risk of disease is measured using the time between a patient's baseline scale-based functional assessment, and the first diagnosis of the disease during follow-up. To develop a reduced uni-dimensional scale useful for risk determination, we evaluate the changes in discrimination accuracy that result from the addition or removal of items from the scale. After investigating the finite sample performance of the proposed procedure in a simulation study, we illustrate the method using data from the study by Tabert et al. (2006) in which 128 patients at risk of developing AD were administered the UPSIT to assess olfactory functioning, and then followed semi-annually for up to nine years to identify incident cases of AD.

Method

Suppose that a full scale has m items in the set W_m and the response on each item is binary, i.e. $X_h \in \{0, 1\}$ for item $h,\ h = 1, \cdots, m$, then the score on the full scale, defined as $S(W_m) = \sum_{h=1}^{m} X_h$, takes integer values between zero and m. For a set W_k with k items, $k = 1, \cdots\ m$; let X_{ih} be the ith subject's binary response on item h in $W_k,\ h = 1, \cdots\ k$; and $S_i(W_k)$ be the ith subject's score for the scale with item set $W_k,\ i = 1, \cdots,\ n$. Let T_i be the length of time between baseline assessment and diagnosis of the disease

for subject i during follow-up. Suppose that a subject who has a higher score tends to develop the disease after a longer period of time or is at a lower risk. We define the discrimination accuracy of the scale consisting of the items in W_k as a conditional probability,

$$DA(W_k) = P(S_i(W_k) < S_j(W_k)|T_i < T_j).$$

The quantity takes values between zero and one. Obviously, $DA(W_k)$ can be applied to a case where the T is a risk measure with a fixed number of ordinal categories. When there are only two risk classes such that with a constant Q, $T_i < Q$ for all subjects in one class and $Q < T_j$ for all subjects in another class, $DA(W_k)$ reduces to the classification accuracy $CA(W_k)$ (Liu and Jin, 2007). Similar to $CA(W_k)$, $DA(W_k)$ retains the invariance property that it remains unchanged with a rank-preserving transformation of the score $S(W_k)$ or the measure of risk. Notice that the change in DA resulting from the addition or deletion of an item may indicate the relative importance of the item to risk discrimination. The estimation of $DA(W_k)$, however, is not straightforward when T is subject to censoring. In the next section, we present a consistent estimator of $DA(W_k)$, along with a simple approach for evaluating the change in the discrimination accuracy that results from adding or removing an item from the item set W_k.

A. Assessment of change in discrimination accuracy

In presence of censoring, i.e. where subject i does not have the disease at the last follow-up time Q_i, the observed time $Y_i = T_i\ d_i + Q_i\ (1 - d_i)$ with $d_i = I(T_i < Q_i)$, where $I(\cdot)$ is an indicator function taking values of 0 or 1. For a given item set W_k, to estimate $DA(W_k)$ with n independent observations $(Y_i,\ d_i,\ S_i(W_k)),\ i = 1, \cdots,\ n$; Harrell's C-index has the form,

$$C(W_k) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} d_i I(Y_i < Y_j) I(S_i(W_k) < S_j(W_k))}{\sum_{i=1}^{n}\sum_{j=1}^{n} d_i I(Y_i < Y_j)}.$$

When Q_i is a constant or $Y_i = T_i$ for all i, then $C(W_k)$ converges to $DA(W_k)$. However, if censoring variable Q is random and independent of variable T, then $C(W_k)$ will converge to $P(S_i(W_k) < S_j(W_k)|T_i < T_j, T_i < Q_i, T_i < Q_j)$, a quantity depending on censoring pattern.

To obtain a consistent estimator for $DA(W_k)$, we may modify $C(W_k)$ by replacing d_i with $b_i = d_i/G^2(y_i)$, where $G(t) = P(t < Q)$ for $t > 0$. In the case that $G(t)$ is unknown, a consistent estimator $\hat{G}(t)$, constructed by

DOI: 10.2202/1557-4679.1094

the Kaplan-Meier product limit method, may be used. Therefore,

$$\widehat{DA}(W_k) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} b_i I(Y_i < Y_j) I(S_i(W_k) < S_j(W_k))}{\sum_{i=1}^{n}\sum_{j=1}^{n} b_i I(Y_i < Y_j)}.$$

For $y_{(n)} = \max_{1\le i\le n} y_i$, we set $b_{(n)} = 0$ if $d_{(n)} = 0$ and $\hat{G}(y_{(n)}) = 0$. Obviously, when Q_i is constant or $Y_i = T_i$ for all i, then the estimator reduces to $C(W_k)$.

Because $\widehat{DA}(W_k)$ is proportional to the quantity

$$A(W_k) = \sum_{i=1}^{n}\sum_{j=1}^{n} b_i I(Y_i < Y_j) I(S_i(W_k) < S_j(W_k)),$$

the change in $\widehat{DA}(W_k)$ will also be proportional to the change in $A(W_k)$.

