Linear Combinations of Biomarkers to Improve Overall Diagnostic Accuracy with Three Ordinal Diagnostic Categories

Le Kang* Lili Tian† Chengjie Xiong‡ Paul Crane§

Abstract

Many researchers have addressed the problem of finding the optimal linear combination of biomarkers to maximize the area under ROC curves (AUC) for scenarios with binary disease status. In practice, many disease processes such as Alzheimer can be naturally classified into three diagnostic categories such as normal, mild cognitive impairment and Alzheimer's disease, and for such diseases the volume under the ROC surface (VUS) is the most commonly used index of diagnostic accuracy. In this article, we propose a few parametric and nonparametric approaches to address the problem of finding the optimal linear combination to maximize the VUS. Simulation studies were carried out to investigate the performance of the proposed methods. All of the investigated approaches are applied to a real data set from a cohort study in early stage Alzheimer's disease (AD).

Key Words: diagnostic accuracy; linear combinations; ordinal categories; volume under the ROC surface

1. Introduction

Multiple diagnostic tests are often performed on the same individual to provide clinicians as much information as possible in order to make more accurate disease diagnosis as it is becoming increasingly clear that one single diagnostic test or biomarker is not sufficient to serve as an optimal screening device for early detection or prognosis Sidransky (2002). It is therefore of critical importance to combine the information available in an optimal way to improve overall diagnostic accuracy Etzioni et al. (2003).

When the diagnostic outcome is binary, i.e., non-diseased and diseased, the receiver operating characteristic (ROC) curves and the area under the ROC curves (AUC) are commonly used diagnostic accuracy measures. Many conditions are conceptualized as having a normal stage, an early/mild/prodromal stage, and a late/diagnosable/fully symptomatic stage. For example, mild cognitive impairment (MCI) and/or early stage Alzheimer's disease (AD) is a transitional stage between the cognitive changes of normal aging and the more serious AD. More details can be seen here Xiong et al. (2006).

With three ordinal diagnostic categories, ROC surface, analogous to ROC curve, as well as the volume under the ROC surface (VUS), analogous to AUC, have been proposed to assess diagnostic accuracy Xiong et al. (2006, 2007). Let S_1, S_2 and S_3 denote the scores resulting from a diagnostic test or biomarker and let F_1, F_2 and F_3 be the corresponding cumulative distribution functions for non-diseased, intermediate and diseased subjects, respectively. Assume the results of a diagnostic test are measured on a continuous scale and higher values indicate greater severity of the disease. Let $p_1 = F_1(c_1)$, $p_3 = 1 - F_3(c_3)$,

*US Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, MD 20993

†Department of Biostatistics, University at Buffalo, 706 Kimball Tower, 3435 Main Street, Buffalo, NY 14214

‡Division of Biostatistics, School of Medicine, Washington University in St. Louis, 660 South Euclid Avenue Box 8067, St. Louis, MO 63110

§Department of Medicine, University of Washington, 325 Ninth Avenue, Campus Box 359780, Seattle, WA 98104

where c_1 and c_3 are threshold values ($c_1 < c_3$), be the true classification rates for non-diseased and diseased category, respectively. Then the probability that a randomly selected subject from intermediate group has a score between c_1 and c_3 is

$$p_2 = F_2(c_3) - F_2(c_1) = F_2\left[F_3^{-1}(1-p_3)\right] - F_2\left[F_1^{-1}(p_1)\right]. \quad (1)$$

The probability p_2 is guaranteed positive due to the imposed order restriction of $c_1 < c_3$ such that $p_3 < 1 - F_3[F_1^{-1}(p_1)]$.

For a pair of thresholds $(c_1,\, c_3)$, we could compute the true classification rate p_2 for the intermediate category. The triplet $(p_1,\, p_2,\, p_3)$, where $p_2 = p_2(p_1,\, p_3)$ being a function of $(p_1,\, p_3)$, would produce an ROC surface in the three-dimensional space for all possible $(c_1,\, c_3) \in \mathbb{R}^2$. The volume under the ROC surface (VUS) is then defined as

$$VUS = \int_0^1 \int_0^{1-F_3\left[F_1^{-1}(p_1)\right]} F_2\left[F_3^{-1}(1-p_3)\right] - F_2\left[F_1^{-1}(p_1)\right] dp_3 dp_1. \quad (2)$$

