9-7-2007

ROC Surfaces in the Presence of Verification Bias

Yueh-Yun Chi
University of Florida, yychi@biostat.ufl.edu

Xiao-Hua (Andrew) Zhou
University of Washington, azhou@u.washington.edu

1 Introduction

Diagnostic tests are important for early detection and guiding treatment of various diseases. A gold standard test, if one exists, ideally provides definitive examination of disease status, but often results in high cost and can be invasive. To counter these drawbacks, less expensive or invasive diagnostic tests are often used for the primary assessment. The accuracy of a diagnostic test can be evaluated and assured by comparing it to the definitive gold standard test. Statistics such as sensitivity and specificity, which respectively account for the proportion of true positives and true negatives, have been commonly used when both the scale of a diagnostic test and true disease status are binary. The Receiver Operating Characteristic (ROC) curve plots a test's sensitivity against its false negative rate (1-specificity) when the test is in either an ordinal- or continuous-scale, and the area under the ROC curve has been long served as a composite index to describe the discriminatory property of a such test. Zhou, Obuchowski, and McClish (2002) provided a comprehensive account of statistical methods for a two-class disease diagnosis, when the true disease status is either the presence or absence of the disease.

In medical practice, there are situations in which the presence or absence of the disease is not sufficient in describing and presenting the severity and progress of the disease. For instance, in the study of Alzheimer's disease (AD), patients are diagnosed by autopsy as in low likelihood of AD, intermediate likelihood of AD, or high likelihood of AD. The definitive diagnosis for differential likelihood indicates AD neuropathology severity, which may be reflected in the degree of cognitive or neuropsychological decline, and may have an impact on the decision of nursing and treatment plans. Diagnostic tests capable of discerning patients among these three classes of severity are then of clinical importance. In the literature, the ROC methodology has been extended to three-class disease status problems by several authors (Scurfield, 1996; Mossman, 1999; Dreiseitl, Ohno-Machado, and Binder, 2000, Obuchowski, 2005, Nakas and Yiannoutsos, 2004 and 2006, and Xiong et al., 2006).

Analogous to the two-way ROC curve, Scurfield (1996) defined the three-way ROC surface as a graph built on the three true classification rates. If we denote $\hat{D}$ and D as the rated and true disease status, the three axes of the surface are the three correct classification rates $P(\hat{D} = k \mid D = k)$, for $k = 1, 2, 3$. There was no direct account to the six false classification rates $P(\hat{D} = k_1 \mid D = k_2)$, for $k_1 \neq k_2$. By varying the decision rules imposing upon a diagnostic test T to determine $\hat{D}$, points on the surface are obtained from the contingency tables between $\hat{D}$ and D. For either an ordinal- or continuous-scaled T, if a higher value of T corresponds to a higher value of D, the decision rule is defined by a pair of ordered decision thresholds (d_1, d_2), such that $\hat{D} = 1$ if $T \leq d_1$, $\hat{D} = 2$ if $d_1 < T \leq d_2$, and $\hat{D} = 3$ if $T > d_2$. The points $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$ are on every ROC surface, and the lines connecting them result in the surface corresponding to a test without any discriminatory power among the three classes. The surface corresponding to a perfect test is the surface of a unit cube. The volume under the ROC surface can be a useful index for the accuracy of a three-class diagnostic test. It has been shown that the volume under the ROC surface equals the probability that diagnostic measurements of any three patients, one from each class, are in an correct order. The volume under the ROC surface of 1/6 corresponds to a test without discriminatory power, and the value of 1 indicates a perfect test. We note that two tests can have similar volumes under the ROC surfaces, but different ROC surfaces, for instance, one test may differentiate disease category 1 very well while the other test may do better in differentiating disease category 3. Caution must be taken when doing the comparison between two ROC volumes.

In practice, it is plausible that only a subgroup of patients who initially are tested subsequently receives the definitive assessment for disease status. Subjects may refuse or simply are not capable of participating in the definitive examination. The mechanism by which patients are selected for verification may be variable. For example, in the study of Alzheimer's disease, the definitive examination is done through a brain autopsy. Patients might still be alive resulting in not being verified, or they died but no autopsy was performed.

The selection for an autopsy may depend on the degree of cognitive impairment, which may also affect the clinical diagnosis of Alzheimer's disease. In the assessment of diagnostic accuracy, omission of those nonverified cases can seriously bias the estimate, and the bias is referred to as verification bias (Begg and Greenes, 1983).

There have been methods proposed in the literature to account for the presence of verification bias for a two-class disease status. Gray et al. (1984) and Zhou (1996) both derived the maximum likelihood estimation for the area under the ROC curve when disease verification is subject to selection bias. Harel and Zhou (2006) adopted a multiple imputation framework. Alonzo and Pepe (2005) proposed and compared imputation and reweighting bias-corrected estimators of ROC curves and area under the ROC curve for continuous tests. Kosinski and Barnhart (2003), and Rotnitzky et al. (2006) suggested a method for correcting for non-ignorable verification bias. In this paper, we extend the methodology to the three-class disease status problems. We formulate the presence of verification bias into the likelihood-based framework. The proposed approach is flexible in allowing for the selection mechanism to depend on initial test results and/or any relevant discrete baseline covariates. Methods for comparing diagnostic tests in the presence of verification bias have also been developed.

The rest of the paper is organized as the following. We introduce the motivating research of Alzheimer's disease in Section 2. In Section 3, we propose our method when the selection is related to the test results, and further incorporate discrete baseline covariates into the proposed method in Section 4. We provide an extension to the comparison between two volumes under the ROC surfaces in the presence of differential verification in Section 5. In Section 6, we apply the proposed method to the motivating application on Alzheimer's disease described in Section 2. In Section 7, we conduct an extensive simulation to demonstrate finite-sample performance of the proposed method. The simulation is set up to cover different true volumes under the ROC surfaces. We discuss some future research directions

in Section 8.

2 The study of Alzheimer's Disease

The National Alzheimer's Coordinating Center (NACC) is funded to facilitate collaborative research and maintain a database of information collected by the Alzheimer's Disease Centers (ADCs) throughout the United States of America. These centers have conducted clinical and laboratory research on the causes and clinical courses of Alzheimer's disease (AD). Most patients at ADCs were referred or self-referred for evaluation of possible dementia, and they were followed over time with periodic clinical evaluation and cognitive testing. For patients who die, permission for brain autopsy was sought.

The definitive examination of Alzheimer's disease was based on the extent of neuritic plaques and neurofibrillary tangles, the hallmarks of AD, at brain autopsy. The NIA/Reagan Institute criteria, based on the frequency of both plaques and tangles in the neocortex to link to the severity of AD pathology, were graded as no or low ($D = 3$), intermediate ($D = 2$), or high likelihood ($D = 1$) of dementia being due to AD. The NIA criteria were absent for patients still alive or who died but had no brain autopsy. The search for clinical diagnostic tools of AD has long been needed and initiated to ensure prompt health care and treatment to AD patients.

In the study, a patient's cognitive function was measured by the latest Mini-Mental State Examination (MMSE) score recorded prior to death. The MMSE (Folstein, Folstein, and McHugh, 1975), is a screening tool that evaluates orientation to place, orientation to time, registration, attention and concentration, recall, language, and visual construction. It is scored as the number of correctly completed items, and could range from 0 (too impaired even to answer questions) to 30 (perfect score, cognitively intact). For elderly adults (the group of patients normally at a higher risk for AD), MMSE scores were used by Reisberg et al. (2003) and Just (2004), to assess patients with either no AD (MMSE>26: dMMSE=4),

mild AD (15≤MMSE≤26: dMMSE=3), moderate AD (10≤MMSE≤14: dMMSE=2), or severe AD (MMSE<10: dMMSE=1). This MMSE-based clinical diagnosis, referred to as dMMSE, provides a gateway for AD assessment. An alternative approach for AD assessment is through clinical evaluation for dementia (CDD) by clinicians. With access to various neuropsychological, cognitive and medical image assessments, experienced clinicians were able to coordinate the information for the evaluation of dementia. Based on the degree of abnormal cognition, patients were categorized into one of the three groups, namely AD dementia (CDD=1), mild cognitive impairment (CDD=2), or non-dementia (CDD=3).