The change in $A(W_k)$ due to excluding $item_h$ from the item set W_k can be written as

$$\Delta A_k(-X_h|W_k) = A(W_k) - A(W_k \setminus \{item_h\}),$$

and the change due to adding $item_h$ into W_{k-1} for a new set W_k,

$$\Delta A_k(+X_h|W_{k-1}) = A(W_{k-1} \cup \{item_h\}) - A(W_{k-1}).$$

Let $e_{ij}(k) = S_i(W_k) - S_j(W_k)$ and $z_{ij}(h) = X_{ih} - X_{jh}$. Then we will have $z_{ij}(h) \in \{-1,\ 0,\ 1\}$. Because $S_i(W_k) = S_i(W_k \setminus \{item_h\}) + X_{ih},\ 1 \le k \le m$, we may write

$$\Delta A_k(-X_h|W_k) = \sum_{i=1}^{n}\sum_{j=1}^{n} U_{-h}(i,j|W_k),$$

where $U_{-h}(i,j|W_k) = b_i I(Y_i \le Y_j)\eta_{ij}^{-}(h,k)$ with

$$\eta_{ij}^{-}(h,k) = I(z_{ij}(h) = -1,\ e_{ij}(k) = -1) - I(z_{ij}(h) = 1,\ e_{ij}(k) = 0).$$

Similarly, we may have

$$\Delta A_k(+X_h|W_{k-1}) = \sum_{i=1}^{n}\sum_{j=1}^{n} U_{+h}(i,j|W_{k-1}),$$

where $U_{+h}(i,j|W_{k-1}) = b_i I(Y_i \le Y_j)\eta_{ij}^{+}(h,\ k-1)$ with

$$\eta_{ij}^{+}(h,\ k-1) = I(z_{ij}(h) = -1,\ e_{ij}(k-1) = 0) - I(z_{ij}(h) = 1,\ e_{ij}(k-1) = -1).$$

In summary, we can write the changes $\Delta A_k(-X_h|W_k)$ and $\Delta A_k(+X_h|W_{k-1})$ in the form

$$\Delta A = \sum_{i=1}^{n}\sum_{j=1}^{n} U_{ij},$$

where U_{ij} will be $U_{-h}(i,j|W_k)$ or $U_{+h}(i,j|W_{k-1})$, accordingly. Under some regularity conditions,

$$\sqrt{n}\ (\Delta A/n^2 - \mu) \to N(0,\ \phi), \quad as\ n \to \infty,$$

where $\mu = E(\Delta A/n^2)$ and ϕ is the limiting variance. The justification is given in the Appendix.

Let $\delta_k = DA(W_k) - DA(W_{k-1})$ with $W_{k-1} \subset W_k$. Since $E(\Delta A_k)$ and δ_k share the same sign and $E(\Delta A_k) = 0$ implies $\delta_k = 0$, we may use ΔA_k to construct a statistic for testing the null hypothesis $H_0 : \delta_k = 0$ that

$$TS = \frac{\Delta A_k}{\hat{se}(\Delta A_k)}.$$

As the Wald type test statistic TS has approximate $N(0,1)$ distribution under the null hypothesis, we propose to use it to guide the risk related item selection. The relevant hypotheses to test are $H_0 : \delta_k \leq 0$ (no improvement in DA) vs. $H_1 : \delta_k > 0$ (DA improved). Specifically, we will use a preset threshold value γ_0 and the test statistic

$$TS(-X_j|W_k) = \frac{\Delta A_k(-X_j|W_k)}{\hat{se}(\Delta A_k(-X_j|W_k))}$$

to decide whether or not to remove $item_j$ from W_k. We will exclude $item_j$ from W_k when $TS(-X_j|W_k) < \gamma_0$. Similarly, we will use a preset threshold value γ_1 and the test statistic

$$TS(+X_h|W_{k-1}) = \frac{\Delta A_k(+X_h|W_{k-1})}{\hat{se}(\Delta A_k(+X_h|W_{k-1}))}$$

to decide whether or not to add $item_h \in W_m \setminus W_{k-1}$ into W_{k-1} for a new set W_k. We will have item set $W_k = \{item_h\} \cup W_{k-1}$ when $TS(+X_h|W_{k-1}) \geq \gamma_1$.

Noting that the test statistic for detecting changes in DA retains the properties of the statistic for detecting the changes in CA, we may use the strategies for reduction of binary risk related items to select the items that are related to an ordinal or a continuous risk measure, possibly subject to random censoring.

B. Item selection procedure

It is obvious that $item_j$ in W_k for a scale with score $S(W_k)$ is not useful in risk discrimination, if excluding it from W_k leads to either no change or an

DOI: 10.2202/1557-4679.1094

increase in the estimated discrimination accuracy. Therefore, we first identify redundant items in the full scale, if any, and then apply a hypothesis test based stepwise selection procedure, to the remaining items.

Starting with the item set W_m of the full scale, we will identify the redundant items, if any, and remove them. For $1 < k \leq m$, we will exclude $item_h$ from W_k if the corresponding change $\Delta A_k(-X_h|W_k) \leq 0$, where

$$\Delta A_k(-X_h|W_k) = \min_{item_j \in W_k} \{A(W_k) - A(W_k \setminus \{item_j\})\}.$$

The deletion process will stop when no more items can be removed. The resulting item set is denoted as $W_J,\ 1 < J < m$. This process will produce a sequence of subsets $\{W_k;\ J \leq k < m\}$ with a sequence of estimated discrimination accuracies $\{\widehat{DA}(W_k);\ J \leq k < m\}$ satisfying $\widehat{DA}(W_J) \geq \cdots \geq \widehat{DA}(W_m)$.

Although the item set W_J has fewer items, it might still have some unstable items that contribute little to discrimination accuracy. It is important to identify relatively stable items in W_J to form a further reduced scale without substantially sacrificing discrimination accuracy. This can be accomplished by the following hypothesis test based selection procedure along with preset positive threshold values γ_0 and γ_1 $(\gamma_0 \leq \gamma_1)$:

(i) Identify the item in W_J that has the largest estimated discrimination accuracy. Let Ω_1 denote the resulting singleton item set.