This is a generalization of the AUC for a binary classification. As in Xiong *et al.* Xiong et al. (2006), under the normality assumption $S_d \sim N(\mu_d, \sigma_d^2)$, $d = 1,\ 2,\ 3$, the VUS can be further expressed as

$$VUS = \int_{-\infty}^{\infty} \Phi\left(as - b\right) \Phi\left(-cs + d\right) \phi\left(s\right) ds, \quad (3)$$

where $a = \sigma_2/\,\sigma_1$, $b = (\mu_1 - \mu_2)/\,\sigma_1$, $c = \sigma_2/\,\sigma_3$, $d = (\mu_3 - \mu_2)/\,\sigma_3$, $\Phi\left(\cdot\right)$ is the standard normal distribution function, and $\phi\left(\cdot\right)$ is the standard normal density function. One could show that VUS is mathematically equivalent to the probability $P(S_1 < S_2 < S_3)$, where S_1, S_2 and S_3 are scores for randomly selected individuals from corresponding diagnostic category. For a useless test (when S_1, S_2 and S_3 have identical distributions), VUS is 1/6. Notice that the unbiased nonparametric Mann-Whitney U statistic of the VUS is given by

$$U = \frac{1}{n_1 n_2 n_3} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} I\left(S_{1i} < S_{2j} < S_{3k}\right), \quad (4)$$

where n_1, n_2 and n_3 are the sample sizes for non-diseased, intermediate and diseased subjects, respectively, and $I\left(\cdot\right)$ stands for the indicator function.

The problem of finding optimal combinations of diagnostic tests and biomarkers with binary diagnostic categories has been well addressed in literatures. Su and Liu Su and Liu (1993) derived an optimal linear combination that maximizes the AUC when the biomarkers in the non-diseased and diseased category follow normal distributions. Without assumptions on the distributions of the biomarkers, Pepe and Thompson Pepe and Thompson (2000) considered an empirical solution of the optimal linear combination that maximizes the Mann-Whitney statistic. However, when the number of biomarkers is large, this approach is computationally formidable. Recently, Liu *et al.* Liu et al. (2011) developed a min-max combination approach which only involves searching for a single coefficient that maximizes the Mann-Whitney U statistic of AUC.

While several studies address optional selection of weights for binary outcomes, the problem of finding the optimal linear combinations has rarely been addressed for outcomes with three ordinal diagnostic categories. Nevertheless, it is of paramount importance to develop such combinations for biomarkers with three disease categories for the purpose of maximizing diagnostic accuracy. The importance can be seen through the data example on Alzheimer's disease. Since Alzheimer's disease is irreversible and no pharmaceutical treatments are effective for late stages, it is critical to accurately diagnose Alzheimer's

disease at its early stage. However, as presented in Xiong *et al.* Xiong et al. (2006), none of the current psychometric tests can be considered as excellent with the estimated VUS ranging from 0.522 to 0.752. Therefore, it is important to develop a composite score derived from a linear combination of biomarkers for better diagnostic accuracy.

The goal of this manuscript is two-fold: 1) to present parametric and nonparametric combination approaches for the purpose of maximizing the most important diagnostic accuracy index for three-category outcomes, namely, the volume under the ROC surface (VUS); 2) to empirical compare the performance of the proposed methods. The rest of our article is organized as follows. In Section 2, two existing combination methods for binary outcomes (i.e., the logistic regression approach and the min-max approach) are extended to maximize VUS for three-category outcomes. In Section 3, a new parametric approach and a new nonparametric approach are proposed. Simulation studies are presented in Section 4 for investigating the performance of different combination methods in maximizing VUS. In Section 5, the proposed approaches as well as the extensions are applied to a real data set of 118 subjects from a cohort study in early stage Alzheimer's disease (AD) from the Washington University Knight Alzheimer's Disease Research Center to combine diagnostic tests to increase the accuracy of discriminating different stages of AD. A broader discussion on deriving linear combinations of diagnostic tests and biomarkers to improve the diagnostic accuracy is presented in Section 6.

2. Extensions of existing methods

Two existing methods for binary outcomes, namely, the logistic regression method and the min-max method, can be easily extended to outcomes with three ordinal disease categories. In the following, Section 2.1 presents notation, and Sections 2.2 & 2.3 will discuss these two extensions.

