Inverse probability weighting estimation of the volume under the ROC surface in the presence of verification bias

Ying Zhang* and **Todd A. Alonzo for the Alzheimer's Disease Neuroimaging Initiative**[†]

Department of Biostatistics, University of Southern California, Keck School of Medicine,
Los Angeles, California 90033, USA

Received 22 October 2015; revised 28 February 2016; accepted 10 March 2016

In diagnostic medicine, the volume under the receiver operating characteristic (ROC) surface (VUS) is a commonly used index to quantify the ability of a continuous diagnostic test to discriminate between three disease states. In practice, verification of the true disease status may be performed only for a subset of subjects under study since the verification procedure is invasive, risky, or expensive. The selection for disease examination might depend on the results of the diagnostic test and other clinical characteristics of the patients, which in turn can cause bias in estimates of the VUS. This bias is referred to as verification bias. Existing verification bias correction in three-way ROC analysis focuses on ordinal tests. We propose verification bias-correction methods to construct ROC surface and estimate the VUS for a continuous diagnostic test, based on inverse probability weighting. By applying U-statistics theory, we develop asymptotic properties for the estimator. A Jackknife estimator of variance is also derived. Extensive simulation studies are performed to evaluate the performance of the new estimators in terms of bias correction and variance. The proposed methods are used to assess the ability of a biomarker to accurately identify stages of Alzheimer's disease.

Keywords: Diagnostic test; Inverse probability weighting; Missing at random; Verification bias; VUS.

Additional supporting information including source code to reproduce the results may be found in the online version of this article at the publisher's web-site

1 Introduction

Diagnostic tests are performed to provide information about the diagnosis or detection of diseases, track disease progression, and understand disease mechanisms (McNeil and Adelstein, 1976; Sox et al., 1988). More than two disease classes are often necessary in many important biomedical diagnostic tasks. Take the Alzheimer's Disease Neuroimaging Initiative (ADNI) study as an example, ADNI investigators used neuropsychological tests to classify disease stages into probable Alzheimer's disease (AD), amnestic mild cognitive impairment (MCI), and normal cognition aging (Misra et al., 2009). One of the goals of ADNI is to develop a new biomarker test that would be less costly substitute for neuropathological findings. The availability of suitable biomarkers tracking the stages of AD progression could markedly accelerate drug development by providing an earlier indication of drug

*Corresponding author: e-mail: janezhang57a@gmail.com, Phone: +1-626-379-4581

[†] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

efficacy. Thus it is critical to develop proper statistical methods to evaluate the diagnostic accuracy of these new biomarkers.

The accuracy of a new diagnostic test can be assessed by comparing it with the gold standard reference test (GST). In practice, often not all patients undergo the definitive disease verification because the GST may be invasive, risky, or expensive. As a result, only a subset of the subjects who received the new test will further have true disease status verified. In the ADNI study, definite diagnosis requires autopsy confirmation. Therefore, the true disease status will be missing because living subjects cannot have disease verified, and furthermore deceased subjects do not always have autopsy performed. On the other hand, the alternative neuropsychological approach is too expensive for regular use. When the decision regarding disease verification depends on the new test results or possibly other clinical characteristics, estimators of the accuracy may be biased when they are only based on data from patients with verified disease status. This bias is usually referred to as verification bias or work up bias (Ransohoff and Feinstein, 1978; Begg and Greenes, 1983).

Receiver operating characteristic (ROC) curves and the summary measure area under the ROC curve (AUC) are the standard approaches for assessing the accuracy of a new diagnostic tests to distinguish between two disease states (e.g., diseased vs. nondiseased individuals). ROC analysis has been extended to accommodate the situation in which the diagnostic task is to classify a case into one of three possible classes (Scurfield, 1996; Mossman, 1999). Specifically, ROC surfaces and the corresponding summary measure volume under the surface (VUS) have been proposed to evaluate the overall performance of the diagnostic test, which represents the capability of the test in correct classification of subjects into three different ordinal disease groups.

Several methods have been proposed for correcting verification bias when the diagnostic task is dichotomous and the new test is continuous (Alonzo et al., 2003; Alonzo and Pepe, 2005; Rotnitzky et al., 2006; He et al., 2009). One approach that has been proposed is to use inverse probability weighting (IPW) where test results for those with disease verification are reweighted by the inverse of the probability of disease verification. Thus far, verification bias correction methods for ROC surface analysis have only been developed for ordinal biomarkers. Specifically, a nonparametric likelihood-based approach has been developed by Chi and Zhou to construct the empirical bias-corrected ROC surface for an ordinal diagnostic test using the missing at random (MAR) assumption (Chi and Zhou, 2008). MAR is the key assumption that has been made in many methods for verification bias correction (Little and Rubin, 1987). The verification process is MAR if the probability of disease verification is purely determined by the test result and other observed clinical covariates. Thus, the probability of receiving disease verification is conditionally independent of unknown true disease status given observed variables. Since their method requires test results can only take finite values ranging from 1 to M, it cannot be applied to a continuous test such as the biomarkers being developed for AD. Therefore, it is imperative to develop proper methods for evaluation of the accuracy of a continuous diagnostic test accounting for the presence of verification bias. Here, we propose IPW estimators of true classification rates (TCRs) that can be used to construct the ROC surface and estimate VUS that correct for verification bias when assessing the accuracy of continuous tests in three-class disease classification problems.

The rest of the paper is organized as follows. In Section 2, we give a brief review of ROC surface construction and estimating VUS with full data. Based on IPW methodology, we propose a new bias-corrected estimator of the ROC surface and VUS in Section 3 that are applicable to continuous diagnostic tests. In Section 4, we present an extensive simulation study, and then in Section 5 we apply the proposed method to the motivating study on AD. The paper concludes with a discussion.

2 Three-class ROC analysis with full data

Parametric and nonparametric approaches have been developed to evaluate accuracy of a diagnostic test in the three-class setting when disease status is available for all subjects under study (Dreiseitl et al., 2000; Heckerling, 2001; Alonzo and Nakas, 2007; Nakas and Alonzo, 2007; Alonzo et al., 2009; Li and Zhou, 2009; Nakas et al., 2010). In this section we provide a brief summary of the nonparametric approaches.

Suppose n subjects are chosen randomly from the target population to assess the accuracy of a diagnostic test. Let T_i denote the continuous result of an investigational diagnostic test, and let D_i denote the true disease status for the i-th subject, $i = 1, \ldots, n$. $D_i = 0$ indicates that the subject does not have the disease, $D_i = 1$ indicates that the subject has mild disease, and $D_i = 2$ indicates that the subject has severe disease. A higher value of T corresponds to a higher value of D.

Since there are three different disease categories to be determined, using the decision rules used by Mossman (1999), two-ordered decision boundaries $c1 < c2$ are selected and the diagnostic decision is made as following:

1. if $T_i \le c1$, then $D_i = 0$: nondisease subject,
2. if $T_i > c1$ and $T_i < c2$, then $D_i = 1$: mild-disease subject,
3. if $T_i \ge c2$, then $D_i = 2$: severe-disease subject,

where T_i and D_i stand for test results and disease status of the i-th subject, respectively. For each pair of ordered thresholds $(c1, c2)$, three TCRs can be generated. The ROC surface plot is made by plotting TCR^0, TCR^1, and TCR^2 in the three-dimensional space for all possible pairs of thresholds $(c1, c2) \in \mathbb{R}^2$.

As a generalization of the AUC for summarizing an ROC curve under a binary diagnostic task, VUS is defined as $P(T_i > T_j > T_k \mid D_i > D_j > D_k)$, where i, j, k are randomly selected subjects from nondisease, mild-disease, and severe-disease subjects, respectively. Thus, a value of 1 corresponds to a perfect test, while a value of 1/6 indicates the test is no better than random chance.

If all of the subjects are selected to obtain disease verification, Nakas and Yiannoutsos (2004) nonparametrically estimated TCRs for each disease group by using

1. $D_i = 0 : \widehat{TCR}^0(c1) = \frac{\sum_{i=1}^n I(T_i \le c1)\, I(D_i = 0)}{\sum_{i=1}^n I(D_i = 0)}$
2. $D_i = 1 : \widehat{TCR}^1(c1,\ c2) = \frac{\sum_{i=1}^n I(c1 < T_i < c2)\, I(D_i = 1)}{\sum_{i=1}^n I(D_i = 1)}$
3. $D_i = 2 : \widehat{TCR}^2(c2) = \frac{\sum_{i=1}^n I(T_i \ge c2)\, I(D_i = 2)}{\sum_{i=1}^n I(D_i = 2)}$.