Among a total of 18,838 patients in a subset of a NACC dataset, only 2,497 had died and agreed to brain autopsy. The selection for autopsy verification was clearly dependent of patients' performance on dMMSE and CDD. For instance, based on dMMSE scores, 28.9% of patients with the diagnosis of severe AD were verified with an autopsy, while only 14.3%, 9.5%, and 9.2% of patients respectively diagnosed as moderate, mild, and no AD underwent an autopsy. The verification rate increased with the severity of the assessment. For CDD, the proportions of verification are respectively 15.4%, 9.0%, and 11.0% for patients determined as AD dementia, mild cognitive impairment, and non-dementia. In this obvious presence of differential verification among patient groups, the accuracy of using dMMSE and CDD as diagnostic tools for AD severity is of interest, as well as the comparison between the two.

To further explore the data, we note that the study participants came from eight distinct ADCs across the United States. Depending on the majority of patient characteristics, which might vary geographically, verification rates may differ across centers. Figure 1 shows the verification rates across the eight centers for each of the classification determined by dMMSE and CDD. It is clear to see that the probability of being verified varies not only among the test categories, but also among the centers. For instance, the rate of verification obviously decreased with the dMMSE classification for the eighth ADC, whereas the rates of verification when dMMSE=1 or dMMSE=4 were about twice as much the rates when dMMSE=2 or

dMMSE=3, for the third ADC. In the evaluation of the overall accuracy of dMMSE and CDD, it may be more plausible to account for center differences in verification. On the other hand, some researchers believe that the correlation between the test measurement (dMMSE or CDD) and the autopsy result would vary across different centers, partly due to the differential dependence of the test and autopsy classification on the center. This predisposition would then lead researchers to focus on the center-specific evaluation instead of the overall evaluation.

3 Test-dependent verification

For three-class disease status problems, Scurfield (1996) defined the three-way ROC surface by points whose coordinates are the three correct classification rates, over all possible pairs of ordered decision thresholds. We let T denote an ordinal test measurement ranging from 1 to M, D denote the true disease status, and a higher value of T corresponds to a higher value of D. For any pair of ordered decision thresholds (d_1, d_2), where $0 \leq d_1 < d_2 \leq M$, the following decision rule may be applied: if $T \leq d_1$ then $\hat{D} = 1$, else if $d_1 < T \leq d_2$ then $\hat{D} = 2$, else $\hat{D} = 3$. Here $\hat{D}$ indicates the random variable of disease diagnosis based on T. Given (d_1, d_2) and the independence among study patients, the three correct classification rates, for $k = 1, 2, 3$, are defined as

$$P(\hat{D} = k \mid D = k) = P(d_{k-1} < T \leq d_k \mid D = k) = \begin{cases} 0, & \bar{d}_{k-1} > \underline{d}_k \\ \sum_{i=\bar{d}_{k-1}}^{\underline{d}_k} \pi_{ik}, & \bar{d}_{k-1} \leq \underline{d}_k \end{cases}, \qquad (3.1)$$

where $\pi_{ik} = P(T = i|D = k)$, $\bar{d}_{k-1}$ is the smallest integer greater than d_{k-1}, and $\underline{d}_k$ is the largest integer less than or equal to d_k. We specify the boundary conditions as, $d_0 = 0$ and $d_3 = M$. If the unknown parameters π_{ik} in (3.1) are replaced by their estimates, we can then obtain the estimates of the three correct classification rates, and further, construct the empirical ROC surface over all possible pairs of decision thresholds, (d_1, d_2). In this section, we derive the maximum likelihood (ML) estimates of π_{ik}'s when differential verification is present.

Nakas and Yiannoutsos (2004) used the following notation system and demonstrated the volume under the ROC surface is given by

$$\theta = P(Y_1 < Y_2 < Y_3) + \frac{1}{2}[P(Y_1 < Y_2 = Y_3) + P(Y_1 = Y_2 < Y_3)] + \frac{1}{6}P(Y_1 = Y_2 = Y_3),$$

where Y_k indicates the test measurement obtained from the kth disease group, for $k = 1, 2, 3$. To describe it in words, the volume under the ROC surface equals the probability that the measurements T of three randomly chosen patients, one from each disease group, are correctly ordered. Their results assumed that each study subject had disease verification. With a finite support of T, the volume θ is equivalent to

$$\theta = \sum_{i=1}^{M-2}\sum_{j=i+1}^{M-1}\sum_{k=j+1}^{M}\pi_{i1}\pi_{j2}\pi_{k3} + \frac{1}{2}\left\{\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}(\pi_{i1}\pi_{j2}\pi_{j3} + \pi_{i1}\pi_{i2}\pi_{j3})\right\} + \frac{1}{6}\sum_{i=1}^{M}\pi_{i1}\pi_{i2}\pi_{i3}. \quad (3.2)$$

Both the ROC surface and its volume can be expressed as functions of π_{ik}'s. Following Bayes' theorem, we know that

$$\pi_{ik} = P(T = i|D = k) = \frac{\tau_i\phi_{ki}}{\sum_{j=1}^{M}\tau_j\phi_{kj}}, \quad (3.3)$$

where $\tau_i = P(T = i)$ and $\phi_{ki} = P(D = k|T = i)$, with $\tau_M = 1 - \sum_{i=1}^{M-1}\tau_i$ and $\phi_{3i} = 1 - \phi_{1i} - \phi_{2i}$. To estimate π_{ik}'s, one may first need to have proper estimates (with account for verification bias) for all τ_i's and ϕ_{ki}'s.

We propose a likelihood-based approach to estimate parameters $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{M-1})$ and $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_M)$, where $\boldsymbol{\phi}_i = (\phi_{1i}, \phi_{2i})$. The observed data with a differential verification status among patients can be summarized as in Table 1. We denote V as the disease verification indicator, which equals one if the patient receives the definitive examination, and equals zero otherwise. When $V = 0$, the patient is not selected for disease verification, and hence, the summarized frequencies are only available for test values collapsed over all three disease states.

We make the following assumption about the selection mechanism. We assume the chance of getting verified is conditionally independent of the unknown true disease status,

given the test measurement. In other words, $P(V|D,T) = P(V|T)$, and it follows $P(D|T) = P(D|V,T)$. Specifically, $P(D|T) = P(D|V=1,T)$ and $\phi_{ki} = P(D=k|T=i,V=1)$. This assumption is a special case of the missing at random (MAR) assumption on the missing data mechanism proposed by Rubin (1976). The chance of getting verified, though might depend on test results, is dealt as the nuisance parameter, and can be excluded from the likelihood function since it is distinct from the parameters of interest, $\boldsymbol{\tau}$ and $\boldsymbol{\phi}$. Conceptually, the likelihood function is the product of $P(T)P(D|T)P(V|D,T)$, which is proportional to $P(T)P(D|T,V=1)$ under the above assumptions. The log-likelihood function based on the observed data in Table 1, can then be derived as

$$l(\boldsymbol{\tau},\boldsymbol{\phi}) = \sum_{i=1}^{M} n_i \log(\tau_i) + \sum_{i=1}^{M}\sum_{k=1}^{3} a_{ki}\log(\phi_{ki}), \tag{3.4}$$

for $i = 1,\ldots,M$. Note that $\boldsymbol{\tau}$ and $\boldsymbol{\phi}_i$, for $i = 1,\ldots,M$, are distinct parameters, $l_1(\boldsymbol{\tau}) = \sum_{i=1}^{M} n_i \log(\tau_i)$ and $l_{2i}(\boldsymbol{\phi}_i) = \sum_{k=1}^{3} a_{ki}\log(\phi_{ki})$ are the log-likelihood functions for multinomial distributions, respectively. Thus, the ML estimates for $\boldsymbol{\tau}$ and $\boldsymbol{\phi}_i$, respectively, are