2.1 Notation

Suppose we have p diagnostic tests or biomarkers available on each individual. The diagnostic category is denoted as $D = d$, where $d = 1,\ 2,\ 3$ stands for non-diseased, intermediate and diseased subjects, respectively. Let

$$\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})\,,\ i = 1, 2, \ldots, n_1,$$

be the p-dimensional observed scores from a random sample of size n_1 in the non-diseased category,

$$\boldsymbol{Y}_j = (Y_{j1}, Y_{j2}, \ldots, Y_{jp})\,,\ j = 1, 2, \ldots, n_2,$$

be the p-dimensional observed scores from a random sample of size n_2 in the intermediate category, and

$$\boldsymbol{Z}_k = (Z_{k1}, Z_{k2}, \ldots, Z_{kp})\,,\ k = 1, 2, \ldots, n_3,$$

be the p-dimensional observed scores from a random sample of size n_3 in the diseased category. The data are often stacked together in a matrix form

$$\begin{pmatrix} \mathbf{1}_{n_1} & [\boldsymbol{X}_i]_{n_1 \times p} \\ \mathbf{2}_{n_2} & [\boldsymbol{Y}_j]_{n_2 \times p} \\ \mathbf{3}_{n_3} & [\boldsymbol{Z}_k]_{n_3 \times p} \end{pmatrix},$$

where the first column indicates the diagnostic category and the other p columns form the matrix of observed scores concatenated from $\boldsymbol{X}_i$, $\boldsymbol{Y}_j$ and $\boldsymbol{Z}_k$ by row. For simplicity, we use $\boldsymbol{M}_p$ to denote p-variate observed scores for an individual from any diagnostic category.

2.2 The cumulative logistic regression approach

When a logistic regression model is used to model a binary outcome, linear coefficients for multiple predictors can be obtained. With three ordinal diagnostic categories, the cumulative logistic model has the form

$$\log \frac{P(D=1)}{P(D=2)+P(D=3)} = \alpha_0 + \boldsymbol{M}_p\boldsymbol{c},$$
$$\log \frac{P(D=1)+P(D=2)}{P(D=3)} = \beta_0 + \boldsymbol{M}_p\boldsymbol{c},$$

where $\boldsymbol{c}$ is a vector coefficient of length p and α_0, β_0 are two intercepts. For modeling a outcome with three or more categories, the multinomial logistic regression is also frequently used, although it is known if the outcome variable is truly ordered, which is the case in this article, cumulative logistic regression will make the model more parsimonious. Also, the multinomial logistic regression would produce more than one set of vector coefficients for predictor variables, which is meaningless for the purpose of combinations. Therefore, the performance of the combined marker using $\boldsymbol{c}$ obtained from cumulative logistic regression is investigated.

For modeling a binary outcome, the logistic regression is used to maximize the logistic likelihood function. For such model, Jin and Lu Jin and Lu (2009) proved that $\boldsymbol{c}$ from a fitted logistic regression is the optimal linear combination in the sense that it provides the highest sensitivity uniformly over the entire range of specificity and therefore yields the largest AUC among all possible linear combinations. This impressive result, however, depends on the strong assumption that the binary response variable (i.e., disease status) is generated through a link function of predictors. As a matter of fact, in practice, disease status is not generated this way. Usually a binary gold standard is used to determine disease status and multiple biomarkers are measured without knowing any information on disease status. Furthermore, this result does not assume any joint distributions for multiple predictors. Therefore, it can not include Su and Liu's Su and Liu (1993) method as a special case, in which multivariate normality is a fundamental assumption.

For three-category outcomes, the result from Jin and Lu Jin and Lu (2009) has not been extended to three-category case. Despite the lack of analytical results, cumulative logistic regression still offers a possible combination method for the scenarios with three-category outcomes. Therefore, it is of interest to investigate the performance of the combination of biomarkers using $\boldsymbol{c}$ from a fitted cumulative logistic regression for the purpose of maximizing the VUS.