(ii) For $1 < k \leq J$, identify the item $item_h$ that has the largest value of the test statistic for $H_0 : \delta_k \leq 0$ vs. $H_1 : \delta_k > 0$ from $W_J \setminus \Omega_{k-1}$. Let $\Omega_k = \Omega_{k-1} \cup \{item_h\}$ if

$$TS(+X_h|\Omega_{k-1}) = \max_{item_j \in W_J \setminus \Omega_{k-1}} TS(+X_j|\Omega_{k-1}) \geq \gamma_1.$$

(iii) Identify the unstable items in Ω_k whose removal leads to little loss or even an improvement in the estimated discrimination accuracy. Specifically, $item_h$ is excluded from $\Omega_k,\ 1 < k \leq J$; if

$$TS(-X_h|\Omega_k) = \min_{item_j \in \Omega_k} TS(-X_j|\Omega_k) < \gamma_0.$$

The exclusion process will stop if no more items can be removed.

(iv) Repeat steps (ii) and (iii) until no more items can be added or removed, or stop the process if an item that has been removed tends to be added again.

The final item set, denoted as Ω_H, will have a set of items appropriate for a reduced scale.

To assess variations in item selection, we may use the bootstrap method (Efron and Tibshirani, 1993). Bootstrap samples can be obtained by sampling with replacement from the original study sample, where the sampling unit is the study subject with a cluster of observed responses to the items on the full scale and the measure of level of risk (such as a risk measure with ordinal categories or observed time to the initial diagnosis of disease, with a censoring indicator if applicable). The empirical distributions of the number of selected items, the estimated discrimination accuracy for the full scale and the reduced scale, as well as the improvement in the estimated discrimination accuracy of the reduced scale over that of the full scale, can be used for inference. Moreover, the selection frequency of each item in a number of bootstrap samples (say 1000) provides an empirical estimate of how often an item is selected. The resulting item spectrum may help identify the most frequently selected items.

A simulation study

To examine the finite sample performance of the selection procedure using different thresholds, we conducted a simulation study for a hypothetical unidimensional scale with 13 items, among which items $\{1, \cdots, 6\}$ are useful for risk discrimination. The sample size $N = 120$, and 240 along with censoring proportions of 50% and 75% were used. In each of the four cases, we generated 1000 data sets. In each data set, we first generated N independent random numbers from exponential distribution with mean of 5 for time variable T, and N independent random numbers from uniform distribution $U(0, \theta)$ for censoring variable Q with θ specified according to the preset censoring proportion. We then calculated the actual time variable $Y = \min(T,\ Q)$ and indicator $d = I(T < Q)$ for the observed event. For item response data, we first generated N independent random numbers from a standard normal distribution for variable Z. Assuming that the latent variable is a function of T and Z, we then generated 13 independent binary responses with probabilities specified by the logistic models for each value of (T, Z),

$$\text{logit P}(X_j = 1|\ T,\ Z) = \alpha_j(T) + \beta_j(T)Z, \quad j = 1, \cdots, 13.$$

The preset values or functions for $\alpha_j(T)$ and $\beta_j(T)$ are listed in Table 1.

DOI: 10.2202/1557-4679.1094

Table 1. Parameters of the logistic models used for data generation

Item j	1	2	3	4	5	6	7	8	9	10	11	12	13
α_j	$a_1(T)$	$a_1(T)$	$a_1(T)$	$a_2(T)$	$a_2(T)$	$a_2(T)$	-1	-0.5	-0.5	0	0.5	0.5	1
β_j	1	$b_1(T)$	$b_2(T)$	1	$b_1(T)$	$b_2(T)$	1	1	2	1	1	2	1

$a_1(T) = -1.5 + 0.4T, \;\; a_2(T) = -1 + 0.3T;$
$b_1(T) = 1 + 0.5I(T < 5), \; b_2(T) = 1 + I(T < 5).$

In each of the four cases, we applied the proposed procedure to the 1000 generated data sets with four sets of threshold values $(\gamma_0, \gamma_1) = (0.524,$ $0.5244)$, $(0.841,\ 0.8416)$, $(1.036,\ 1.0364)$, $(1.281, 1.2816)$ according to 70th, 80th, 85th and 90th percentiles of standard normal distribution, respectively.

Table 2 shows that in each case, the mean discrimination accuracy of the full scale with W_{13} is lower than that of the "true" scale with $W = \{1, \cdots,\ 6\}$ and those of the item sets selected by the four criteria. As expected, increased threshold values result in fewer selected items, lower discrimination accuracies and less improvement in all the four cases. The mean scale size, mean discrimination accuracy and mean improvement, however, vary least with threshold values in the case with the larger sample size ($N = 240$) and uncensored proportion (50%). For a given set of threshold values, averaged scale size and percent of positive improvement increase with sample size and with proportions of uncensored subjects.

Table 3 lists the frequencies of items selected based on different selection criteria. It is interesting to note that in all the cases, the most frequently selected items are the six "true" items. The numbers of correctly selected items increase when the threshold is lowered or when the uncensored proportion or the sample size increases. In contrast, the numbers of incorrectly selected items decrease with increasing threshold, uncensored proportion, and sample size.