VUS can then be estimated as

$$\widehat{VUS} = \frac{\sum_{i=1}^n \sum_{i=1}^n \sum_{i=1}^n I(T_i < T_j < T_k)\, I(D_i < D_j < D_k)}{\sum_{i=1}^n \sum_{i=1}^n \sum_{i=1}^n I(D_i < D_j < D_k)}.$$

Let V_i denote the verification status for the i-th subject, where $i = 1, \ldots, n$. $V_i = 1$ if the subject has the true disease status verified, and $V_i = 0$ otherwise. If only a subset of subjects are selected to obtain disease verification, the naïve estimator of VUS is obtained as

$$\widehat{VUS}_{\text{naïve}} = \frac{\sum_{i=1}^n \sum_{i=1}^n \sum_{i=1}^n I(T_i < T_j < T_k)\, I(D_i < D_j < D_k)\, V_i V_j V_k}{\sum_{i=1}^n \sum_{i=1}^n \sum_{i=1}^n I(D_i < D_j < D_k)\, V_i V_j V_k}.$$

This naïve estimator is unbiased only if the subjects are completely randomly selected for disease verification. Otherwise, this naïve estimator is biased. This estimator is also referred to as a complete case estimator because it only uses data from subjects with true disease status verified without bias correction.

When the diagnostic task is binary, estimators for sensitivity and specificity of a diagnostic test were derived by incorporating methodologies including the IPW approach that reweight observed data by the inverse of the probability of disease verification (Alonzo and Pepe, 2005). The bias-corrected ROC curve can be made by plotting bias-corrected sensitivity versus bias-corrected (1-specificity) for all possible thresholds. By applying trapezoidal rule for integration, AUC can be empirically estimated from a bias-corrected ROC curve. Alternatively, He et al. achieved direct estimation of the AUC based on U-statistics and the IPW approach (He et al., 2009). Based on the MAR assumption, in the next section, we use IPW methodology to construct a bias-corrected ROC surface and propose a new estimator of bias-corrected VUS when the test results are continuous.

3 IPW estimation of ROC surface

Let $\boldsymbol{A_i}$ be a vector of observed covariates for the subject that may be informative about D_i. Under the defined setting, the observed data are composed of n independent and identical distributed (i.i.d) samples $\boldsymbol{S_i}$ of the random vector $\boldsymbol{S_i} = (V_i, D_i, T_i, \boldsymbol{A_i})$. $(T_i, V_i, \boldsymbol{A_i})$ is observed for each subject, but D_i is observed only if $V_i = 1$. Here we also make the MAR assumption, that is, $V \perp\!\!\!\perp D \,|\, (T, A)$ (ignorable missingness). Let $\lambda_0, \lambda_1, \lambda_2$ be the prevalence for the three disease subgroups and let $\pi_i = Pr(V_i = 1|T_i, \boldsymbol{A_i})$ be the verification probability. We further define $F_d(t)$ to be the corresponding distribution function of the new test result T for subjects with $D = d, d = 0, 1, 2$. Finally let $G_d(t, a)$ be the corresponding joint distribution function of new test result T and covariates A for subjects with $D = d, d = 0, 1, 2$.

3.1 ROC surface

Based on the MAR assumption, we construct bias-corrected ROC surface by reweighting each observation from verified subjects using the inverse of probability of disease verification, that is, π_i^{-1}. First, we need to estimate TCRs, for each pair of two-ordered decision boundaries $c1 < c2$. Based on MAR assumption, we can show that

$$\mathrm{E}\left[I(T_i \le c1)\, V_i I(D_i = 0)\, \pi_i^{-1}\right] = \mathrm{E}\left\{\mathrm{E}\left[I(T_i \le c1)\, V_i I(D_i = 0)\, \pi_i^{-1} \,|\, T_i \boldsymbol{A_i}\right]\right\} =$$

$$= \mathrm{E}\left\{I(T_i \le c1)\mathrm{E}\left[V_i I(D_i = 0)\, \pi_i^{-1} \,|\, T_i \boldsymbol{A_i}\right]\right\}.$$

Since $\pi_i^{-1} = Pr(V_i = 1|T_i, \boldsymbol{A_i})^{-1}$, we have

$$\mathrm{E}\left[V_i \pi_i^{-1} \,|\, T_i \boldsymbol{A_i}\right] = Pr(V_i = 1|T_i, \boldsymbol{A_i})^{-1} Pr(V_i = 1|T_i, \boldsymbol{A_i}) = 1.$$

In addition, based on the MAR assumption, $V \perp\!\!\!\perp D \,|\, (T, A)$, so

$$\mathrm{E}\left[V_i I(D_i = 0)\, \pi_i^{-1} \,|\, T_i \boldsymbol{A_i}\right] = \mathrm{E}\left[I(D_i = 0) \,|\, T_i \boldsymbol{A_i}\right] \mathrm{E}\left[V_i \pi_i^{-1} \,|\, T_i \boldsymbol{A_i}\right].$$

Therefore, we can show that

$$\mathrm{E}\left[I(T_i \le c1)\, V_i I(D_i = 0)\, \pi_i^{-1}\right] = \mathrm{E}\left\{I(T_i \le c1)\mathrm{E}\left[V_i I(D_i = 0)\, \pi_i^{-1} \,|\, T_i \boldsymbol{A_i}\right]\right\} =$$

$$= \mathrm{E}\left\{I(T_i \le c1)\mathrm{E}\left[I(D_i = 0) \,|\, T_i \boldsymbol{A_i}\right] \mathrm{E}\left[V_i \pi_i^{-1} \,|\, T_i \boldsymbol{A_i}\right]\right\} = Pr(T_i \le c1 \,|\, D_i = 0) Pr(D_i = 0).$$

By similar argument, we have

$$\mathrm{E}\left[V_i I(D_i = 0)\pi_i^{-1}\right] = Pr(D_i = 0).$$

Therefore,

$$\begin{aligned}
D_i = 0 &: \widehat{TCR}^{0}_{IPW}(c1) = \frac{\sum_{i=1}^{n} I(T_i \le c1)\, V_i I(D_i = 0)\,/\,\pi_i}{\sum_{i=1}^{n} V_i I(D_i = 0)\,/\,\pi_i}\\
D_i = 1 &: \widehat{TCR}^{1}_{IPW}(c1,\ c2) = \frac{\sum_{i=1}^{n} I(c1 < T_i < c2)\, V_i I(D_i = 1)/\pi_i}{\sum_{i=1}^{n} V_i I(D_i = 1)\,/\,\pi_i}\\
D_i = 2 &: \widehat{TCR}^{2}_{IPW}(c2) = \frac{\sum_{i=1}^{n} I(T_i \ge c2)\, V_i I(D_i = 2)\,/\,\pi_i}{\sum_{i=1}^{n} V_i I(D_i = 2)\,/\,\pi_i}.
\end{aligned}$$

Then the bias-corrected ROC surface plot is made by plotting TCRs in the three-dimensional space for all possible pairs of thresholds. The estimation of $\widehat{VUS}_{IPW}$ assumed that $\pi_i = Pr(V_i = 1 \mid T_i, \boldsymbol{A_i})$, $i = 1, \ldots, n$ were known. At the design stage of a study when a protocol dictates which subjects get disease verified the sampling fractions (π) are known. In some studies, such as observational studies, however, the actual selection probabilities may be unknown and thus need to estimate the sampling fractions. Parametric models such as logistic regression models can be used to estimate the selection probabilities. In these cases, we could substitute $\widehat{\pi_i}$, the estimate of π_i, for it in the expressions for the estimator.