2.3 The min-max combination approach

With binary diagnostic categories, Pepe and Thompson Pepe and Thompson (2000) proposed to estimate the optimal linear combination coefficient $\boldsymbol{c}$ by maximizing the Mann-Whitney U statistic (i.e., the empirical estimate of AUC) as follows,

$$U(\boldsymbol{c}) = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} I\left(c_1X_{i1}+\cdots+c_pX_{ip} < c_1Y_{j1}+\cdots+c_pY_{jp}\right), \qquad (5)$$

where $I(\cdot)$ stands for the indicator function. Pepe and Thompson Pepe and Thompson (2000) also pointed out that since the Mann-Whitney statistic estimate of AUC is not a continuous function of $\boldsymbol{c}$, a search rather than a derivative-based method is required for this maximization. It means general-purpose optimization algorithms such as conjugate-gradient or Newton-type methods are not appropriate for this maximization. They illustrated the idea with an application involving only two markers. In that case, the computation

is relatively easy. However, when the number of markers is large, i.e., ≥ 3, this approach is computationally inaccessible.

To address such computational difficulty, Liu *et al.* Liu et al. (2011) proposed a non-parametric min-max approach that linearly combines only the minimum and maximum values of the p markers to maximize the AUC, i.e.,

$$U(c) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I\left(X_{i,\max} + cX_{i,\min} < Y_{j,\max} + cY_{j,\min}\right), \tag{6}$$

where

$$X_{i,\max} = \max_{1 \leq l \leq p} X_{il},\ X_{i,\min} = \min_{1 \leq l \leq p} X_{il};$$

and

$$Y_{j,\max} = \max_{1 \leq l \leq p} Y_{jl},\ Y_{j,\min} = \min_{1 \leq l \leq p} Y_{jl}.$$

Such a combination only involves searching for a single combination coefficient and thus is computationally efficient. They showed under certain circumstances, the proposed min-max combination may yield larger AUC than empirical search of $\boldsymbol{c}$ by Pepe and Thompson Pepe and Thompson (2000). This min-max combination approach can be easily extended to the cases with three ordinal diagnostic categories by maximizing

$$U(c) = \frac{1}{n_1 n_2 n_3} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} I\left(X_{i,\max} + cX_{i,\min} < Y_{j,\max} + cY_{j,\min} < Z_{k,\max} + cZ_{k,\min}\right),$$

where $X_{i,\max}, X_{i,\min}, Y_{j,\max}$ and $Y_{j,\min}$ are defined as above and

$$Z_{k,\max} = \max_{1 \leq l \leq p} Z_{kl},\ Z_{k,\min} = \min_{1 \leq l \leq p} Z_{kl}.$$

3. The Proposed Methods

In this section, two new approaches for linearly combining markers to improve the VUS will be proposed. The first approach requires the assumption of multivariate normality and is designed to maximize the *penalized/scaled* stochastic distance between three ordinal diagnostic categories. The second distribution-free stepwise approach aims to find the optimal combination empirically by maximizing the Mann-Whitney statistic of the VUS at each step.

3.1 The penalized/scaled stochastic distance method based on normality

Assume that $\boldsymbol{X}_i, \boldsymbol{Y}_j, \boldsymbol{Z}_k$ follow a multivariate normal distribution $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ and $N_p(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$, respectively. The problem of interest is to obtain a vector combination coefficient $\boldsymbol{c}$ such that the univariate scores $S_1 = \boldsymbol{X}_i\boldsymbol{c},\ S_2 = \boldsymbol{Y}_j\boldsymbol{c}$ and $S_3 = \boldsymbol{Z}_k\boldsymbol{c}$ have the largest overall discriminating ability to classify subjects into their corresponding disease category, in this case, yielding the largest VUS. Notice that under normality assumption, S_d $(d = 1, 2, 3)$ follows a univariate normal distribution $N(\boldsymbol{c}'\boldsymbol{\mu}_d, \boldsymbol{c}'\boldsymbol{\Sigma}_d\boldsymbol{c})$.

Because the VUS is equal to $P(S_1 < S_2 < S_3)$, where S_1, S_2 and S_3 are univariate scores after combination for a randomly selected individual from each diagnostic category, it is reasonable to conclude that the larger stochastic distance between S_d $(d = 1, 2, 3)$, the larger VUS would be. Due to the fact that mean and variance completely characterize the normal distribution, we will define stochastic distance between normally distributed random variables based on functions of mean and variance.