It is noticeable in Tables 2 and 3 that the impact of threshold values on the selection of risk related items is smaller in the case of larger sample size ($n = 240$) with lower censored proportion (50%) than in the case of smaller sample size ($n = 120$) with higher censored proportion (75%).

Table 2. Performance of selected scales

Sample size (% Censored)	Criteria	Scale size H Mean (SD)	$DA(W_H)$ Mean (SD)	ΔDA (%) Mean (SD)	$\Delta DA > 0$ %
N=120	Full scale	13	0.5990 (0.0638)		
(75%)	Scale(W)	6	0.6617 (0.0588)	10.88 (6.77)	98.4
	I	4.85 (1.01)	0.6731 (0.0564)	12.87 (7.44)	99.6
	II	4.41 (0.94)	0.6693 (0.0564)	12.23 (7.46)	99.3
	III	4.20 (0.90)	0.6669 (0.0569)	11.83 (7.52)	99.0
	IV	3.87 (0.86)	0.6619 (0.0576)	10.96 (7.53)	98.3
N=120	Full scale	13	0.5908 (0.0372)		
(50%)	Scale(W)	6	0.6539 (0.0351)	10.87 (4.33)	99.7
	I	5.56 (0.86)	0.6574 (0.0353)	11.42 (4.17)	100
	II	5.10 (0.85)	0.6542 (0.0359)	10.87 (4.31)	100
	III	4.85 (0.82)	0.6517 (0.0364)	10.45 (4.35)	100
	IV	4.57 (0.80)	0.6481 (0.0370)	9.84 (4.43)	99.8
N=240	Full scale	13	0.5950 (0.0459)		
(75%)	Scale(W)	6	0.6595 (0.0430)	11.04 (4.91)	99.5
	I	5.43 (0.94)	0.6640 (0.0419)	11.81 (4.42)	100
	II	5.11 (0.89)	0.6625 (0.0415)	11.55 (4.50)	100
	III	4.93 (0.88)	0.6611 (0.0417)	11.31 (4.54)	100
	IV	4.67 (0.86)	0.6586 (0.0425)	10.89 (4.59)	99.9
N=240	Full scale	13	0.5879 (0.0266)		
(50%)	Scale(W)	6	0.6523 (0.0251)	11.03 (3.13)	100
	I	5.83 (0.59)	0.6530 (0.0249)	11.17 (3.02)	100
	II	5.63 (0.62)	0.6522 (0.0251)	11.02 (3.07)	100
	III	5.49 (0.64)	0.6513 (0.0253)	10.87 (3.10)	100
	IV	5.29 (0.67)	0.6496 (0.0258)	10.58 (3.15)	100

$\Delta DA = \frac{(DA(W_H) - DA(W_{13}))}{DA(W_{13})} \times 100\%$: per cent improvement.
$W = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$.
Criteria (γ_0, γ_1): I=(0.524, 0.5244), II=(0.841, 0.8416), III=(1.036, 1.0364), IV=(1.281, 1.2816).

In summary, the proposed procedure showed satisfactory finite sample performance in the simulation study. Empirically, the threshold values can be in the range of 0.8 to 1.3. However, increasing the sample size and the uncensored proportion can reduce the impact of threshold values and allow more stable results to be obtained.

DOI: 10.2202/1557-4679.1094

Table 3. Frequencies of selection of specific items in 1000 simulated data sets

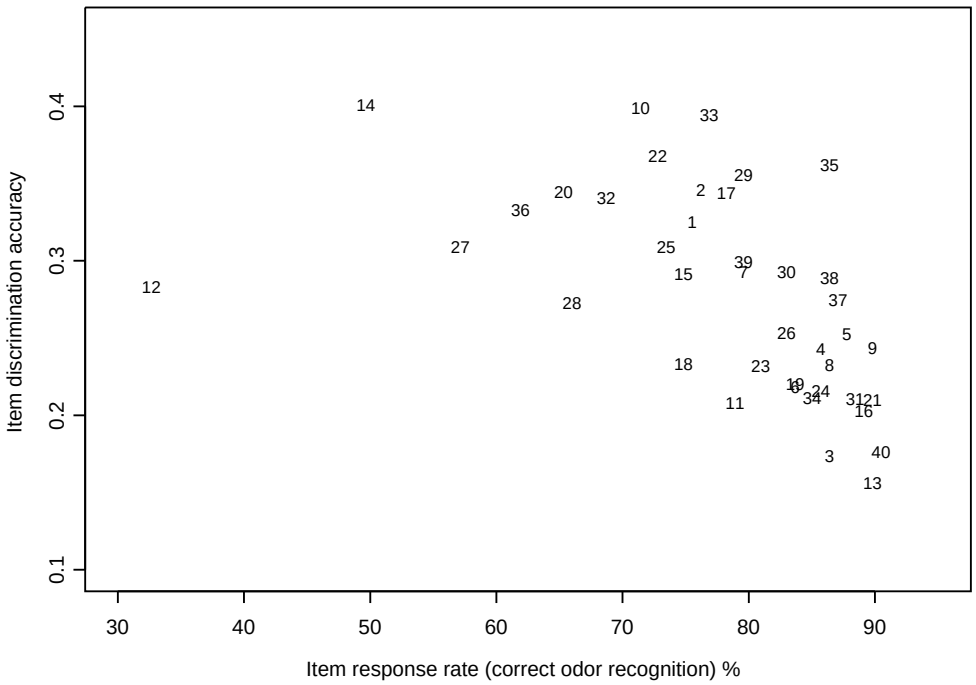
Criteria (γ_0, γ_1)	1	2	3	4	5	6	7	8	9	10	11	12	13
$N = 120$, 75% censored													
I: (0.524, 0.5244)	894	841	745	726	644	536	88	70	29	64	91	21	97
II: (0.841, 0.8416)	865	795	686	675	576	470	72	50	19	50	63	14	71
III: (1.036, 1.0364)	856	768	661	651	555	430	58	41	15	40	53	12	63
IV: (1.281, 1.2816)	830	718	610	615	506	385	43	31	9	31	37	6	45
$N = 120$, 50% censored													
I: (0.524, 0.5244)	983	959	915	922	808	703	55	33	7	46	51	17	56
II: (0.841, 0.8416)	972	918	871	877	713	601	35	14	6	22	25	11	34
III: (1.036, 1.0364)	958	892	841	843	661	555	22	12	4	11	21	9	22
IV: (1.281, 1.2816)	946	864	800	793	602	493	17	9	1	7	16	5	17
$N = 240$, 75% censored													
I: (0.524, 0.5244)	969	946	884	883	768	687	65	50	11	40	49	10	64
II: (0.841, 0.8416)	956	927	845	858	719	608	45	38	6	31	31	4	45
III: (1.036, 1.0364)	953	909	820	824	682	573	41	28	5	25	25	3	38
IV: (1.28, 1.2816)	943	891	778	796	620	523	26	18	5	18	19	2	27
$N = 240$, 50% censored													
I: (0.524, 0.5244)	998	992	987	979	929	835	26	22	3	11	17	3	23
II: (0.841, 0.8416)	997	987	979	966	881	762	13	13	2	7	11	2	11
III: (1.036, 1.0364)	997	983	963	950	852	706	7	8	2	5	9	2	8
IV: (1.281, 1.2816)	994	972	940	927	799	629	4	4	1	4	4	1	6