3.2 IPW estimator of VUS

Next, we develop an estimator of VUS that accounts for the verification biased sampling. The estimation of VUS is based on following observation:

$$\mathrm{E}\left[\pi_i^{-1}\ \pi_j^{-1}\ \pi_k^{-1} V_i\ \ V_j\ \ V_k I(T_i > T_j > T_k) I(D_i > D_j > D_k)\right] =$$

$$= \mathrm{E}\left\{\mathrm{E}\left[\pi_i^{-1}\ \pi_j^{-1}\ \pi_k^{-1} V_i\ \ V_j\ \ V_k I(T_i > T_j > T_k) I(D_i > D_j > D_k) | T_i,\ T_j,\ T_k,\ \boldsymbol{A}_i,\ \boldsymbol{A}_j,\ \boldsymbol{A}_k\right]\right\} =$$

$$= \mathrm{E}\left\{I(T_i > T_j > T_k)\mathrm{E}\left[\pi_i^{-1}\ \pi_j^{-1}\ \pi_k^{-1} V_i\ \ V_j\ \ V_k I(D_i > D_j > D_k) | T_i,\ T_j,\ T_k,\ \boldsymbol{A}_i, \boldsymbol{A}_j,\ \boldsymbol{A}_k\right]\right\} =$$

$$= \mathrm{E}\left\{I(T_i > T_j > T_k)\mathrm{E}\left[I(D_i > D_j > D_k) | T_i,\ T_j,\ T_k,\ \boldsymbol{A}_i,\ \boldsymbol{A}_j,\ \boldsymbol{A}_k\right]\right.$$

$$\left. E\left[\pi_i^{-1}\ \ V_i\ \mid\ T_i,\ \boldsymbol{A}_i\right] E\left[\pi_j^{-1}\ \ V_j\ \mid\ T_j,\ \boldsymbol{A}_j\right] E\left[\pi_k^{-1}\ \ V_k\ \mid\ T_k,\ \boldsymbol{A}_k\right]\right\} =$$

$$= \mathrm{E}\left\{I(T_i > T_j > T_k)\mathrm{E}\left[I(D_i > D_j > D_k) | T_i,\ T_j,\ T_k,\ \boldsymbol{A}_i,\ \boldsymbol{A}_j,\ \boldsymbol{A}_k\right]\right\} =$$

$$= \mathrm{E}\left\{I(T_i > T_j > T_k) I(D_i > D_j > D_k)\right\} =$$

$$= \mathrm{E}\left\{I(T_i > T_j > T_k) | I(D_i > D_j > D_k)\right\} P(D_i > D_j > D_k) =$$

$$= VUS \cdot Pr(D_i = 2) Pr(D_j = 1) Pr(D_k = 0) =$$

$$= VUS \cdot \lambda_0\, \lambda_1\, \lambda_2.$$

By similar argument, we can prove that

$$\mathrm{E}\left[\pi_i^{-1}\ \pi_j^{-1}\ \pi_k^{-1} V_i\ \ V_j\ \ V_k I(D_i > D_j > D_k)\right] = Pr(D_i = 2)Pr(D_j = 1)Pr(D_k = 0).$$

We propose direct estimation of the VUS in the presence of verification bias using the following:

$$\widehat{VUS}_{IPW} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\pi_i^{-1}\ \pi_j^{-1}\ \pi_k^{-1}\ \ V_i\ \ V_j\ \ V_k\ \ I(T_i > T_j > T_k)I(D_i > D_j > D_k)}{\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\pi_i^{-1}\ \pi_j^{-1}\ \pi_k^{-1}\ \ V_i\ \ V_j\ \ V_k\ \ I(D_i > D_j > D_k)}.$$

To express it as a function of U-statistics, we need to rewrite (1) in symmetric form as follows:

$$\widehat{VUS}_{IPW} = \frac{\sum_{i=1}^{n}\ \sum_{j=1}^{n}\ \sum_{k=1}^{n}\ \phi(S_i,\ \ S_j,\ \ S_k)}{\sum_{i=1}^{n}\ \sum_{j=1}^{n}\ \sum_{k=1}^{n}\ \psi(S_i,\ \ S_j,\ \ S_k)}\ ,$$

where

$$\begin{aligned}\phi(\boldsymbol{S}_i,\ \ \boldsymbol{S}_j,\ \ \boldsymbol{S}_k) = \frac{1}{6}\pi_i^{-1}\ \pi_j^{-1}\ \pi_k^{-1} V_i\ \ V_j\ \ V_k \cdot \Big\{ & I\left[(T_i - T_j)\ (D_i - D_j)\ >\ 0\right]\cdot \\ & \cdot I\left[(T_k - T_j)\ (D_k - D_j)\ >\ 0\right]\cdot I\left[(T_k - T_i)\ (D_k - D_i)\ >\ 0\right]\Big\} \\ \psi(\boldsymbol{S}_i,\ \ \boldsymbol{S}_j,\ \ \boldsymbol{S}_k) = \frac{1}{6}\pi_i^{-1}\ \pi_j^{-1}\ \pi_k^{-1} V_i\ \ V_j\ \ V_k \cdot \Big\{ & \left[I(D_i > D_j) + I(D_i < D_j)\right]\cdot \\ & \cdot\left[I(D_k > D_j) + I(D_j < D_k)\right]\cdot\left[I(D_i > D_k) + I(D_k < D_i)\right]\Big\}. \qquad (1)\end{aligned}$$

We can prove that the new estimator $\widehat{VUS}_{IPW}$ is consistent.

Theorem 1. *The IPW VUS estimator is consistent. The Appendix contains a sketch of the proof of Theorem 1.*

3.3 Variance estimation

Based on the asymptotic theory of U-statistics, we can obtain the asymptotic distribution of $\widehat{VUS}_{IPW}$.

Theorem 2. *Let* $\boldsymbol{A} = \left(\frac{1}{\lambda_0\lambda_1\lambda_2}, -\frac{VUS}{\lambda_0\lambda_1\lambda_2}\right)$*, then* $\sqrt{n}(\widehat{VUS}_{IPW} - VUS) \xrightarrow{p} N(0, \boldsymbol{A}^T\ \boldsymbol{\Sigma}\ \boldsymbol{A})$*, where* $\boldsymbol{\Sigma}$ *is*

$$\boldsymbol{\Sigma} = \begin{cases}\mathrm{Cov}\left(\phi(\boldsymbol{S}_i,\ \ \boldsymbol{S}_j,\ \ \boldsymbol{S}_k), \phi(\boldsymbol{S}_i,\ \ \boldsymbol{S}_m,\ \ \boldsymbol{S}_n)\right) & \mathrm{Cov}\left(\phi(\boldsymbol{S}_i,\ \ \boldsymbol{S}_j,\ \ \boldsymbol{S}_k), \psi(\boldsymbol{S}_i,\ \ \boldsymbol{S}_m,\ \ \boldsymbol{S}_n)\right) \\ \mathrm{Cov}\left(\psi(\boldsymbol{S}_i,\ \ \boldsymbol{S}_j,\ \ \boldsymbol{S}_k), \phi(\boldsymbol{S}_i,\ \ \boldsymbol{S}_m,\ \ \boldsymbol{S}_n)\right) & \mathrm{Cov}\left(\psi(\boldsymbol{S}_i,\ \ \boldsymbol{S}_j,\ \ \boldsymbol{S}_k), \psi(\boldsymbol{S}_i,\ \ \boldsymbol{S}_m,\ \ \boldsymbol{S}_n)\right).\end{cases}$$

A sketch of the proof of Theorem 2 is provided in the Supporting Information, Section A.

$\widehat{\lambda_d}$ can be estimated by $\frac{\sum_{j=1}^{n}\ \pi_i^{-1}\ \ V_i\ I(D_i=d)}{\sum_{j=1}^{n}\ \pi_i^{-1}\ \ V_i}$, and the empirical density functions (EDFs) $\widehat{F}_d(t),\ \widehat{G}_d(t, a)$ can also be estimated empirically using observed data for $d\ =\ 0,\ 1,\ 2$.

Next, we apply the Jackknife method to obtain variance estimation for our IPW estimator. First, we briefly describe Jackknife variance estimation. Detailed theory can be found in, for example, Tukey (1958) and Quenouille (1956).