$$\hat{\tau}_i = \frac{n_i}{n}, \quad i = 1,\ldots,M, \tag{3.5}$$

$$\hat{\phi}_{ki} = \frac{a_{ki}}{a_{1i}+a_{2i}+a_{3i}}, \quad k = 1,2; \quad i = 1,\ldots,M, \tag{3.6}$$

where $n = \sum_{i=1}^{M} n_i$. The observed Fisher information matrix defined on $(\boldsymbol{\tau},\boldsymbol{\phi})$ is a block diagonal matrix given by

$$I(\boldsymbol{\tau},\boldsymbol{\phi}) = \mathrm{diag}(I_1(\boldsymbol{\tau}), I_{21}(\boldsymbol{\phi_1}),\ldots,I_{2M}(\boldsymbol{\phi_M})),$$

where $I_1(\boldsymbol{\tau})$ and $I_{2i}(\boldsymbol{\phi_i})$ are respectively the observed Fisher's information matrix on the log-likelihood $l_1(\boldsymbol{\tau})$ and $l_{2i}(\boldsymbol{\phi}_i)$. The ML estimates of π_{ik} can be obtained by replacing τ_i and ϕ_{ki} in (3.3) by $\hat{\tau}_i$ and $\hat{\phi}_{ki}$. Subsequently substituting the unknown parameters π_{ik} in (3.2) by their ML estimates through the Bayes' theorem, we can then obtain the ML estimator $\hat{\theta}$

for the volume under the ROC surface as

$$\hat{\theta} = \frac{\sum_{i=1}^{M-2}\sum_{j=i+1}^{M-1}\sum_{k=j+1}^{M}\hat{\tau}_i\hat{\tau}_j\hat{\tau}_k\hat{\phi}_{1i}\hat{\phi}_{2j}\hat{\phi}_{3k}}{\left(\sum_{i=1}^{M}\hat{\tau}_i\hat{\phi}_{1i}\right)\left(\sum_{i=1}^{M}\hat{\tau}_i\hat{\phi}_{2i}\right)\left(\sum_{i=1}^{M}\hat{\tau}_i\hat{\phi}_{3i}\right)} + \frac{\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}\hat{\tau}_i\hat{\tau}_j^2\hat{\phi}_{1i}\hat{\phi}_{2j}\hat{\phi}_{3j} + \hat{\tau}_i^2\hat{\tau}_j\hat{\phi}_{1i}\hat{\phi}_{2i}\hat{\phi}_{3j}}{2\left(\sum_{i=1}^{M}\hat{\tau}_i\hat{\phi}_{1i}\right)\left(\sum_{i=1}^{M}\hat{\tau}_i\hat{\phi}_{2i}\right)\left(\sum_{i=1}^{M}\hat{\tau}_i\hat{\phi}_{3i}\right)}$$
$$+ \frac{\sum_{i=1}^{M}\hat{\tau}_i^3\hat{\phi}_{1i}\hat{\phi}_{2i}\hat{\phi}_{3i}}{6\left(\sum_{i=1}^{M}\hat{\tau}_i\hat{\phi}_{1i}\right)\left(\sum_{i=1}^{M}\hat{\tau}_i\hat{\phi}_{2i}\right)\left(\sum_{i=1}^{M}\hat{\tau}_i\hat{\phi}_{3i}\right)}. \quad (3.7)$$

This estimate is equivalent to the one proposed by Nakas and Yiannoutsos (2004) when no verification bias exists. By the use of the Delta method (Agresti, 1990, p56-58), the variance of $\hat{\theta}$, can be estimated as

$$\frac{\partial\theta^T}{\partial\boldsymbol{\tau}}I_1^{-1}(\boldsymbol{\tau})\frac{\partial\theta}{\partial\boldsymbol{\tau}} + \frac{\partial\theta^T}{\partial\boldsymbol{\phi}_1}I_{21}^{-1}(\boldsymbol{\phi}_1)\frac{\partial\theta}{\partial\boldsymbol{\phi}_1} + \ldots + \frac{\partial\theta^T}{\partial\boldsymbol{\phi}_M}I_{2M}^{-1}(\boldsymbol{\phi}_M)\frac{\partial\theta}{\partial\boldsymbol{\phi}_M}, \quad (3.8)$$

evaluated at $\boldsymbol{\tau} = \hat{\boldsymbol{\tau}}$ and $\boldsymbol{\phi}_i = \hat{\boldsymbol{\phi}}_i$, where $I_1^{-1}(\boldsymbol{\tau})$, $I_{2i}^{-1}(\boldsymbol{\phi}_i)$, $\partial\theta/\partial\boldsymbol{\tau}$ and $\partial\theta/\partial\boldsymbol{\phi}_i$ are given in the Appendix.

An alternative way to estimate the variance of $\hat{\theta}$ is to use the jackknife method. As illustrated by Efron and Tibshirani (1993), estimates of θ with the exclusion of a single patient allows for the estimation of the variance of $\hat{\theta}$ as

$$\frac{n-1}{n}\left[\sum_{i=1}^{M}\left\{b_{\cdot i}(\hat{\theta}_{(4i)} - \hat{\theta}_{(\cdot)})^2 + \sum_{k=1}^{3}a_{ki}(\hat{\theta}_{(ki)} - \hat{\theta}_{(\cdot)})^2\right\}\right], \quad (3.9)$$

where $\hat{\theta}_{(ki)}$ is the estimate of θ after deleting a patient with $V = 1$, $T = i$, and $D = k$, for $k = 1, 2, 3$, $\hat{\theta}_{(4i)}$ is the estimate of θ after deleting a patient with $V = 0$, $T = i$, and $\hat{\theta}_{(\cdot)} = (1/n)\left(\sum_{i=1}^{M}b_{\cdot i}\hat{\theta}_{(4i)} + \sum_{k=1}^{3}a_{ki}\hat{\theta}_{(ki)}\right)$.

4 Incorporating covariates in the selection mechanism

For some studies, such as the one on Alzheimer's Disease, there often exist baseline covariates $\boldsymbol{X}$, which may either inter-correlate with the test value T and true disease status D, or affect the selection for verification. We focus our discussion on the incorporation of discrete covariates with a finite total number of covariate patterns. The covariate-specific volume

under the ROC surface can be estimated by the method derived in the previous section to each pattern of $\boldsymbol{X}$. If the goal is to obtain a common index of accuracy across all covariate patterns, the method derived in the previous section needs to be modified to account for the possible covariate-dependent differential verification.

Assume that $\boldsymbol{X}$ comprises of P discrete covariates, each with N_j possible categories, for $j = 1, \ldots, P$, and we let $N = \prod_{j=1}^{P} N_j$ indicate the total number of different covariate patterns of $\boldsymbol{X}$. We further assume that $\boldsymbol{X}$ is a random sample from a discrete space $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$ with probabilities $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_N)$. The parameter π_{ik} in (3.3) is then given by

$$\pi_{ik} = \sum_{j=1}^{N} P(T = i, \boldsymbol{X} = \boldsymbol{X}_j \mid D = k) = \frac{\sum_{j=1}^{N} \tau_{ij}\phi_{kij}\delta_j}{\sum_{l=1}^{M}\sum_{j=1}^{N} \tau_{lj}\phi_{klj}\delta_j}, \tag{4.1}$$

where $\tau_{ij} = P(T = i \mid \boldsymbol{X} = \boldsymbol{X}_j)$, $\phi_{kij} = P(D = k \mid T = i, \boldsymbol{X} = \boldsymbol{X}_j)$, and $\delta_j = P(\boldsymbol{X} = \boldsymbol{X}_j)$, for $i = 1, \ldots, M$, $j = 1, \ldots, N$, and $k = 1, 2, 3$. Deviated from the previous derivation, the distributions of T and $[D \mid T]$ are now related to the covariates $\boldsymbol{X}$. With a similar ignorable assumption on the selection mechanism, except now in the presence of $\boldsymbol{X}$, we have $P(V|D, T, \boldsymbol{X}) = P(V|T, \boldsymbol{X})$, which can be shown leads to $P(D|T, \boldsymbol{X}) = P(D|T, \boldsymbol{X}, V = 1)$, equivalently $\phi_{kij} = P(D = k \mid T = i, \boldsymbol{X} = \boldsymbol{X}_j, V = 1)$.