Application

To illustrate the proposed method, we used the olfaction-test data collected from the patients at risk of developing AD in a prospective study (Tabert et al., 2006). There were 128 patients aged 55 and older with mild cognitive impairment (MCI) who were administered UPSIT at baseline assessment and then followed semi-annually for up to nine years. During the follow-ups, 38 of them met criteria for AD diagnosis with the time to AD conversion varying between 6 months and 5.5 years. The censored proportion of the sample was about 70%. The risk factors of AD such as baseline age, UPSIT score and Mini-Mental State Examination test scores (Folstein et al., 1975) were associated with the time to AD conversion, but unrelated to the time to censoring, in a survival analysis for the time to event.

In this sample, Cronbach's Coefficient Alpha estimate (Cronbach, 1951) for the UPSIT items is 0.8706, indicating a good consistency among

Figure 1. Item response rate and discrimination accuracy



items in relation to the latent variable for olfactory function. Figure 1 reveals a wide range of percentages of odors correctly identified ($28.91\% - 89.84\%$), as well as a wide range of item discrimination accuracies ($0.1532 - 0.4063$).

To identify items in the UPSIT for a reduced scale that is related to the risk of AD in the MCI patients, we first applied commonly used backward, forward or stepwise selection procedures based on Cox proportional hazards models. Using the criterion of a significance level of 0.1 with backward selection, 12 items were selected with parameter estimates 6 positive and 6 negative; with forward or stepwise selection, 8 items were selected, the estimated parameters were 3 positive and 5 negative. When adaptive LASSO for Cox proportional hazards models (Zhang and Lu, 2007) was applied, 27 items were selected with 12 parameters estimated to be positive, and 15 negative. Obviously, the Cox regression model based variable selection procedures did not produce results meaningful for risk related item reduction.

We then applied the proposed procedure to produce a reduced uni-dimensional scale that may efficiently discriminate the risk of AD. The 21 items initially identified as the least useful items were excluded from the

DOI: 10.2202/1557-4679.1094

full 40-item scale. Using the threshold value $(\gamma_0, \gamma_1) = (1.281,\ 1.2816)$, the item set finally selected is $W_6 = \{X8,\ X14,\ X22,\ X33,\ X35,\ X37\}$ with a discrimination accuracy estimate of 0.7111, close to the 0.7166 of the full scale. The choice of this threshold was suggested by the results of the simulation study: among the four sets of criteria, $(\gamma_0, \gamma_1) = (1.281,\ 1.2816)$ selected scales having mean discrimination accuracy closest to that of the "true" subscale when using a sample of size $N = 120$ and a censored proportion of 75%, which is similar to the censored proportion of 70.31% of the study sample of 128 MCI patients.

To evaluate the variation in item selections, we applied the proposed selection procedure to 1000 bootstrap samples obtained from the original data set ($n = 128$). The mean discrimination accuracy of the full scale is 0.7150 ($se = 0.0448$), similar to the estimate of 0.7166 from the original sample. With criteria $(\gamma_0, \gamma_1) = (1.281,\ 1.2816)$, the average number of selected items for the reduced scale is 6.37 ($se = 1.16$) with a mean discrimination accuracy of 0.7728 ($se = 0.0412$).

Table 4 shows the item selection spectrum with bootstrap samples. Note that five of the six items selected from the original sample are among those most frequently selected using the bootstrap samples. Items $X8,\ X33,\ X37$ appear to be the most important, followed by $X14$, while $X17,\ X20,\ X35,\ X38$ might deserve some attention as well.