Let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ be independent random variables and define

$$Q_n = Q(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n) = \binom{n}{m}^{-1} \sum_{1 \le i_1 < \cdots < i_m \le n} h(\boldsymbol{Z}_{i_1}, \ldots, \boldsymbol{Z}_{i_m})$$

as an one-sample U-statistic of degree m as a consistent estimator of the parameter θ. Let the Jackknife pseudovalues be $\hat{P}_i = nQ_n - (n-1)Q_{n-1}^{(-i)}$, where

$$\begin{aligned} Q_{n-1}^{(-i)} &= Q(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{i-1}, \boldsymbol{Z}_{i+1}, \ldots, \boldsymbol{Z}_n) = \\ &= \binom{n-1}{m}^{-1} \sum_{1 \le j_1 < \cdots < j_m \le n}^{-i} h(\boldsymbol{Z}_{j_1}, \ldots, \boldsymbol{Z}_{j_m}), \end{aligned}$$

which is the statistic Q_{n-1} computed using the $n-1$ sample from the original data set without the i-th data point. Then the Jackknife estimator of θ is the average of the pseudovalues:

$$\hat{Q}_n(jack) \cong \frac{1}{n} \sum_{i=1}^{n} \hat{P}_i$$

and the Jackknife estimate of the variance of Q_n is given by

$$\widehat{\text{Var}}(Q_n) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\hat{P}_i - \hat{Q}_n(jack))^2.$$

Recall that we proved that

$$VUS = \frac{E\left[\pi_i^{-1}\ \pi_j^{-1}\ \pi_k^{-1} V_i \;\; V_j \;\; V_k I(T_i > T_j > T_k) I(D_i > D_j > D_k)\right]}{E\left[\pi_i^{-1}\ \pi_j^{-1}\ \pi_k^{-1} V_i \;\; V_j \;\; V_k I(D_i > D_j > D_k)\right]}.$$

Let

$$\phi = E\left[\pi_i^{-1}\ \pi_j^{-1}\ \pi_k^{-1} V_i \;\; V_j \;\; V_k I(T_i > T_j > T_k) I(D_i > D_j > D_k)\right]$$

and

$$\psi = E\left[\pi_i^{-1}\ \pi_j^{-1}\ \pi_k^{-1} V_i \;\; V_j \;\; V_k I(D_i > D_j > D_k)\right]$$

be the numerator and denominator, respectively. Then estimators of ϕ and ψ are

$$\begin{aligned} \hat{\phi} &= \frac{1}{n(n-1)(n-2)} \sum^{i \ne j \ne k} \pi_i^{-1}\ \pi_j^{-1}\ \pi_k^{-1} V_i \;\; V_j \;\; V_k I(T_i > T_j > T_k)(D_i > D_j > D_k) \\ \hat{\psi} &= \frac{1}{n(n-1)(n-2)} \sum^{i \ne j \ne k} \pi_i^{-1}\ \pi_j^{-1}\ \pi_k^{-1} V_i \;\; V_j \;\; V_k I(D_i > D_j > D_k). \end{aligned}$$

The i-th Jackknife pseudovalue for $\hat{\phi}$ is $\hat{\phi}^i_{PS} = n\hat{\phi} - (n-1)\hat{\phi}^{(-i)}_{n-1}$ and the Jackknife estimator of ϕ, $\hat{\phi}_{JK} = \frac{1}{n}\sum_{i=1}^n \hat{\phi}^i_{PS}$. Thus the Jackknife estimator of variance of ϕ is

$$\widehat{\mathrm{Var}}(\phi) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\hat{\phi}^i_{PS} - \hat{\phi}_{JK})^2.$$

Similarly, the Jackknife estimator of the variance of ψ is

$$\widehat{\mathrm{Var}}(\psi) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\hat{\psi}^i_{PS} - \hat{\psi}_{JK})^2$$

and the Jackknife estimator of the covariance between ϕ and ψ is

$$\widehat{\mathrm{Cov}}(\phi,\psi) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left[(\hat{\phi}^i_{PS} - \hat{\phi}_{JK})(\hat{\psi}^i_{PS} - \hat{\psi}_{JK})\right].$$

By the multivariate delta method, we have

$$\widehat{\mathrm{Var}}(VUS_{IPW}) = \frac{\hat{\phi}^2}{\hat{\psi}^4}\,\widehat{\mathrm{Var}}(\psi) + \frac{1}{\hat{\psi}^2}\,\widehat{\mathrm{Var}}(\phi) + \frac{2\hat{\phi}}{\hat{\psi}^3}\,\widehat{\mathrm{Cov}}(\phi,\psi).$$

4 Simulation studies

Let us define short notations for our proposed estimators. We consider two IPW estimators. "IPW(E)" corresponds to $\widehat{VUS}_{IPW}$ using estimated π by assuming a logistic regression model for the verification process, while "IPW(K)" corresponds to $\widehat{VUS}_{IPW}$ using known π based on prior knowledge or study design. "CC" represents the naïve estimator of VUS only using complete cases without bias correction, $\widehat{VUS}_{\text{naïve}}$. "Full data" represents the estimates based on complete data when every subject gets disease verified, which should be unbiased. "True" estimates are the average "full data" estimates when sample size is 5000 across 1000 realizations, which can represent the true value of VUS.

The simulation setup of Alonzo et al. (2003, 2005) is modified to accommodate three-ordered disease stages rather than binary disease status. Specifically, the disease status is formed based on an underlying continuous pathology process, which remains subclinical or mild disease until it reaches certain thresholds and hence progresses to the next disease stage. In this simulation study, the disease status D with three stages, that is, nondisease, mild disease, and severe disease, is generated based on the comparison between random variable $Z \sim N(0,1)$ and two-ordered decision boundaries $p_1 < p_2$. That is, nondisease ($D = 0$) if $Z < p_1$; severe disease ($D = 2$) if $Z > p_2$, and mild disease ($D = 1$), otherwise. In other words, the thresholds p_1 and p_2 determine the prevalence of the disease. In practice, there are usually multiple factors contributing to the development of disease, so it is reasonable for us to view Z as the sum of two independent random variables Z_1 and Z_2 where both of them $\sim N(0, 0.5)$. The aim of the continuous new diagnostic test results is to capture the information in those factors, and here, we construct T as a linear combination of Z_1, Z_2 and random normal error, ϵ_1. In particular,

$$T = \alpha_1 Z_1 + \beta_1 Z_2 + \epsilon_1, \qquad \epsilon_1 \sim N(0, 0.25).$$

Test accuracy varies according to the values of α_1 and β_1 (see Supporting Information Table SB1). The new diagnostic test can be highly accurate when $\alpha_1 = \beta_1 = 1$. Decreasing the value of either α_1 or

Table 1 Mean estimated VUS from 1000 realizations of the simulation with different sample sizes.

Method	Sample size					
	100	400	600	800	1000	2000
Full data	0.798 (0.7%)	0.791 (0.2%)	0.791 (−0.2%)	0.794 (−0.2%)	0.792 (−0.1%)	0.793 (0.1%)
CC	0.752 (−5.1%)	0.745 (−6.0%)	0.745 (−5.9%)	0.748 (−5.6%)	0.746 (−5.8%)	0.748 (−5.7%)
IPW(E)	0.800 (1.0%)	0.791 (0.2%)	0.791 (0.1%)	0.795 (−0.4%)	0.793 (0.0%)	0.793 (0.1%)
IPW(K)	0.797 (0.6%)	0.791 (0.2%)	0.791 (0.2%)	0.795 (−0.4%)	0.792 (0.0%)	0.793 (0.1%)

Relative bias, (mean VUS − true estimate of 0.792)/0.792 is provided in parentheses.

β_1 to zero, causes the test to capture less information about the disease status and thus leads to lower accuracy. Similarly, when both the values of α_1 and β_1 decrease to 0.5, the inherent accuracy of T also decreases. By letting $\alpha_1 = \beta_1 = 0$, the test becomes no better than random guessing.

One covariate A is generated, in a way similar to T, according to the following relation:

$$A = \alpha_2 Z_1 + \beta_2 Z_2 + \epsilon_2, \qquad \epsilon_2 \sim N(0, 0.25),$$

where ϵ_1 and ϵ_2 are independent. Again, by varying the value of α_2 and β_2, one can change the degree to which the covariate is correlated with disease status.