In order to estimate θ in (3.2) and the coordinates of the empirical ROC surface in (3.1), both of which are functions of the π_{ik}'s, we need to obtain the ML estimates of τ_{ij}, ϕ_{kij}, and δ_j within each pattern of covariates. We note that for the observed data with $\boldsymbol{X} = \boldsymbol{X}_j$, the jth contingency table among random variables, V, D, and T can be formed, and the data structure is an analog of the observed data displayed in Table 1, except for an additional subscript j for all elements. The log-likelihood function can be derived as

$$l(\boldsymbol{\tau}, \boldsymbol{\phi}, \boldsymbol{\delta}) = \sum_{j=1}^{N} \left\{ \sum_{i=1}^{M} n_{ij} \log(\tau_{ij}) + \sum_{i=1}^{M}\sum_{k=1}^{3} a_{kij} \log(\phi_{kij}) + n_j \log(\delta_j) \right\}, \tag{4.2}$$

where n_{ij}, a_{kij}, and n_j are respectively the total number of patients with $T = i$ and $\boldsymbol{X} = \boldsymbol{X}_j$, the total number of verified patients with $D = k$, $T = i$ and $\boldsymbol{X} = \boldsymbol{X}_j$, and the total number

of patients with $\boldsymbol{X} = \boldsymbol{X}_j$. The ML estimates for parameters $\boldsymbol{\delta}$, $\boldsymbol{\tau}$, and $\boldsymbol{\phi}$ are

$$\hat{\delta}_j = \frac{n_j}{n}, \quad j = 1, \ldots, N-1, \tag{4.3}$$

$$\hat{\tau}_{ij} = \frac{n_{ij}}{n_j}, \quad i = 1, \ldots, M-1; \quad j = 1, \ldots, N, \tag{4.4}$$

$$\hat{\phi}_{kij} = \frac{a_{kij}}{a_{1ij} + a_{2ij} + a_{3ij}}, \quad k = 1, 2; \quad i = 1, \ldots, M; \quad j = 1, \ldots, N, \tag{4.5}$$

where $n = \sum_{j=1}^{N} n_j$. The ML estimates of π_{ik} can be obtained by replacing δ_j, τ_{ij} and ϕ_{kij} in (4.1) respectively by $\hat{\delta}_j$, $\hat{\tau}_{ij}$ and $\hat{\phi}_{kij}$.

For any given pair of decision thresholds, (d_1, d_2), the ML estimates of the three correct classification rates are then given by

$$\hat{P}(\hat{D} = k \mid D = k) = \begin{cases} 0, & \bar{d}_{k-1} > \underline{d}_k \\ \sum_{i=\bar{d}_{k-1}}^{\underline{d}_k} \hat{\pi}_{ik}, & \bar{d}_{k-1} \leq \underline{d}_k \end{cases}, \tag{4.6}$$

for $k = 1, 2, 3$. These three correct classification rates can then be used as coordinates in the construction of the empirical three-way ROC surface. Similarly, substituting the unknown parameters π_{ik} by their ML estimates in (3.2), we can then obtain the ML estimator $\hat{\theta}$ for the volume under the ROC surface. The corresponding variance estimator can be obtained by either Jackknife method or the Fisher's information method as described in the previous section.

5 Comparison between volumes under ROC surfaces

When comparing two diagnostic tests, one of the efficient designs is a paired design in which patients receive both tests for the diagnosis of the same disease. Since for a test, the volume under the ROC surface can be used to measure its accuracy, the difference between the two volumes naturally provides a means to assess the diagnostic discrepancy in accuracy. In order to account for possible correlation in addition to adjust for verification bias, we need to modify our derivations. We first assume test T_1 ranges from 1 to M_1, and test T_2 ranges

from 1 to M_2. The intermediate parameter η_{ijk} is given by

$$\eta_{ijk} = P(T_1 = i, T_2 = j | D = k) = \frac{\alpha_{ij}\psi_{kij}}{\sum_{s=1}^{M_1}\sum_{t=1}^{M_2}\alpha_{st}\psi_{kst}}, \tag{5.1}$$

where $\alpha_{ij} = P(T_1 = i, T_2 = j)$, and $\psi_{kij} = P(D = k | T_1 = i, T_2 = j)$. For $k = 1, 2, 3$, we then let $\pi_{ik1} = P(T_1 = i | D = k) = \sum_{j=1}^{M_2}\eta_{ijk}$, for $i = 1, \ldots, M_1$, and $\pi_{ik2} = P(T_2 = i | D = k) = \sum_{j=1}^{M_1}\eta_{jik}$, for $i = 1, \ldots, M_2$. Given π_{ik1}'s, the volume under the ROC surface for T_1 is

$$\theta_1 = \sum_{i=1}^{M_1-2}\sum_{j=i+1}^{M_1-1}\sum_{k=j+1}^{M_1}\pi_{i11}\pi_{j21}\pi_{k31} + \frac{1}{2}\left\{\sum_{i=1}^{M_1-1}\sum_{j=i+1}^{M_1}\left(\pi_{i11}\pi_{j21}\pi_{j31} + \pi_{i11}\pi_{i21}\pi_{j31}\right)\right\} + \frac{1}{6}\sum_{i=1}^{M_1}\pi_{i11}\pi_{i21}\pi_{i31}, \tag{5.2}$$

and given π_{ik2}'s, the volume under the ROC surface for T_2 is

$$\theta_2 = \sum_{i=1}^{M_2-2}\sum_{j=i+1}^{M_2-1}\sum_{k=j+1}^{M_2}\pi_{i12}\pi_{j22}\pi_{k32} + \frac{1}{2}\left\{\sum_{i=1}^{M_2-1}\sum_{j=i+1}^{M_2}\left(\pi_{i12}\pi_{j22}\pi_{j32} + \pi_{i12}\pi_{i22}\pi_{j32}\right)\right\} + \frac{1}{6}\sum_{i=1}^{M_2}\pi_{i12}\pi_{i22}\pi_{i32}, \tag{5.3}$$

The observed data with differential selection for definitive examination is shown in Table 2. The cross classification of T_1 and T_2 gives a total of M_1M_2 frequency counts for the group of verified patients. By assuming $P(V|D, T_1, T_2) = P(V|T_1, T_2)$, similar to the one in Section 3, we can obtain the observed log-likelihood as

$$l(\boldsymbol{\alpha}, \boldsymbol{\psi}) = \sum_{i=1}^{M_1}\sum_{j=1}^{M_2} m_{ij}\log(\alpha_{ij}) + \sum_{i=1}^{M_1}\sum_{j=1}^{M_2}\sum_{k=1}^{3} c_{kij}\log(\psi_{kij}), \tag{5.4}$$

where $\boldsymbol{\alpha} = [\alpha_{11}, \ldots, \alpha_{1M_2}, \alpha_{21}, \ldots, \alpha_{2M_2}, \ldots, \alpha_{M_11}, \ldots, \alpha_{M_1(M_2-1)}]$, $\boldsymbol{\psi_{ij}} = (\psi_{1ij}, \psi_{2ij})$, and $\boldsymbol{\psi} = (\boldsymbol{\psi}_{11}, \ldots, \boldsymbol{\psi}_{M_1M_2})$. The ML estimates for $\boldsymbol{\alpha}$ and $\boldsymbol{\psi}$, respectively, are

$$\hat{\alpha}_{ij} = \frac{m_{ij}}{m}, \quad i = 1, \ldots, M_1; \quad j = 1, \ldots, M_2, \tag{5.5}$$

$$\hat{\psi}_{kij} = \frac{c_{kij}}{c_{1ij} + c_{2ij} + c_{3ij}}, \quad k = 1, 2; \quad i = 1, \ldots, M_1; \quad j = 1, \ldots, M_2, \tag{5.6}$$

where $m = \sum_{i=1}^{M_1}\sum_{j=1}^{M_2} m_{ij}$. Subsequently substituting the unknown parameters α_{ij} and ψ_{kij} by their ML estimates to obtain estimates $\hat{\eta}_{ijk}$, and $\hat{\pi}_{ik1}$ and $\hat{\pi}_{ik2}$, we can ultimately compute the ML estimates $\hat{\theta}_1$ from (5.2), and $\hat{\theta}_2$ from (5.3).