To examine the predictive performance of the item selection method for this sample, as a reviewer suggested, we use the leave-one-out cross-validation method. Using $n-1$ observations for item selection under a given criterion, we calculate a predicted score for each of the left out observations based on the selected item set. These predicted scores are then used to generate a DA estimate, $\widehat{DA}_{cv}$, to compare with the DA estimate of the item set selected under the same criteria using all n observations. With item selection criteria $(\gamma_0, \gamma_1) = (1.281,\ 1.2816)$, the selected item sets based on $n-1$ observations had a mean size of 6.9 with the most frequently selected items being the members of W_6. The estimate $\widehat{DA}_{cv}$ for the predicted scores was 0.5053, much lower than $\widehat{DA}(W_6) = 0.7111$ based on all n observations. When reducing the threshold values to be $(\gamma_0, \gamma_1) = (1.036,\ 1.0364)$, the selected item sets with $n-1$ observations were larger in size with a mean of 9.1 and $\widehat{DA}_{cv}$ was improved to 0.5976, but still lower from $\widehat{DA}(W_9) = 0.7535$ estimated using the whole sample. Using criteria $(\gamma_0, \gamma_1) = (0.524,\ 0.5244)$, the selected item sets with $n-1$ observations had a mean size of 12.0, and $\widehat{DA}_{cv}$ was further improved to 0.6582, while still not close to $\widehat{DA}(W_{12}) = 0.7675$ based on the whole sample. This analysis suggests that the cross-validation estimate $\widehat{DA}_{cv}$

can be improved by lowering the item selection criteria, as observed in DA estimates based on all observations.

Table 4. Frequencies of selection of specific UPSIT items in 1000 bootstrap samples

	Item	Frequency		Item	Frequency
1	pizza	13	21	lilac	240
2	bubble gum	98	22	turpentine	255
3	menthol	6	23	peach	80
4	cherry	11	24	root beer	276
5	motor oil	29	25	dill pickle	13
6	mint	182	26	pineapple	18
7	banana	2	27	lime	3
8	clove	550	28	orange	10
9	leather	35	29	wintergreen	161
10	coconut	229	30	watermelon	45
11	onion	5	31	paint thinner	38
12	fruit punch	241	32	grass	215
13	licorice	3	33	smoke	634
14	cheddar cheese	482	34	pine	16
15	cinnamon	221	35	grape	393
16	gasoline	2	36	lemon	106
17	strawberry	347	37	soap	660
18	cedar	0	38	natural gas	332
19	chocolate	31	39	rose	17
20	gingerbread	316	40	peanut	59

Evidently, the 40−item scale can be greatly reduced. However, to obtain a confirmative result of item selection with good predictive performance, we need to increase the sample size and the uncensored proportion.

Discussion

We have extended the nonparametric method for selecting binary risk related items from a uni-dimensional scale for screening, to accommodate cases where risk is quantified in ordinal categories or measured as time to event possibly subject to random censoring. The method is invariant to rank-preserving transformations of the scale score and the risk measure. The extended method is also applicable where items on a scale have $K(> 2)$ response levels, because

DOI: 10.2202/1557-4679.1094

$K-1$ binary indicators can be produced; for example, $I(X=j)$ for $j=2,\cdots,K$ can be used. If the K response levels are ordinal, then indicators $I(X\geq j)$ for $j=2,\cdots,K$ may be used.

By evaluating, at every step, changes in discrimination accuracy, the proposed item selection procedure enables us to begin by removing the least useful items from the full scale, and then apply stepwise selection to the remaining items. To decide whether or not to include an item in the reduced scale, the proposed stepwise selection requires pre-set values for thresholds γ_0 and γ_1 which will determine the size of the reduced scale. An upper bound on the threshold is necessary for the estimated discrimination accuracy of the reduced scale to exceed that of the full scale. Based on the simulation study, we recommend using threshold values between 0.8 and 1.3. The results of the simulation study indicate that we can reduce the impact of the threshold values by increasing sample size and the uncensored proportion. Meanwhile, examination of item selection frequencies in a number of bootstrap samples may help assess the variation in item selection. The most frequently selected items can be used for the reduced scale. This resembles the 'bootstrap model averaging' approach to survival analysis discussed by Augustin et al. (2005).

In application, it is important to examine the predictive performance of the reduced scale selected using specific criteria. This can be done by using the leave-one-out cross-validation method, as shown in the example.

A reviewer has pointed out that a test-based backward selection, which eliminates items present in the larger item set at the previous steps, is simpler than the proposed one and can be used as an alternative. Because the descending procedure evaluates each item conditioned on the other items in the set, when the criterion is met, some less important items may not be eliminated. Consequently, the descending procedure could select more items though it is possible that not all the selected items will contribute significantly to DA. In contrast, the stepwise selection method builds up the item set dynamically, allowing for both the addition of important items and the elimination of some previously selected items whose importance was lessened by the introduction of a new item into the set. To demonstrate this, we applied the descending procedure to the data sets previously used for stepwise selection procedure. As expected, for each given criterion, with the descending procedure, the average size of the selected subscales was larger, while the mean DA was similar and the number of DA improved cases was slightly smaller. For example, with a sample size of $n=120$ and censored proportion of 75%, when using criterion I, the mean size of selected subscales with the descending selection procedure was 0.55 larger than that of the scales generated

by proposed procedure. The discrepancy decreased to 0.23 when criterion IV was used. The result suggests that as an alternative, the descending procedure may choose slightly more items without improving DA under the same criterion, compared to the proposed procedure.