For $S_d \sim N(\boldsymbol{c}'\boldsymbol{\mu}_d, \boldsymbol{c}'\boldsymbol{\Sigma}_d\boldsymbol{c})$ $(d = 1, 2, 3)$, $\sum_{d=1}^{3}(\boldsymbol{c}'\boldsymbol{\mu}_d - \boldsymbol{c}'\overline{\boldsymbol{\mu}})^2$ measures the between group variation, where $\overline{\boldsymbol{\mu}}$ is the mean of $\boldsymbol{\mu}_d$'s, and $\sum_{d=1}^{3}\boldsymbol{c}'\boldsymbol{\Sigma}_d\boldsymbol{c}$ measures the total within group variation. In an ideal situation, we want the quantity $\sum_{d=1}^{3}(\boldsymbol{c}'\boldsymbol{\mu}_d - \boldsymbol{c}'\overline{\boldsymbol{\mu}})^2$ as large as possible while at the same time keep $\sum_{d=1}^{3}\boldsymbol{c}'\boldsymbol{\Sigma}_d\boldsymbol{c}$ the minimal, because these are two necessary conditions for large separation of distributions underlying $S_d, d = 1, 2, 3$. An intuitive penalized stochastic distance (PSD) could be defined as

$$PSD = \sum_{d=1}^{3}\left(\boldsymbol{c}'\boldsymbol{\mu}_d - \boldsymbol{c}'\overline{\boldsymbol{\mu}}\right)^2 - \sum_{d=1}^{3}\boldsymbol{c}'\boldsymbol{\Sigma}_d\boldsymbol{c}, \tag{7}$$

such that $\boldsymbol{c}$ maximizing $\sum_{d=1}^{3}(\boldsymbol{c}'\boldsymbol{\mu}_d - \boldsymbol{c}'\overline{\boldsymbol{\mu}})^2$ and simultaneously minimizing $\sum_{d=1}^{3}\boldsymbol{c}'\boldsymbol{\Sigma}_d\boldsymbol{c}$ may be obtained once by maximizing PSD. With some rearrangement,

$$PSD = \boldsymbol{c}'\left[\sum_{d=1}^{3}(\boldsymbol{\mu}_d - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_d - \overline{\boldsymbol{\mu}})' - \sum_{d=1}^{3}\boldsymbol{\Sigma}_d\right]\boldsymbol{c}.$$

Lemma 3.1 *Let $\boldsymbol{A}$ be a $p \times p$ real symmetric matrix with (real) eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ and a corresponding set of orthonormal eigenvectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_p$, i.e., $\boldsymbol{u}_i'\boldsymbol{u}_j = I(i{=}j)$, where $I(\cdot)$ stands for the indicator function, such that $\boldsymbol{A}\boldsymbol{u}_i = \lambda_i\boldsymbol{u}_i$. Then for any $\boldsymbol{x} \in \Re^p$ and $\boldsymbol{x} \neq \boldsymbol{0}$, $\max_{\|\boldsymbol{x}\|=1}\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \lambda_1$, and the maximum occurs when $\boldsymbol{x} = \boldsymbol{u}_1$.*

Lemma 3.1 directly follows from Raleigh-Ritz Theorem Golub and van der Vorst (2000). Therefore, the $\boldsymbol{c}$ which maximizes PSD is the eigenvector corresponding to the largest eigenvalue of $\left[\sum_{d=1}^{3}(\boldsymbol{\mu}_d - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_d - \overline{\boldsymbol{\mu}})' - \sum_{d=1}^{3}\boldsymbol{\Sigma}_d\right]$. Notice that it is not necessary to normalize the eigenvector to obtain $\boldsymbol{c}$ as indicated in Lemma 3.1, because the eigenvectors are unique apart from a scalar, and the VUS associated with the linear combination $\boldsymbol{c}$ is invariant to a scaling constant.

However, this newly defined penalized stochastic distance (PSD) might have some potential problems. For example, in an extreme case, $\sum_{d=1}^{3}(\boldsymbol{\mu}_d - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_d - \overline{\boldsymbol{\mu}})' - \sum_{d=1}^{3}\boldsymbol{\Sigma}_d$ could be singular. For this reason, we also consider a scaled stochastic distance (SSD) defined as follows,