The observed Fisher information matrix defined on $(\boldsymbol{\alpha}, \boldsymbol{\psi})$ is a block diagonal matrix given by

$$I(\boldsymbol{\alpha}, \boldsymbol{\psi}) = \text{diag}(I_1(\boldsymbol{\alpha}), I_{211}(\boldsymbol{\psi_{11}}), \ldots, I_{2M_1M_2}(\boldsymbol{\psi_{M_1M_2}})),$$

where $I_1(\boldsymbol{\psi})$ and $I_{2ij}(\boldsymbol{\psi_{ij}})$ are respectively the observed Fisher information matrix on the log-likelihood $l_1(\boldsymbol{\alpha}) = \sum_{i=1}^{M_1}\sum_{j=1}^{M_2} m_{ij}\log(\alpha_{ij})$ and $l_{2ij}(\boldsymbol{\psi}_{ij}) = \sum_{k=1}^{3} c_{kij}\log(\psi_{kij})$. By the Delta method, we can estimate the variance-covariance matrix $\Sigma_{\hat{\boldsymbol{\theta}}}$ of $\hat{\boldsymbol{\theta}} = [\hat{\theta_1}, \hat{\theta_2}]^T$ by

$$\frac{\partial \boldsymbol{\theta}^T}{\partial \boldsymbol{\alpha}} I_1^{-1}(\boldsymbol{\alpha}) \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\alpha}} + \frac{\partial \boldsymbol{\theta}^T}{\partial \boldsymbol{\psi}_{11}} I_{211}^{-1}(\boldsymbol{\psi}_{11}) \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\psi}_{11}} + \ldots + \frac{\partial \boldsymbol{\theta}^T}{\partial \boldsymbol{\psi}_{M_1M_2}} I_{2M_1M_2}^{-1}(\boldsymbol{\psi}_{M_1M_2}) \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\psi}_{M_1M_2}}, \qquad (5.7)$$

evaluated at $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\psi}_{ij} = \hat{\boldsymbol{\psi}_{ij}}$, where $I_1^{-1}(\boldsymbol{\alpha})$, $I_{2ij}^{-1}(\boldsymbol{\psi}_{ij})$, $\partial\boldsymbol{\theta}/\partial\boldsymbol{\alpha}$ and $\partial\boldsymbol{\theta}/\partial\boldsymbol{\psi}_{ij}$ are given in the Appendix. Another alternative way to estimate $\Sigma_{\hat{\boldsymbol{\theta}}}$ is by the Jackknife method. With the exclusion of a single patient, the variance-covariance matrix can be estimated as

$$\hat{\Sigma}_{\hat{\boldsymbol{\theta}}} = \frac{n-1}{n}\left[\sum_{i=1}^{M_1}\sum_{j=1}^{M_2}\left\{d_{\cdot ij}(\hat{\boldsymbol{\theta}}_{(4ij)} - \hat{\boldsymbol{\theta}}_{(\cdot)})(\hat{\boldsymbol{\theta}}_{(4ij)} - \hat{\boldsymbol{\theta}}_{(\cdot)})^T + \sum_{k=1}^{3} c_{kij}(\hat{\boldsymbol{\theta}}_{(kij)} - \hat{\boldsymbol{\theta}}_{(\cdot)})(\hat{\boldsymbol{\theta}}_{(kij)} - \hat{\boldsymbol{\theta}}_{(\cdot)})^T\right\}\right],$$

where $\hat{\boldsymbol{\theta}}_{(kij)}$ is the estimate of $\boldsymbol{\theta} = [\theta_1, \theta_2]^T$ after deleting a patient with $V = 1$, $T_1 = i$, $T_2 = j$, and $D = k$, for $k = 1, 2, 3$, $\hat{\boldsymbol{\theta}}_{(4ij)}$ is the estimate of $\boldsymbol{\theta}$ after deleting a patient with $V = 0$, $T_1 = i$, $T_2 = j$, and $\hat{\boldsymbol{\theta}}_{(\cdot)} = (1/n)\left(\sum_{i=1}^{M_1}\sum_{j=1}^{M_2} d_{\cdot ij}\hat{\boldsymbol{\theta}}_{(4ij)} + \sum_{k=1}^{3} c_{kij}\hat{\boldsymbol{\theta}}_{(kij)}\right)$.

A test of statistical significance of diagnostic accuracy difference can then be calculated as $z = \boldsymbol{c}^T\hat{\boldsymbol{\theta}}/\sqrt{\boldsymbol{c}^T\hat{\Sigma}_{\hat{\boldsymbol{\theta}}}\boldsymbol{c}}$, where the contrast $\boldsymbol{c} = [1, -1]^T$. Given the asymptotic normality of the ML estimate $\hat{\boldsymbol{\theta}}$, the test statistic z asymptotically follows a standard normal distribution.

6 The analysis of NACC data

We now analyze the NACC dataset to evaluate the accuracy of using Mini-Mental State Examination and clinical diagnosis of dementia in assessing AD severity. Figure 2 displays the empirical ROC surfaces for dMMSE and CDD, after accounting for the verification bias. The graphs were drawn with respect to the three correct classification rates, across all

possible decision thresholds. We note that both plots are bounded in a unit cube, and both surfaces contain the points of $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$.

Based on dMMSE scores, our estimate of the volume under the ROC surface was 0.4027 with standard deviation of 0.0104 by both the information and Jackknife approach. Without account for differential verification, the estimate, proposed by Nakas and Yiannoutsos (2004), based only on verified patients was 0.4035 with bootstrap (size=100) standard deviation of 0.0108. Only a slight difference was present between the two estimates, which might be due to a small discrepancy between center-specific estimates with and without considering verification bias, as shown in Figure 3. Our estimate was more efficient due to the account for both verified and unverified cases in the analysis. For CDD, the discrepancy between the two estimates became noticeable. Our estimate of the ROC surface volume was 0.3707 with standard deviation of 0.0091 by both the information and Jackknife approach, while the estimate ignoring unverified patients was 0.3503 with standard deviation of 0.0094. For both the analyses of dMMSE and CDD, our estimates of standard deviation are quite close between the information and Jackknife approach. By comparing the point estimates, dMMSE seems to have a better power than CDD in assessing the severity of AD, after accounting for verification bias.

To properly infer and compare the accuracy between dMMSE and CDD, we then worked on the joint cross-classification table by the two tests and verification status. This observed cross table implicitly permits the incorporation of any correlation structure between dMMSE and CDD (due to a paired design) in the calculation of our statistics. After accounting for differential verification, the difference between the two volumes under the ROC surfaces was 0.0124 with 95% confidence limits of (-0.0089, 0.0337) by the information method and the confidence limits of (-0.0091, 0.0339) by the Jackknife approach. There was no significant discrepancy in the accuracy between dMMSE and CDD, and the same conclusion was drawn by the two proposed approaches for variance estimation. We note that, despite the unified

MMSE procedure may have shown a competitive power in assessing the severity of AD, the clinical implication based on CDD may still be indispensable. The diagnosis based on CDD not only guides the assessment of AD severity, but is aimed to direct the etiology of dementia to other variants.