The proposed method has limitations. To select a reduced scale with good predictive performance, it is crucial to apply the selection procedure to a large sample with low proportion of censored observations. As indicated in the illustration example, predictive performance may not be acceptable when the sample size, especially the number of uncensored observations, is not large. When using the proposed method on a relatively small sample with few risk categories, it is possible that the reduced scale will have a DA estimate close to one, suggesting an over-fit. To avoid the problem one has to increase the sample size, especially of the uncensored sample, refine the risk categories, and use various criteria in the statistical test based selection process. In the presence of random censoring, the method requires censoring time to be independent of risk predictors. Since this assumption may not always hold, a generalized method allowing for dependence on predictors is desirable. In longitudinal studies, the subjects may be assessed over time by the same uni-dimensional scale. Efforts to develop an extension of the method for use with repeated measures would be worthwhile.

Appendix

We show asymptotic normality of $\Delta A_k(-X_h|W_k)$. The asymptotic normality of $\Delta A_k(+X_h|W_{k-1})$ can be shown similarly.

Let $\Lambda_G(u)$ denote the common cumulative hazard function of censoring time Q and

$$\eta_{ij}(h,k) = I(z_{ij}(h) = -1,\ e_{ij}(k) = -1) - I(z_{ij}(h) = 1,\ e_{ij}(k) = 0).$$

With the use of $(\hat{G} - G)/G$ martingale integral representation, we have

$$\begin{aligned}
\Delta A_k(-X_h|W_k) &= \sum_{i=1}^{n}\sum_{i=1}^{n} \frac{d_i}{G^2(Y_i)} I\{Y_i \le Y_j\}\eta_{ij}(h,k) \\
&\quad +2\sum_{i=1}^{n}\sum_{i=1}^{n} \frac{d_i}{G^2(Y_i)} I\{Y_i \le Y_j\}\eta_{ij}(h,k)\frac{G(Y_i) - \hat{G}(Y_i)}{G(Y_i)} \\
&\quad +o_p(1) \\
&= \sum_{i=1}^{n}\sum_{i=1}^{n} \frac{d_i}{G^2(Y_i)} I\{Y_i \le Y_j\}\eta_{ij}(h,k)
\end{aligned}$$

DOI: 10.2202/1557-4679.1094

$$+2n\sum_{i=1}^{n}\int_0^\infty \frac{\xi(t)}{\pi(t)}dM_i(t) + o_p(1)$$

where

$$\xi(t) = \lim_{n\to\infty}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{i=1}^{n}\frac{d_i}{G^2(Y_i)}I\{Y_i \le Y_j\}I\{Y_i > t\}\eta_{ij}(h,k)$$

$$\pi(t) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} I\{Y_i > t\}$$

$$M_i(t) = I\{Y_i \le t\}(1-d_i) - \int_0^t I\{Y_i > u\}d\Lambda_G(u).$$

Therefore,

$$\begin{aligned}\sqrt{n}\,\left(\frac{\Delta A_k(-X_h|W_k)}{n^2} - \mu\right) &= n^{-3/2}\sum_{i=1}^{n}\sum_{i=1}^{n}\left(\frac{d_i}{G^2(Y_i)}I\{Y_i \le Y_j\}\eta_{ij}(h,k) - \mu\right)\\ &\quad +2n^{-1/2}\sum_{i=1}^{n}\int_0^\infty \frac{\xi(t)}{\pi(t)}dM_i(t) + o_p(1).\end{aligned}$$

By the standard U-statistic asymptotic theory, it follows that the quantity $\sqrt{n}(\Delta A_k(-X_h|W_k)/n^2 - \mu)$ is asymptotically normal with mean 0.

Below we show how to obtain the asymptotic variance of the quantity $\sqrt{n}(\Delta A_k(-X_h|W_k)/n^2 - \mu)$. It is easy to see that the first term

$$n^{-3/2}\sum_{i=1}^{n}\sum_{i=1}^{n}\left(\frac{d_i}{G^2(Y_i)}I\{Y_i \le Y_j\}\eta_{ij}(h,k) - \mu\right) \tag{1}$$

is a U-statistic, and its variance can be estimated easily. From the martingale representation of the second term

$$2n^{-1/2}\sum_{i=1}^{n}\int_0^\infty \frac{\xi(t)}{\pi(t)}dM_i(t), \tag{2}$$

it follows that its asymptotic variance is

$$4\int_0^\infty \frac{\xi^2(t)}{\pi(t)}d\Lambda_G(t).$$

Notice that

$$E\left\{\frac{d_i}{G^2(Y_i)}I\{Y_i \le Y_j\}\eta_{ij}(h,k)(1-d_j)\frac{\xi(Y_j)}{\pi(Y_j)}\right\}$$

$$
\begin{aligned}
&= E\left\{\frac{d_i}{G^2(Y_i)} I\{Y_i \le C_j\}\eta_{ij}(h,k) I\{C_j \le Y_j\}\frac{\xi(C_j)}{\pi(C_j)}\right\} \\
&= E\left\{\int_0^\infty \frac{d_i}{G^2(Y_i)} I\{Y_i \le t\}\eta_{ij}(h,k) I\{t \le Y_j\}\frac{\xi(t)}{\pi(t)} dG(t)\right\} \\
&= E\left\{\int_0^\infty \frac{d_i}{G^2(Y_i)} I\{Y_i \le t\}\eta_{ij}(h,k) I\{t \le Y_j\}\frac{\xi(t)}{\pi(t)} G(t) d\Lambda_G(t)\right\}.
\end{aligned}
$$