The verification mechanism is designed as a Bernoulli random variable with the verification probability related to the test results and covariate using the probability model

$$\begin{aligned}\text{logit}(P(V = 1)) = \gamma_0 + \gamma_1 I(t^{q_1} < T \leq t^{q_2}) + \\ + \gamma_2 I(T > t^{q_2}) + \gamma_3 I(a^{q_1} < A \leq a^{q_2}) + \gamma_4 I(A > a^{q_2}),\end{aligned} \qquad (2)$$

where t^{q_1} and t^{q_2} are the q_1-th and q_2-th quantiles of the distribution of T or A ($q_1 < q_2$), respectively. In this setting, verification status is related to the disease status through the new test results and covariate. Higher T or A increases the probability of disease verification. Using ordered thresholds t^{q_1} and t^{q_2}, patients are classified into different categories according to the values of their T and A. Different verification probabilities are given to patients in different categories to introduce informative missingness in the disease status. In the estimation procedure, the correct models are assumed to hold for the verification probability.

The diagnostic test under study is highly accurate (true VUS = 0.792) with $\alpha_1 = \beta_1 = 1$. The value of $(p_1,\ p_2)$ is chosen to yield disease prevalence rates of 5%, 15%, and 80%, for nondisease, mild-disease, and severe-disease status, respectively. $(q_1,\ q_2)$ and $(\gamma_0,\ \gamma_1,\ \gamma_2, \gamma_3,\ \gamma_4)$ are chosen to be (70%, 90%) and (−1.4, 1, 2, 0.8, 2), respectively. Note that under this verification model, subjects with higher test results or higher covariate values or both are more likely to receive disease verification. Averaging over 1000 realizations of simulation with a sample size of 2000, the percentage of disease verification among subjects in nondisease ($D = 0$), mild-disease ($D = 1$), and severe-disease ($D = 2$) groups are 26%, 67%, and 89%, respectively. Overall, the probability of verification in the population is about 35%. The default simulations have a sample size of 1000. Simulation results under different scenarios are summarized over 1000 replications.

4.1 Performance of VUS: bias

Table 1 reports average estimated VUS of the diagnostic test across 1000 replications of the simulation for sample sizes ranging from 100 to 2000. As expected, the CC estimator is biased (5.1–6.0%), compared to the true value of VUS ("true" estimates). Both IPW(E) and IPW(K) correct the noticeable bias of the CC estimator and are quite close to the full data estimates. The relative bias of IPW and full data to "true" estimates are negligible. As sample size increases, their performance become better.

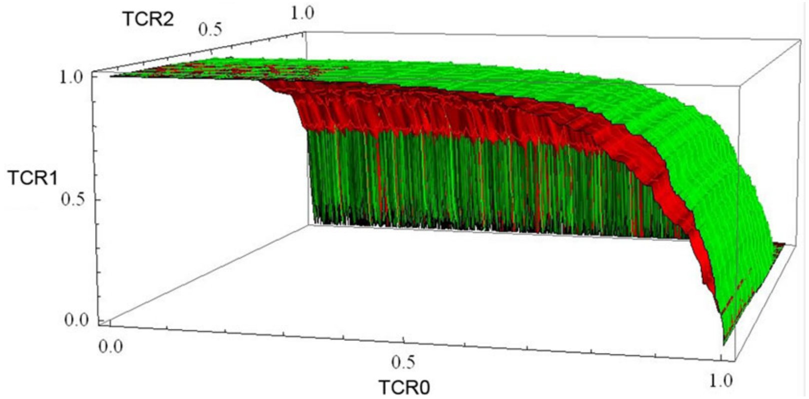

Figure 1 Full data (green) and CC (red) ROC surface from a randomly chosen realization of the simulation study; TCR0, TCR1, and TCR2 for $D = 0$, 1, and 2, respectively.

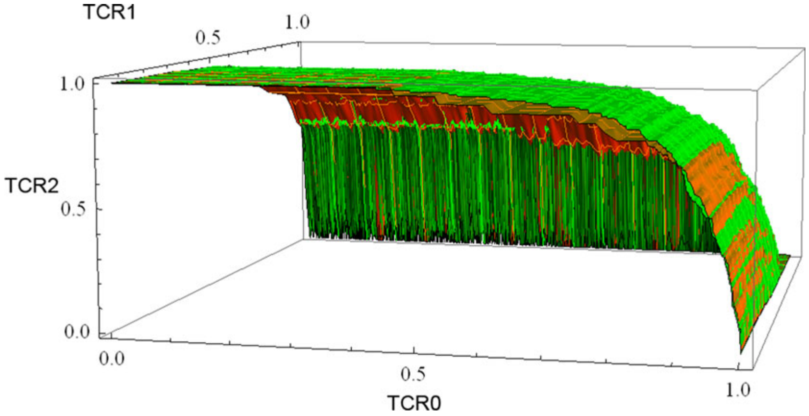

Figure 2 Full data (green) and IPW (K) (red) ROC surface from a randomly chosen realization of the simulation study; TCR0, TCR1, and TCR2 for $D = 0$, 1, and 2, respectively.

The verification bias caused by ignoring the biased sampling procedure for disease verification is also apparent in Fig. 1, where the empirical ROC surface based on the CC (red surface) and full data (green surface) for one randomly chosen data realization with sample size of 1000 is presented. The CC ROC surface is biased downwards relative to the full data curve (VUS value: 0.772 vs. 0.807). The graphs are drawn with respect to the three correct classification rates, across all possible decision thresholds. Figure 2 provides the bias-corrected ROC surface using IPW(E) (red surface) and the full data (green surface) ROC surface. It is not surprising that IPW ROC surface lies closely next to the full data ROC surface and the corresponding VUS is similar to the full data value (0.812 vs. 0.807). The ROC surface for IPW(K) is not presented because it is very similar to that for IPW(E).

When sample size is as small as 100 and the overall disease verification rate is 34%, approximately only 34 subjects have disease status verified. Although the CC and IPW estimators both only require data from verified subjects, the IPW estimators exhibited substantially smaller bias than the naïve CC estimator. This further justifies the useful application of the IPW estimators in correcting verification bias and properly assessing the test accuracy with small sample size.

Table 2 The ratio of the Monte Carlo mean of the estimated SD to simulation SD of the estimators.

Method	Sample size				
	400	600	800	1000	2000
Full data	0.972 (88.6)	0.991 (90.4)	0.987 (90.2)	1.028 (91.3)	1.034 (91.7)
CC	0.998 (89.7)	1.008 (89.9)	1.013 (90.0)	1.017 (91.6)	1.033 (91.5)
IPW(E)-J	0.986 (90.2)	1.008 (90.3)	1.007 (91.0)	1.015 (91.7)	1.043 (91.8)
IPW(K)-J	0.983 (89.7)	1.001 (90.2)	1.002 (90.5)	1.012 (91.5)	1.034 (91.6)
IPW(E)-CF	0.714 (76.9)	0.833 (83.9)	0.875 (87.2)	0.911 (87.8)	0.992 (89.4)
IPW(K)-CF	0.917 (87.0)	0.949 (88.5)	0.966 (88.6)	0.980 (90.2)	1.019 (91.0)

90% CI coverage probabilities of VUS variance estimator are provided in parentheses. Jackknife variance estimator used for full data, CC, IPW(E)-J, and IPW(K)-J. Closed-form variance estimator used for IPW(E)-CF and IPW(K)-CF.

4.2 Performance of VUS: SD

In Section 3.3, we developed a closed-form asymptotic variance estimator as well as Jackknife variance estimator for the IPW estimators of VUS. Here we further investigate the performance of these asymptotic standard deviation (SD) estimators for the VUS for a range of sample sizes. The ratio between the Monte Carlo mean of the estimated SD to the Monte Carlo SD of the VUS estimators (simulation SD) is presented. The simulation SD of the VUS estimators is calculated from 1000 realizations and can adequately represent the variability of the estimators.