Finally, in order to account for center differences in verification selection and in the correlation between the test and autopsy examination, we include the AD center as a covariate in the evaluation of dMMSE and CDD. The center-specific estimates of the volumes under the ROC surfaces were drawn in the first row of Figure 3. There was clear variation across the eight research centers in the accuracy of using dMMSE scores to assess AD severity, and the differences among centers became less prominent in the assessment of CDD. An overall evaluation of CDD, by combining the empirical information from all centers, may reasonably be used to achieve a higher efficiency in the estimate as compared to the center-specific ones.

After adjusting for the center effect, the volume under the ROC surface was estimated as 0.3961 (with standard deviation of 0.0120) for dMMSE and 0.3932 (with standard deviation of 0.0117) for CDD. The volume estimate for dMMSE was a bit inflated if center differences in verification were ignored, while on the other hand, the estimate for CDD was reduced if the center effect was overlooked. Based on the asymptotic property of ML estimates, the 95% confidence limits for volumes under the ROC surfaces, after adjusting for the center effect, were (0.3726, 0.4196) and (0.3702, 0.4161), respectively for dMMSE and CDD. The exclusion of the null value of 1/6 in both confidence limits suggests that both dMMSE and CDD possess discriminatory power in assessing the severity of AD.

7 Simulation study

We conducted sets of simulation study to examine the performance of the proposed method. The data were first generated as follows. We simulated a 5-point rating scale test measurement for each of the three disease states. The distributions of test categories can differ across

different disease status, and the five different probabilities were chosen and given in Table 3, aimed to have diverse true volumes under the ROC surfaces. After generating the complete data, we then randomly chose a subgroup of patients and set their disease status as unknown. The selection probability was chosen to be an increasing function of a test result. The selection probabilities for verification are (1) $P(V = 1|T = 1) = 0.4$, (2) $P(V = 1|T = 2) = 0.6$, (3) $P(V = 1|T = 3) = 0.7$, (4) $P(V = 1|T = 4) = 0.8$, (5) $P(V = 1|T = 5) = 0.9$.

For each run, we assumed a sample of 300 patients for each disease status, and the results were summarized over 1,000 replications. Table 4 displays the true volumes under the ROC surfaces, and the performance of the proposed estimate. The bootstrap approach proposed by Nakas and Yiannoutsos (2004), without account for differential verification, was also implemented with 100 bootstrap samples for the purpose of comparison. It is clear to see in Table 4, that our estimates were up to 5.4% more accurate than naive estimates ignoring non-verified cases, and the improvement remained in four of the five different parameter settings. When distributions of test scores were the same among the three disease groups (no diagnostic power), the bias of ignoring differential verification was gone, as expected. The 95% coverage of confidence intervals was almost successfully retained by both the Fisher's information and Jackknife approach of standard deviation estimation. A poor coverage appeared when the selection mechanism was ignored. We further compared the proposed information and Jackknife approach for the estimation of standard deviation with the empirical estimates. The Jackknife estimates were consistently closer to the empirical estimates, suggesting that standard deviation estimated by Jackknife method may be better than the one estimated by the information method. Consistently larger estimates in standard deviation by the information approach also resulted in a wider coverage in confidence intervals, as compared to the one by the Jackknife method.

To further investigate the validity and robustness of the proposed approach against the violation of the MAR verification, we modified the selection probabilities to have them vary

with the true disease states. We had $P(V = 1|T = i, D = k) = 0.3 + 0.1i + 0.05k$, for $i = 1, \ldots, 5$ and $k = 1, 2, 3$. The results were summarized in Table 5 over 1,000 replications. We observed that the absolute biases increased when the MAR assumption was violated; nevertheless, our estimates remained to be more accurate than the naive estimates. The coverage of 95% confidence intervals was not affected very much by this non-ignorable verification. This simulation supports that our estimates remain valid and robust if the MAR assumption is moderately violated.

For the comparison between volumes under two related ROC surfaces, we extended the simulation setups to two ordinal-scale tests. We supposed that a disease could be diagnosed either by a 5-point rating test tool T_1, or by another 4-point rating test measurement T_2. We chose five different parameter settings for the joint probability of the two test scores to achieve a diverse spectrum in the true volumes under the ROC surfaces and their differences. The joint probabilities were specified for each individual cell in the cross table analog to Table 2 as $V = 1$. We assumed that the chance for the definite verification depends on the joint distribution of the two test scores, namely, $P(V = 1|T_1 = i, T_2 = j) = 0.64$, for $i, j \leq 2$, $P(V = 1|T_1 = i, T_2 = j) = 0.81$, for $i, j > 2$, and $P(V = 1|T_1 = i, T_2 = j) = 0.72$ otherwise. The results were summarized over 1,000 replications in Table 6, with the size of 1,000 patients for each of the three disease groups. In Table 6, each row corresponds to a specific distributional setup for T_1 and T_2, and the true volumes under the ROC surfaces, θ_1 and θ_2, were designed to differ across rows. It was clear to see that biases in the difference between volumes were small in all five settings. The coverage of confidence intervals was satisfactorily around 95% by both the information and Jackknife method for the estimation of the standard deviation.

Several other selection probability and sample sizes had also been chosen in our simulation, but were skipped for presentation, due to the similar conclusion. We only note that biases induced by ignoring non-verified cases in the analysis increased with decreased

verification rates.

8 Discussion

In this paper, the ML estimate of the volume under the ROC surface is derived when definitive disease verification is subject to selection. The verification bias induced by differential selection is numerically demonstrated in the simulation over different diagnostic abilities. Under the consideration of ignorable verification, this bias can be adjusted in the framework of likelihood principle, and the standard deviation for the volume estimate can be obtained from either the Fisher's information or Jackknife method. There are limitations in the proposed approach. Like chi-square tests for the analysis of contingency tables, our estimate is not suitable to handle sparse data, particularly when a zero number of verified cases is present for some test measurements. The standard deviation estimated by the information method would fail if the observed cross table between D and T, e.g. Table 1, for verified cases has any zero entry. Similar limitations apply to the extension to the comparison between two diagnostic tools. It requires a nonzero number of verified cases for each cross classification of T_1 and T_2 in Table 2. A more stringent requirement of all positive c_{kij}'s is needed to carry out the variance estimation by the information method.

We applied the proposed method to the largest national database on Alzheimer's disease to compare the relative accuracy between Mini-Mental State Examination and clinical evaluation of dementia in grading the severity of Alzheimer's disease. We found an interesting result that Mini-Mental State Examination and clinical evaluation of dementia had similar accuracy in predicting neuropathological severity of Alzheimer's disease, even though the clinical evaluation of dementia was much more expensive and required more clinical information. The results were drawn from the recruitment of the eight ADCs across the United States, and may need further investigation by community-based studies. Nevertheless, the clinical evaluation of dementia still has important implication in guiding the diagnosis of

dementia to other variants, such as Lewy body disease and vascular dementia.

The proposed methodology may be easily modified to diagnostic problems with more than three disease classes. There are some issues that remain open for future investigation. For instance, in the presence of verification bias, how can the accuracy of a continuous test measurement be evaluated? Rotnitzky et al. (2006) developed a doubly robust estimation for the area under the ROC curve that adjusts for selection to verification for markers measured on any scale. The extension of their methodology to the volume under the ROC surface may be helpful in diagnostic medicine when the severity of the disease is of interest. The comparison of accuracy between tests measured in different scales would be another direction to pursue. Furthermore, the relaxation of the assumption about ignorable verification by presuming the selection to verification directly associated with the true disease status, may be more plausible as adjusting for verification bias.

Acknowledgement

We thank for the data and scientific support provided by the National Alzheimer's Coordinating Center (NACC). Both Chi and Zhou's work was supported by grants from NACC (U01 AG16976) and from the National Institute of Health (R01 EB005829).

9 References

Agresti, A. (1990) *Categorical data analysis*, New York: John Wiley & Sons.