Thus, the limiting covariance between the terms (1) and (2) is

$$-4\int_0^\infty \frac{\xi^2(t)}{\pi(t)} d\Lambda_G(t).$$

As a result, the limiting covariance for $\sqrt{n}(\Delta A_k(-X_h|W_k)/n^2 - \mu)$ is the variance of the first term (1) minus $4\int_0^\infty \frac{\xi^2(t)}{\pi(t)} d\Lambda_G(t)$. It is easy to see that a consistent estimator for this quantity can be obtained by replacing the unknown $G(\cdot)$, $\xi(\cdot)$ and $\pi(\cdot)$ with corresponding empirical estimators $\hat{G}(\cdot)$, $\hat{\xi}(\cdot)$ and $\hat{\pi}(\cdot)$.

References

Antolini, L. Boracchi, P. and Biganzoli, E. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 2005, **24**(24), 3927-3944.

Augustin, N., Sauerbrei, W. and Schumacher, M. The practical utility of incorporating model selection uncertainty into prognostic models for survival data. *Statistical Modelling*, 2005, **5**(2), 95-118.

Brier, G.W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 1950, **75**(1), 1-3.

Chambless, L. and Diao, G. Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Statistics in Medicine*, 2006, **25**(20), 3474-3486.

Cronbach, L.J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, **16**(3), 297-334.

Devanand, D.P., Michaels-Marston, K.S., Liu, X., Pelton, GH., Padilla, M., Marder, K., Bell, K., Stern, Y. and Mayeux, R. Olfactory deficits in mild cognitive impairments predict Alzheimer's disease on follow-up.

DOI: 10.2202/1557-4679.1094

American Journal of Psychiatry, 2000, **157**, 1399-1405.

Doty, R.L., Reyes, P.F. and Gregor, T. Presence of both odour identification and detection deficits in Alzheimer's disease. *Brain Research Bulletin*, 1987, **18**, 597-600.

Doty, R.L., Shaman, P. and Dann, M. Development of the University of Pennsylvania Smell Identification Test: a standardized microencapsulated test of olfactory function. *Physiology Behavior*, 1984, **32**, 489-502.

Efron, B. and Tibshirani, R.J. *An Introduction to the Bootstrap*, 1993, Chapman & Hall, New York.

Folstein, M.F., Folstein, S.E. and McHugh, P.R. "Mini-Mental State". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 1975, **12**(3), 189-198.

Gönen, M. and Heller, G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 2005, **92**(4), 965-970.

Hanley, J.A. and McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 1982, **143**(1), 29-36.

Harrell, F.E., Lee, K.L., Calife, R.M., Pryor, D.B. and Rosati, R.A. Regression modeling strategies for improved prognostic prediction. *Statistics in Medicine*, 1984, **3**(2), 143-152.

Harrell, F.E., Lee, K.L. and Mark, D.B. 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.

Hastie, T., Tibshirani, R. and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2001, Springer, New York.

Heagerty, P.J., Lumley, T. and Pepe, M.S. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 2000, **56**(2), 337-344.

Heagerty, P.J. and Zheng, Y. Survival model predictive accuracy and ROC curves. *Biometrics*, 2005, **61**(1), 92-105.

Kendall, M.G. *Rank Correlation Methods*, 1970, Griffin, London.

Liu, X. and Jin, Z. Item reduction in a scale for screening. *Statistics in Medicine*, 2007, **26**(23), 4311-4327.

Lord, F.M. and Novick, M.R. *Statistical Theories of Mental Test Scores*, 1968, Addison-Wesley Publishing, Massachusetts.

Pencina, M.J. and D'Agostino, R.B. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine*, 2004, **23**(13), 2109-2123.

Pepe, M.S. *The Statistical Evaluation of Medical Tests for Classification and Prediction*, 2003, Oxford University Press, New York.

Somers, R.H. A similarity between Goodman and Kruskal's Tau and Kendall's Tau, with a partial interpretation of the latter. *Journal of the American Statistical Association*, 1962, **57**, 804-812.

Tabert, M.H., Liu, X., Doty, R.L., Serby, M., Albers, M., Zamora, D., Pelton, G., Marder, K. and Devanand, D.P. A 10-item smell identification scale related to risk of Alzheimer's disease. *Annals of Neurology*, 2005, **58**, 155-160.

Tabert, M.H., Manly, J.J., Liu, X., Pelton, G.H., Rosenblum, S., Jacobs, M., Zamora, D., Goodkind, M., Bell, K., Stern, Y. and Devanand, D.P. Neuropsychological prediction of conversion to Alzheimer disease in patients with mild cognitive impairment. *Archive of General Psychology*, 2006, **63**, 916-924.

Zhang, H.H. and Lu, W. Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, 2007, **94**(3), 691-703.

Zheng, Y., Cai, T. and Feng, Z. Application of the time-dependent ROC curves for prognostic accuracy with multiple biomarkers. *Biometrics*, 2006, **62**(1), 279-287.

Zhou, X.H., Obuchowski, N.A. and McClish, D.K. *Statistical Methods in Diagnostic Medicine*, 2002, Wiley, New York.

DOI: 10.2202/1557-4679.1094