The Jackknife variance estimator performs well for sample size as small as 100, with the ratio of estimated SD to simulation SD ranging from 0.830 to 1.043. When sample size is small (<400), the closed-form IPW(E) VUS variance estimator substantially underestimates the variance of IPW(E) (SD ratio <30%). On the contrary, the closed-form IPW(K) VUS variance estimator is close to empirical estimates. The SD ratio is around 91% when sample size is 100. The poor performance of IPW(E) VUS variance estimator is likely due to the fact that the estimated variance does not account for the variation in estimating the verification probability. Naturally the extension to consider the variation in estimating the verification probability is worth further investigation, which is discussed in Section 6. As sample size increases, the improvement in the performance of the closed-form variance estimator for IPW(E) demonstrates that this should not be a significant problem with sample sizes greater than 600. This is not surprising because the variance in the estimated verification probability is relatively small, compared with the variance of VUS.

The performance of the variance estimators is further assessed by calculating coverage probabilities of the confidence interval (CI) corresponding to the variance estimator. Table 2 summarizes nominal 90% CI coverage probabilities of VUS for several sample sizes. For the Jackknife variance estimators, nominal 90% coverage probabilities between 89.7% and 91.8% are achieved. The nominal 90% coverage probabilities for the closed-form variance estimators for IPW(K) have comparable performance (87.0–91.0%). Not surprising the closed-form variance estimator for IPW(E) did not obtain 90% coverage rate (71.4–83.3%) when sample size is small ($\leq$ 600) but has good coverage for sample size of 2000. In Table 3, the IPW mean squared error (MSE) is smaller than the CC MSE (between 45% and 78%) but is about 35% more than the full data MSE, which is reasonable as about 66% of the observations are missing the disease status.

In general, the Jackknife variance estimator and the closed-form variance estimator of $\widehat{VUS}_{IPW}$ using known π perform very well under small samples. The closed-form variance estimator of $\widehat{VUS}_{IPW}$ using estimated π, however, requires relatively larger sample size to achieve good performance.

Table 3 MSE $\times 10^{-3}$ of VUS estimators.

Method	Sample size				
	400	600	800	1000	2000
Full data	2.21	1.46	1.06	0.86	0.42
CC	5.42	4.30	3.46	3.34	2.59
IPW(E)	3.03	1.97	1.43	1.15	0.57
IPW(K)	2.98	1.97	1.41	1.14	0.57

Jackknife variance estimator used for full data, CC, IPW(E), and IPW(K).

Table 4 Detailed information about verification rate: varying test accuracy.

	Scenario A	Scenario B	Scenario C	Scenario D
Verification rate (overall)	35%	36%	36%	36%
Verification rate ($D = 0$)	26%	28%	28%	31%
Verification rate ($D = 1$)	67%	59%	62%	51%
Verification rate ($D = 2$)	89%	80%	84%	67%

4.3 Varying test accuracy and verification rates

The inherent accuracy of the new diagnostic test may influence the degree of verification bias (Begg and Greenes, 1983). In this setting, by changing values of α_1 and β_1, we can vary the accuracy of the test. We consider the following four scenarios,

Scenario A (Default setting): $\alpha_1 = \beta_1 = 1$.
Scenario B: $\alpha_1 = 1$, $\beta_1 = 0$.
Scenario C: $\alpha_1 = 0.5$, $\beta_1 = 0.5$.
Scenario D: $\alpha_1 = 0$, $\beta_1 = 0$.

Detailed information about verification percentages for each scenario can be found in Table 4.

Simulation results are obtained by averaging over 1000 realizations of simulation with a sample size of 1000. All other simulation setup parameters, including verification mechanism and disease prevalence, stay the same as previous simulation setup. Results are presented in Table 5. Since the CC estimator does not account for biased sampling, as expected, CC yields biased estimates of VUS when there is a potential for verification bias. In each scenario, CC underestimates the true VUS. Conversely, IPW is very close to the true VUS. Note that as the test accuracy decreases, the relative bias tends to increase which is a reflection of the smaller value of base value (true value of VUS). Jackknife variance estimators are similar to the simulation variance and the 90% coverage rates are near their nominal values.

Table 5 also provides MSE of the VUS estimators for scenarios A–D. Under all scenarios, IPW shows smaller MSE relative to CC. The two IPW estimators give comparable performance.

We also investigate the performance of the VUS estimators when we vary the verification percentage. By changing the gamma values in (4), we generate four different scenarios representing low, medium, high verification rate as well as random verification are considered (see Supporting Information Tables SB2 and SB3). As expected, we find little small sample bias (less than 0.3%) in the IPW estimator. Conversely, CC underestimates VUS (−1.2% to −5.4%). When verified subjects are selected completely

Table 5 Comparison of the VUS estimators: varying test accuracy.

	Method	VUS[a)]	SD[b)]	SE[c)]	Coverage (%)[d)]	MSE[e)]
Scenario A	Full data	0.792 (−0.1%)	0.028	0.029	91.3	0.86
	CC	0.746 (−5.8%)	0.034	0.035	91.6	3.34
	IPW(E)	0.793 (0.1%)	0.033	0.034	91.7	1.15
	IPW(K)	0.792 (0.0%)	0.033	0.034	91.5	1.14
Scenario B	Full data	0.458 (0.2%)	0.036	0.036	90.1	1.27
	CC	0.402 (−12.1%)	0.041	0.040	88.7	4.69
	IPW(E)	0.459 (0.5%)	0.046	0.045	88.6	1.99
	IPW(K)	0.459 (0.4%)	0.046	0.044	88.4	1.98
Scenario C	Full data	0.565 (−0.0%)	0.036	0.036	90.6	1.32
	CC	0.506 (−10.5%)	0.041	0.042	90.0	5.23
	IPW(E)	0.567 (0.2%)	0.045	0.045	89.9	2.00
	IPW(K)	0.566 (0.2%)	0.045	0.045	89.7	1.99
Scenario D	Full data	0.167 (−0.0%)	0.023	0.023	90.0	0.52
	CC	0.136 (−18.1%)	0.025	0.025	90.0	1.55
	IPW(E)	0.168 (0.9%)	0.030	0.029	89.4	0.86
	IPW(K)	0.168 (0.9%)	0.030	0.029	89.7	0.86

a) Relative bias to "true" estimates is provided in parentheses.
b) Simulation SD.
c) The average of the SD estimator (Jackknife).
d) 90% CI coverage probabilities calculated using the SD estimator (Jackknife).
e) MSE $\times 10^{-3}$ of VUS estimators calculated using the SD estimator (Jackknife).

at random, there is little potential for verification bias. And since each verified observation is attached with the same weight (π_i^{-1}), the IPW estimator and CC estimator are both unbiased and approximately the same. This simulation supports that IPW remains valid and robust under different verification mechanisms.

5 Data application

We illustrate our proposed approaches with data from the ADNI study. ADNI is a large, multicenter, longitudinal neuroimaging study launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies, and nonprofit organizations (Misra et al., 2009). For disease verification, due to the fact that it is unrealistic to perform autopsy on all the subjects under study, ADNI investigators used neuropsychological tests to identify disease stages into probable AD, amnestic MCI, and normal cognition. Over 1600 participants with MCI, probable AD, and elderly, cognitively normal control (CN) have been identified and recruited over three phases: ADNI 1, ADNI GO, and ADNI 2.

One of the goals of the ADNI study is to develop a cerebrospinal fluid (CSF) biomarker signature for AD stage identification (Misra et al., 2009). Among all the potential biomarkers under study, CSF tau protein concentrations are considered to be a promising biomarker that may be diagnostic for different AD stages relative to degree of cognitive impairment (Frank et al., 2003). Also, recent studies showed that CSF tau changes may predict the conversion to AD in MCI subjects (Hansson et al., 2006). However, before introducing it into clinical diagnosis, it is critically important to perform a

Table 6 Summary statistics of ADNI data.

Variable	Summary statistics
D	Nondisease: 25%; mild disease: 18%; severe disease: 56%
T[a)]	Mean: 0.002; Median: -0.261; SD: 0.999
A[b)]	Mean: 0.002; Median: 0.246; SD: 1.000

a) Tau protein: values are standardized.
b) Aβ1-42: values are standardized and multiplied by -1.

study to determine the accuracy of using tau level in CSF as a biomarker to predict different stages of AD.