Alonzo, T. A. and Pepe, M. S. (2005) Assessing accuracy of a continuous screening test in the presence of verification bias, *Applied Statistics*, **54**, 173-190.

Begg, C. B. and Greenes, R. A. (1983) Assessment of diagnostic tests when disease verification is subject to selection bias, *Biometrics*, **39**, 207-216.

Dreiseitl, S., Ohno-Machado, L., and Binder, M. (2000) Comparing three-class diagnostic tests by three-way ROC analysis, *Medical Decision Making*, **20**, 323-331.

Efron, B. and Tibshirani, R. J. (1993) *An introduction to the Bootstrap*, New York: Chapman & Hall.

Folstein M., Folstein S., and McHugh S. (1975) Mini-mental state. A practical method for grading the cognitive state of patients for the clinician, *Journal of Psychiatric Research*, **12**, 189-198

Gray, R., Begg, C. B., and Greenes, R. A. (1984) Construction of receiver operating characteristic curves when disease verification is subject to selection bias, *Medical Decision Making*, **4**, 151-164.

Harel, O. and Zhou, X. H. (2007) Multiple imputation for correcting verification bias, *Statistics in Medicine*, in press.

Just, S. (2004) Namenda (Memantine) for Moderate-to-Severe Alzheimer's Disease, *Pharmacotherapy update newsletter*, Cleveland Clinic at http://www.clevelandclinicmeded.com/.

Kosinski, A. S. and Barnhart, H. X. (2003) Accounting for nonignorable verification bias in assessment of diagnostic tests, *Biometrics*, **59**, 163-171.

Mossman, D. (1999), Three-way ROCs *Medical Decision Making*, **19**, 78-89.

Nakas, C. T. and Yiannoutsos, C. T. (2004) Ordered multiple-class ROC analysis with continuous measurements, *Statistics in Medicine*, **23**, 3437-3449.

Nakas, C. T. and Yiannoutsos, C. T. (2006) Ordered multiple-class ROC analysis, *Encyclopedia of Biopharmaceutical Statistics*, DOI: 10.1081/E-EBS-120041740.

Obuchowski, N. A. (2005) Estimating and comparing diagnostic tests' accuracy when the gold standard is not binary, *Academic Radiology*, **12**, 1198-1204.

Reisberg B., Doody R., Stoffler A., Schmitt F., Ferris S., and Mobius H. J. (2003) Memantine in moderate-to-severe Alzheimer's disease, *New England Journal of Medicine*, **348**, 1333-1341.

Rotnitzky, A., Faraggi, D., and Schisterman, E. (2006) Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verification bias, *Journal of the American Statistical Association*, **101**, 1276-1288.

Rubin, D. B. (1976) Inference and missing data, *Biometrika*, **63**, 581-592.

Scurfield, B. K. (1996) Multiple-event forced-choice tasks in the theory of signal detectability, *Journal of Mathematical Psychology*, **40**, 253-269.

Zhou, X. H. (1996) A nonparametric maximum likelihood estimator for the receiver operation characteristic curve area in the presence of verification bias, *Biometrics*, **52**, 299-305.

Zhou, X. H., Obuchowski, N. A., and McClish, D. K. (2002) *Statistical methods in diagnostic medicine*, New York: John Wiley & Sons.

Xiong, C., van Belle, G., Miller, J. P., and Morris, J. C. (2006) Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups, *Statistics in Medicine*, **25**, 1251-1273.

Table 1: Observed data of an ordinal-scaled test data

		$T=1$	$T=2$	$\cdots$	$T=M$
$V=1$	$D=1$	a_{11}	a_{12}	$\cdots$	a_{1M}
	$D=2$	a_{21}	a_{22}	$\cdots$	a_{2M}
	$D=3$	a_{31}	a_{32}	$\cdots$	a_{3M}
$V=0$		$b_{\cdot 1}$	$b_{\cdot 2}$	$\cdots$	$b_{\cdot M}$
Total		n_1	n_2	$\cdots$	n_M

Table 2: Observed data of the two ordinal-scaled test data.

		$T_1=1$			$\cdots$	$T_1=M_1$		
		$T_2=1$	$\cdots$	$T_2=M_2$	$\cdots$	$T_2=1$	$\cdots$	$T_2=M_2$
$V=1$	$D=1$	c_{111}	$\cdots$	c_{11M_2}	$\cdots$	c_{1M_11}	$\cdots$	$c_{1M_1M_2}$
	$D=2$	c_{211}	$\cdots$	c_{21M_2}	$\cdots$	c_{2M_11}	$\cdots$	$c_{2M_1M_2}$
	$D=3$	c_{311}	$\cdots$	c_{31M_2}	$\cdots$	c_{3M_11}	$\cdots$	$c_{3M_1M_2}$
$V=0$		$d_{\cdot 11}$	$\cdots$	$d_{\cdot 1M_2}$	$\cdots$	$d_{\cdot M_11}$	$\cdots$	$d_{\cdot M_1M_2}$
Total		m_{11}	$\cdots$	m_{1M_2}	$\cdots$	m_{M_11}	$\cdots$	$m_{M_1M_2}$

Table 3: Probabilities for test scores in the three disease groups.

	$P(T=1,T=2,T=3,T=4,T=5)$		
	$D=1$	$D=2$	$D=3$
I	(0.20, 0.20, 0.20, 0.20, 0.20)	(0.20, 0.20, 0.20, 0.20, 0.20)	(0.20, 0.20, 0.20, 0.20, 0.20)
II	(0.30, 0.30, 0.20, 0.10, 0.10)	(0.10, 0.20, 0.25, 0.25, 0.20)	(0.05, 0.05, 0.20, 0.30, 0.40)
III	(0.50, 0.20, 0.20, 0.05, 0.05)	(0.10, 0.25, 0.30, 0.25, 0.10)	(0.05, 0.05, 0.20, 0.20, 0.50)
IV	(0.80, 0.05, 0.05, 0.05, 0.05)	(0.05, 0.10, 0.70, 0.10, 0.05)	(0.05, 0.05, 0.05, 0.05, 0.80)
V	(0.95, 0.02, 0.01, 0.01, 0.01)	(0.02, 0.03, 0.90, 0.03, 0.02)	(0.01, 0.01, 0.01, 0.02, 0.95)

Table 4: Volume under a single ROC surface with MAR verification, averaged over 1,000 replications.

		Our estimates						Naive estimates	
		Absolute	Standard deviation			% CI coverage		Absolute	% CI
	θ	bias	Empirical	Delta	Jackknife	Delta	Jackknife	bias	coverage
I	0.1667	0.0010	0.0157	0.0167	0.0166	96.4	96.4	0.0010	95.7
II	0.3903	0.0012	0.0233	0.0248	0.0233	96.4	95.2	0.0217	80.8
III	0.5164	0.0008	0.0249	0.0277	0.0252	97.2	95.4	0.0196	88.0
IV	0.7270	0.0001	0.0276	0.0296	0.0273	96.3	95.0	0.0392	76.7
V	0.9312	0.0004	0.0166	0.0202	0.0160	98.5	93.1	0.0122	90.8

Table 5: Volume under a single ROC surface with non-ignorable verification, averaged over 1,000 replications.

		Our estimates						Naive estimates	
		Absolute	Standard deviation			% CI coverage		Absolute	% CI
	θ	bias	Empirical	Delta	Jackknife	Delta	Jackknife	bias	coverage
I	0.1667	0.0051	0.0153	0.0156	0.0154	93.3	92.9	0.0047	92.6
II	0.3903	0.0032	0.0228	0.0242	0.0226	96.1	95.1	0.0195	85.3
III	0.5164	0.0058	0.0247	0.0271	0.0245	96.1	93.7	0.0145	89.9
IV	0.7270	0.0059	0.0266	0.0291	0.0268	96.6	95.5	0.0263	85.9
V	0.9312	0.0029	0.0164	0.0202	0.0162	98.5	94.6	0.0090	92.5

Table 6: The numerical comparison between two ROC surfaces, with $\Delta_\theta = \theta_1 - \theta_2$.