$$SSD = \frac{\sum_{d=1}^{3}\left(\boldsymbol{c}'\boldsymbol{\mu}_d - \boldsymbol{c}'\overline{\boldsymbol{\mu}}\right)^2}{\sum_{d=1}^{3}\boldsymbol{c}'\boldsymbol{\Sigma}_d\boldsymbol{c}} = \frac{\boldsymbol{c}'\left[\sum_{d=1}^{3}(\boldsymbol{\mu}_d - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_d - \overline{\boldsymbol{\mu}})'\right]\boldsymbol{c}}{\boldsymbol{c}'\left[\sum_{d=1}^{3}\boldsymbol{\Sigma}_d\right]\boldsymbol{c}}, \tag{8}$$

such that, again, $\boldsymbol{c}$ maximizing $\sum_{d=1}^{3}(\boldsymbol{c}'\boldsymbol{\mu}_d - \boldsymbol{c}'\overline{\boldsymbol{\mu}})^2$ and simultaneously minimizing $\sum_{d=1}^{3}\boldsymbol{c}'\boldsymbol{\Sigma}_d\boldsymbol{c}$ may be obtained by maximizing SSD. This definition of SSD is similar to a natural extension of Fisher discriminant for multi-category linear discriminant analysis Johnson and Wichern (2002), except that here we do not assume the common variance matrix $\boldsymbol{\Sigma}_d = \boldsymbol{\Sigma}, d = 1, 2, 3$, since it is a too strong assumption across three ordinal diagnostic categories.

Lemma 3.2 *Let $\boldsymbol{A}$ be a real $p \times p$ symmetric matrix, and let $\boldsymbol{B}$ be any $p \times p$ positive definite matrix. Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ be the eigenvalues of $\boldsymbol{B}^{-1}\boldsymbol{A}$ with a corresponding set of right eigenvectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_p$ (all of which are real), i.e., $\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{u}_i = \lambda_i\boldsymbol{u}_i$. Then for any $\boldsymbol{x} \in \Re^p$ and $\boldsymbol{x} \neq \boldsymbol{0}$, $\max_{\boldsymbol{x}\neq\boldsymbol{0}}\dfrac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{B}\boldsymbol{x}} = \lambda_1$, with the bounds being attained when $\boldsymbol{x} = \boldsymbol{u}_1$. In particular, for any $\boldsymbol{a}$ we have $\max_{\boldsymbol{x}\neq\boldsymbol{0}}\dfrac{\boldsymbol{x}'\boldsymbol{a}\boldsymbol{a}'\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{B}\boldsymbol{x}} = \boldsymbol{a}'\boldsymbol{B}^{-1}\boldsymbol{a}$, and the maximum occurs when $\boldsymbol{x} = \boldsymbol{B}^{-1}\boldsymbol{a}$, apart from some scaling constant.*

Lemma 3.2 follows from Theorem 6.59 (Seber, pp. 109–110) Seber (2008). To obtain the maximum of SSD in Equation (8), $\boldsymbol{c}$ can be obtained as the eigenvector corresponding to the largest eigenvalue of $\left(\sum_{d=1}^{3}\boldsymbol{\Sigma}_d\right)^{-1}\left[\sum_{d=1}^{3}\left(\boldsymbol{\mu}_d-\overline{\boldsymbol{\mu}}\right)\left(\boldsymbol{\mu}_d-\overline{\boldsymbol{\mu}}\right)'\right]$ based on Lemma 3.2. In practice, the mean and variance for each disease category can be estimated from the data and then the estimates can be substituted into the above formulas for calculating the combination coefficient $\boldsymbol{c}$.

Remark: For the scenarios with binary disease status,

$$SSD=\frac{\boldsymbol{c}'\left[\sum_{d=1}^{2}\left(\boldsymbol{\mu}_d-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\right)\left(\boldsymbol{\mu}_d-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\right)'\right]\boldsymbol{c}}{\boldsymbol{c}'\left[\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2\right]\boldsymbol{c}}=\frac{\boldsymbol{c}'\left[\frac{1}{2}\left(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1\right)\left(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1\right)'\right]\boldsymbol{c}}{\boldsymbol{c}'\left[\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2\right]\boldsymbol{c}},$$

the maximum occurs when $\boldsymbol{c}=\left(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2\right)^{-1}\left(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1\right)/\sqrt{2}$ from Lemma 3.2. Apart from the constant $1/\sqrt{2}$, this result is exactly the same as that in Su and Liu Su and Liu (1993). In this sense, our proposed SSD method coincides with Su and Liu's method for binary disease outcomes.