In our analysis, we are interested in evaluating the diagnostic accuracy of CSF tau protein in predicting stages of AD in terms of cognitive impairment, which ranges from normal aging (nondisease), to MCI (mild disease), and probable AD (severe disease). Our purpose is to evaluate whether the proposed estimators can properly assess the accuracy of CSF tau protein without requiring disease verification on all the subjects. In practice, the clinical decision about disease verification can be made based on a combination of new test results and available clinical characteristics. To mimic the real-life application without overcomplication, we introduce nonrandom missingness in disease status based on levels of tau protein (new biomarker test) in CSF and a clinical covariate. We then apply the proposed methods to estimate the ROC surface and VUS. We have the full data results so we can compare our estimators to the full data estimator, which is not subject to verification bias.

Here we consider tau protein levels as the new diagnostic test. Test results are positively related with disease status (i.e., larger values of test results indicate more severe disease stages). Reduced CSF levels of Aβ1-42 are believed to result from large-scale accumulation of this Aβ peptide into insoluble plaques in the AD brain, which makes Aβ1-42 concentration in CSF an informative clinical covariate (Frank et al., 2003). We denote D to be the true AD stages ($D = 0$: normal aging; $D = 1$: MCI; $D = 2$: probable AD). T denotes tau protein levels, A denotes Aβ1-42, and V denotes verification status. We continue to use notation defined before.

Excluding subjects without data on tau protein (T) and Aβ1-42 (A), the full data set in our analysis includes AD stages (D), tau protein, and Aβ1-42 for 1078 subjects. As mentioned before, a decreased level of A implies higher degree of disease stages. For convenience, T and $-A$ are standardized. Note that there are tied values in both T and A. Since our methods apply to continuous test results, in practice, one reasonable approach is to add a small amount of uniformly distributed random “jitter” to each test result to break ties. We study the sensitivity of our proposed estimators to different amount of jitter added to the test results. We estimate VUS over a series of values of jitter by adding a uniform number between [−jitter, +jitter] to the test result. We tried four values: 0.001, 0.0001, 0.00001, and 0.000001 and they all give almost identical results in terms of bias and variance (Euclidean distance between average estimate of VUS is less than -10^{-10}). Therefore, we only present results with jitter randomly selected from the uniform distribution $[-10^{-5}, 10^{-5}]$.

Table 6 shows the summary statistics for all the subjects used in our analysis. We observe that T and A are not normally distributed with large variation relative to the mean. It is appropriate to apply our proposed estimators since our approach is nonparametric. Also due to the large variation, we expect wider 95% CI than that of simulation results with similar setting. Subjects with higher T and A are more likely to have higher degree of disease status, which can be shown in boxplots of T and A for stratified disease status (see Supporting Information Fig. SC1).

Only a fraction of the subjects receive the GST and have disease status observed. T and A are obtained on all subjects. The verification status, V is generated with a model as following:

Table 7 Estimates of the VUS using data from ADNI.

Method	VUS	JK SD (×100)	Asymptotic SD (×100)	95 % CI (JK)
Full data	0.356	1.872	—	(0.320, 0.393)
CC	0.296	2.986	—	(0.237, 0.354)
IPW(E)	0.357	5.358	5.109	(0.252, 0.462)
IPW(K)	0.357	5.285	4.950	(0.253, 0.460)

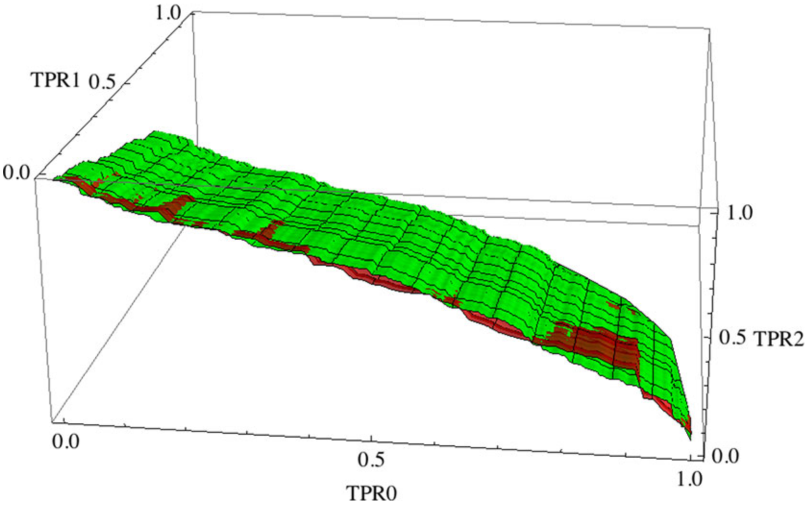

Figure 3 Full data (green) and CC (red) ROC surface of tau protein; TCR0, TCR1, and TCR2 for $D = 0$, 1, and 2, respectively.

$$\log\left\{\frac{P(V=0 \mid D\ ,\ T\ ,\ A)}{P(V=1 \mid D\ ,\ T\ ,\ A)}\right\} = \gamma_0 + \gamma_1 T + \gamma_2 A.$$

In particular, we choose $(\gamma_0, \gamma_1, \gamma_2) = (-0.6, 1.6, 1)$ for the verification model. Overall, 39% of the subjects receive disease verification. The percentage of disease verification among subjects in nondisease ($D = 0$), mild-disease ($D = 1$), and severe-disease ($D = 2$) groups are 20%, 37%, and 69%, respectively. Since the conditional probability of having disease status verified is in the denominator of the model, subjects with higher T and A are related to increased verification rates (see Supporting Information Fig. SC2).

The resulting estimates of VUS are presented in Table 7. The asymptotic variance result from Theorem 2 as well as the Jackknife variance estimate are presented. As expected, CC underestimates the VUS in this case by nearly 17% and its 95% CI does not even include the true value. The bias of CC estimator is also apparent in Fig. 3, where the empirical ROC surface based on the complete cases (red surface) and full data (green surface) is presented. The CC ROC surface is consistently biased downwards relative to the full data curve. Figure 4 provides the bias-corrected ROC surface using IPW(K) (red surface) and the full data (green surface) ROC surface. The bias-correction method shows underestimation and overestimation evenly across different pairs of decision thresholds, which yields VUS estimates similar to the full data value, with values of bias close to 0.3%. The ROC surfaces for IPW(E) is not presented because they are very similar to that for IPW(K). Clearly, the naïve estimator using verified subjects without taking into account the verification bias could lead to inaccurate conclusions about the accuracy of the test.

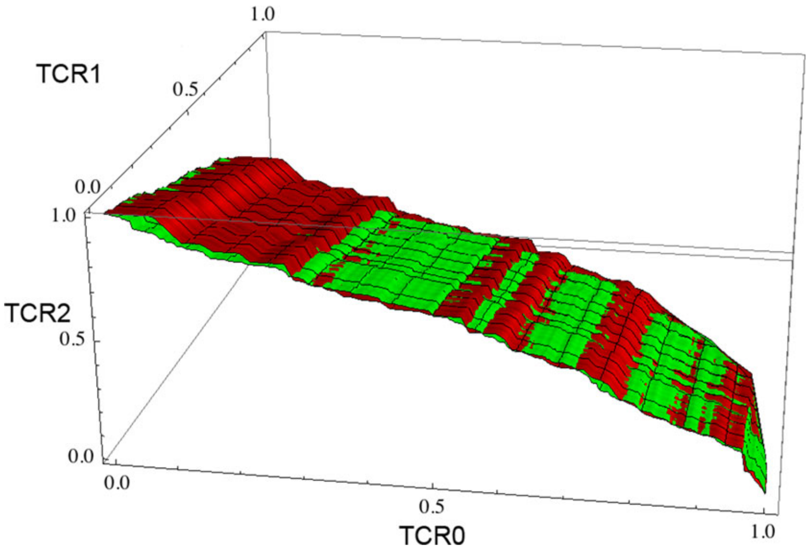

Figure 4 Full data (green) and IPW(K) (red) ROC surface of tau protein; TCR0, TCR1, and TCR2 for $D = 0$, 1, and 2, respectively.

Comparing IPW(K) and IPW(E), the 95 % CIs obtained based on Jackknife methods are very close to each other and both of them contain the full data value of 0.357. The variance estimates obtained via the Jackknife methods and the asymptotic theorem are very similar, which further supports the application of the asymptotic variance in real-life data.