			Absolute bias	Delta method		Jackknife	
θ_1	θ_2	Δ_θ	for Δ_θ	CI coverage %	CI length	CI coverage %	CI length
0.1667	0.1667	0.0000	0.0003	94.6	0.0441	94.7	0.0442
0.1667	0.4680	-0.3013	0.0005	95.9	0.0573	95.9	0.0575
0.1667	0.7693	-0.6026	0.0004	95.8	0.0578	95.9	0.0580
0.4693	0.4680	0.0013	0.0006	94.1	0.0711	94.2	0.0713
0.4693	0.7693	-0.3000	0.0004	95.8	0.0710	96.0	0.0714

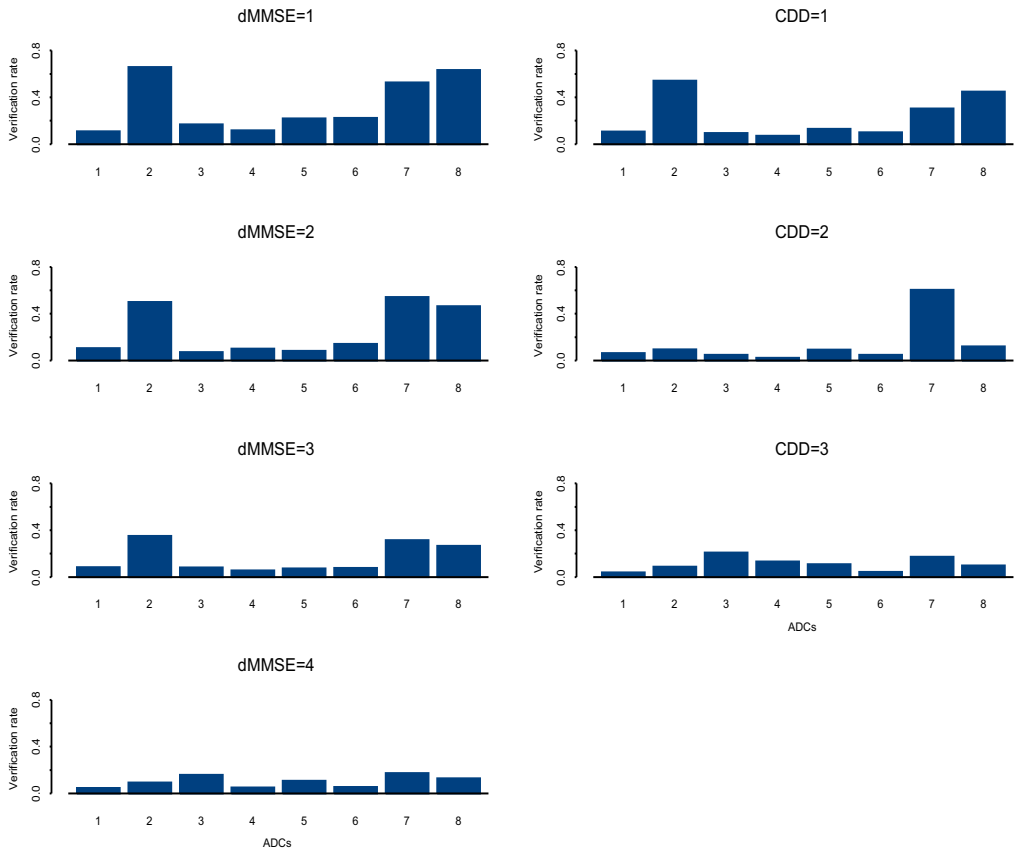

Figure 1: The selection rates for verification across eight ADCs and categories by dMMSE (left) and CDD (right).

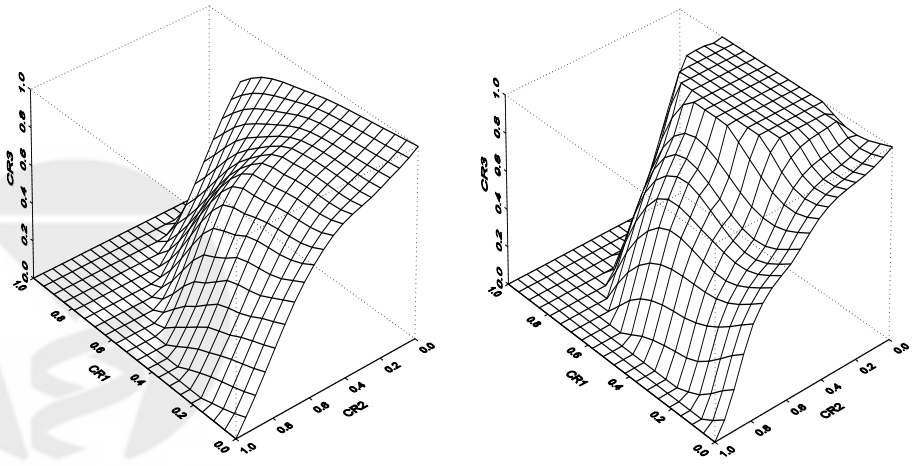

Figure 2: The empirical ROC surfaces of dMMSE (left) and CDD (right), after accounting for verification bias. The three axes are the three correct classification rates.

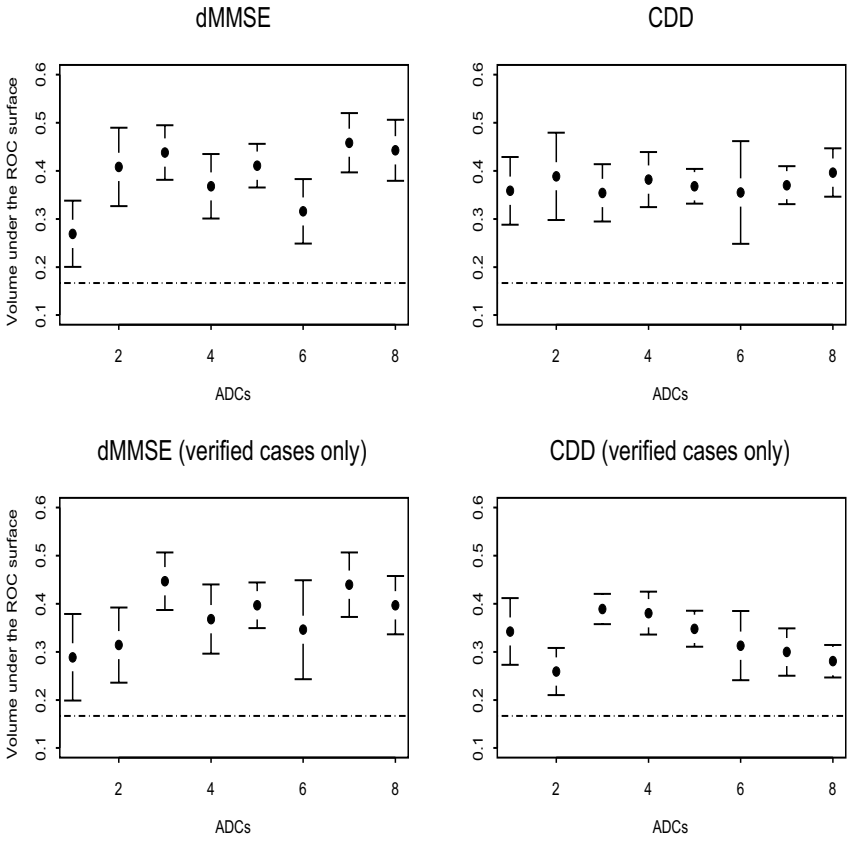

Figure 3: The means and 95% asymptotic confidence intervals of the center-specific ROC volume estimates for dMMSE and CDD. The top two graphs were with account for verification bias, while the bottom two used only verified cases. The dashed line is the null line of no discrimination power.