6 Discussion

This paper focuses on the important problem of assessing the accuracy of a continuous medical diagnostic test with biased sampling in disease verification for three-class classification problems. We explored the estimation of the ROC surface and the summary measure of the ROC surface, VUS, in the presence of verification bias. Through theoretical framework and simulation studies, we showed that proposed IPW estimator performs quite well. In addition, Jackknife variance estimators were derived and shown to work well in finite sample setting. A closed-form expression for the asymptotic variance of the IPW estimator was also derived and it resembled the simulation variance when the verification process is known. With relative large sample size (> 600) and reasonably estimated verification probabilities, the closed-form variance estimator with estimated verification rate performs well. In practice, when we know the verification process well, we recommend to use the closed-form variance estimator for IPW VUS with known verification probabilities. When we do not have a good understanding about the selection process for disease verification and sample size is small (< 400), we recommend the Jackknife approach to estimate variance of our proposed estimators because it gives good performance in terms of closeness to the simulated variance and maintaining required coverage probabilities. Otherwise, when sample size is large, the closed-form variance estimator for IPW with estimated verification probabilities is preferred as it is comparable to the simulated variance and is much less computer intensive.

The proposed estimators are easy to implement and only require a subset of subjects under study to obtain disease verification. When the GST is too expensive, our approach can reduce the cost by only requiring partial disease verification but at the same time still presents valid conclusions about the accuracy of the new diagnostic test. If the studies are designed so that the verification probabilities are known or can be estimated reasonably well, the performance of our estimators can be further ensured. The methodology proposed may be easily modified to diagnostic problems with more than three classes of disease.

There are some issues that remain open for future investigation. In simulation studies, we found that the variance estimator has good performance when the verification probabilities, that is, π_i, are known. But when π_i are not known, the variance estimator tends to underestimate the variance, which probably is due to the fact that we did not account for the variance in estimating the π_i. Although this should not be a significant problem in large samples, developing variance estimators that account for the variance in $\hat{\pi}_i$ warrants further attention. In addition, here we developed bias-corrected estimators based on the MAR assumption. In practice, however, such assumption may not hold since the doctor's decision to refer a patient to receive disease verification may depend on comprehensive information on the patient's health condition, which may be not be guaranteed to be fully captured by the test results or other measured auxiliary data. In such a situation, we call the underlying missingness in disease status nonignorable. Rotnitzky et al. (2006) developed a doubly robust bias-correction procedure for the AUC under nonignorable missingness. 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.

Acknowledgments The authors thank Dr. Helena Chang Chui for providing her insightful comments and knowledge about Alzheimer's disease. The author also thank the Alzheimer's disease Neuroimaging Initiative research group for kindly permitting access to the data analyzed in this paper. In addition, the authors thank an editor and two anonymous reviewers for helpful comments that led to an improved version of the article. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC; Johnson & Johnson Pharmaceutical Research & Development LLC; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Conflict of interest

The authors have declared no conflict of interest.

Appendix

Proof of Theorem 1

We already know that,

$$E\left[\phi(\boldsymbol{S}_i,\ \boldsymbol{S}_j,\ \boldsymbol{S}_k)\right] = VUS\ \lambda_0\,\lambda_1\,\lambda_2$$

$$E\left[\psi(\boldsymbol{S}_i,\ \boldsymbol{S}_j,\ \boldsymbol{S}_k)\right] = \lambda_0\,\lambda_1\,\lambda_2.$$

Therefore, by the law of large numbers, we have

$$\frac{\sum_{i\neq j\neq k}\ \phi(\boldsymbol{S}_i,\ \boldsymbol{S}_j,\ \boldsymbol{S}_k)}{n(n-1)(n-2)} \xrightarrow{p} VUS\ \lambda_0\,\lambda_1\,\lambda_2$$

$$\frac{\sum_{i\neq j\neq k}\ \psi(\boldsymbol{S}_i,\ \boldsymbol{S}_j,\ \boldsymbol{S}_k)}{n(n-1)(n-2)} \xrightarrow{p} \lambda_0\,\lambda_1\,\lambda_2.$$

Hence,

$$\widehat{VUS}_{IPW} = \frac{\sum_{i\neq j\neq k}\ \phi(\boldsymbol{S}_i,\ \boldsymbol{S}_j,\ \boldsymbol{S}_k)}{\sum_{i\neq j\neq k}\ \psi(\boldsymbol{S}_i,\ \boldsymbol{S}_j,\ \boldsymbol{S}_k)} \xrightarrow{p} VUS.$$

References

Alonzo, T. A. and Nakas, C. T. (2007). Comparison of ROC umbrella volumes with an application to the assessment of lung cancer diagnostic markers. *Biometrical Journal* **49**, 654–664.

Alonzo, T. A., Nakas, C. T., Yiannoutsos, C. T. and Bucher, S. (2009). A comparison of tests for restricted orderings in the three-class case. *Statistics in Medicine* **28**, 1144–1158.

Alonzo, T. A. and Pepe, M. S. (2005). Assessing accuracy of a continuous screening test in the presence of verification bias. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **54**, 173–190.

Alonzo, T. A., Pepe, M. S. and Lumley, T. (2003). Estimating disease prevalence in two-phase studies. *Biostatistics* **4**, 313–326.

Begg, C. and Greenes, R. (1983). Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* **39**, 207–215.

Chi, Y. and Zhou, X. (2008). Receiver operating characteristic surfaces in the presence of verification bias. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **57**, 1–23.

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.

Frank, R. A., Galasko, D., Hampel, H., Hardy, J., De Leon, M. J., Mehta, P. D., Rogers, J., Siemers, E. and Trojanowski, J. Q. (2003). Biological markers for therapeutic trials in Alzheimer's disease: proceedings of the biological markers working group; NIA initiative on neuroimaging in Alzheimer's disease. *Neurobiology of Aging* **24**, 521–536.

Hansson, O., Zetterberg, H., Buchhave, P., Londos, E., Blennow, K. and Minthon, L. (2006). Association between CSF biomarkers and incipient Alzheimer's disease in patients with mild cognitive impairment: a follow-up study. *Lancet Neurology* **5**, 228–234.

He, H., Lyness, J. and McDermott, M. (2009). Direct estimation of the area under the receiver operating characteristic curve in the presence of verification bias. *Statistics in Medicine* **28**, 361–376.

Heckerling, P. S. (2001). Parametric three-way receiver operating characteristic surface analysis using mathematica. *Medical Decision Making* **21**, 409–417.

Li, J. and Zhou, X. (2009). Nonparametric and semiparametric estimation of the three way receiver operating characteristic surface. *Journal of Statistical Planning and Inference* **139**, 4133–4142.

Little, R. and Rubin, D. (1987). *Statistical Analysis with Missing Data*. Wiley, New York, NY.

McNeil, B. and Adelstein, S. (1976). Determining the value of diagnostic and screening tests. *Journal of Nuclear Medicine* **17**, 439–486.

Misra, C., Fan, Y. and Davatzikos, C. (2009). Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI. *NeuroImage* **44**, 1415–1422.

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

Nakas, C. T. and Alonzo, T. A. (2007). ROC graphs for assessing the ability of a diagnostic marker to detect three disease classes with an umbrella ordering. *Biometrics* **63**, 603–609.

Nakas, C. T. and Alonzo, T. A. (2010). Accuracy and cut-off point selection in three-class classification problems using a generalization of the Youden index. *Statistics in Medicine* **29**, 2946–2955.

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

Pepe, M. S. and Alonzo, T. A. (2001). Comparing disease screening tests when true disease status is ascertained only for screen positives. *Biostatistics* **2**, 249–260.

Quenouille, M. H., (1956). Notes on bias in estimation. *Biometrika* **43**, 353–360.

Ransohoff, D. and Feinstein, A. (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New England Journal of Medicine* **299**, 926–930.

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.

Scurfield, B. (1978). Multiple-event forced-choice tasks in the theory of signal detectability. *Journal of Mathematical Psychology* **40**, 253–269.

Sox, H. C., Blatt, M. A., Higgins, M. C. and Marton, K. I. (1988). *Medical Decision Making*. Butterworths-Heinemann, Boston, MA.

Tukey, J. (1958). Bias and confidence in not-quite large samples. *Annals of Mathematical Statistics* **29**, 614.