Generally speaking, the term $\sum_{d=1}^{3}\left(\boldsymbol{\mu}_d-\overline{\boldsymbol{\mu}}\right)\left(\boldsymbol{\mu}_d-\overline{\boldsymbol{\mu}}\right)'$ cannot be written as $\boldsymbol{a}\boldsymbol{a}'$ for some $\boldsymbol{a}$, thus a closed-form solution does not exist. However, eigenvalues and eigenvectors of a square matrix can be easily computed using statistical packages, such as *eigen()* in R and *call eigen()* in SAS/IML, and therefore obtaining the vector combination coefficient $\boldsymbol{c}$ using these proposed PSD or SSD methods is numerically straightforward.

3.2 The distribution-free stepwise approach

The above approach makes use of the assumption of multivariate normality. We now consider maximizing VUS without normality assumption. The empirical estimate of VUS of the combination $\boldsymbol{c}$ is

$$U\left(\boldsymbol{c}\right)=\frac{1}{n_1n_2n_3}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}I\left(c_1X_{i1}+\cdots+c_pX_{ip}<c_1Y_{j1}+\cdots+c_pY_{jp}<c_1Z_{k1}+\cdots+c_pZ_{kp}\right).$$

This is a three-category generalization of Pepe and Thompson Pepe and Thompson (2000). When the number of markers p is large, i.e., ≥ 3, the empirical search for $\boldsymbol{c}$ is computationally inaccessible. The nonparametric min-max procedure by Liu *et al.* Liu et al. (2011) is easy to implement; however, it comes with a few drawbacks: 1) feasibility might be an issue when not all biomarkers are measured on the same scale; 2) the approach may be an inefficient use of the data as it only considers the minimum and maximum values; 3) interpretation of the estimated combination coefficient is difficult.

To overcome all the shortcomings of the current existing nonparametric combination methods, we will develop a distribution-free approach that combines the diagnostic tests or the scores of all the biomarkers in a stepwise fashion. Two stepwise proceeding procedures are considered, i.e., step-down and step-up, which we describe in details in the following, using the step-down procedure as an example:

1. Estimate VUS for each of p diagnostic tests or biomarkers based on the Mann-Whitney statistic by Equation (4);

2. Assign the order from 1 to p for each diagnostic test or biomarker based on their estimated VUS from the largest to the smallest.

3. Combine the first two markers (i.e., markers with first two largest VUS) using empirical search for combination coefficients presented by Pepe and Thompson Pepe and Thompson (2000).

4. Having derived the univariate composite score in Step 3 by linearly combining first two markers, combine it with the third marker (i.e., marker with the third largest VUS) using empirical searching combination again.

5. Proceed in this fashion until the ordered p^{th} marker (i.e., marker with smallest VUS) is included in the linear combination.

The estimated combination coefficient by searching needs to be saved in each step, and in the end, the order of p's combination coefficients needs to be adjusted to match their corresponding markers. The step-up procedure is exactly the same as the step-down one except that in Step 2, the order from 1 to p for each diagnostic test or biomarker is assigned based on their estimated VUS from the smallest to the largest.

Given p biomarkers, there exist $p!$ ways of permuting them, and hence there exist $p!$ stepwise procedures. The proposed step-down and step-up procedures are just two out of those $p!$ ways. However, when p is relatively large, it is not feasible to carry out all the $p!$ ways. For example, when $p \geq 50$, there exist $50! \approx 3 \times 10^{64}$ stepwise procedures. Another reason that we only consider step-down and step-up procedures is rooted in order restricted inference Robertson et al. (1988), where it is argued that any other stepwise method selecting different proceeding orders would have performance somewhere in between the step-down and step-up procedures.

The advantages of our proposed stepwise approach are: 1) it is distribution-free and therefore it is robust; 2) it is easy to implement with computer iterations and therefore it offers a relief from the computational burden in the empirical search of combination coefficients in p-dimensional space as $p > 2$ as encountered in Pepe and Thompson Pepe and Thompson (2000); 3) simulation studies in Section 4 demonstrate that the stepwise approach (especially the step-down one) may outperform the other methods under some scenarios, and for other scenarios, its performance is comparable to that of other methods.

4. Simulation Studies

Simulations are conducted to investigate the performance of the different combination methods as it is difficult, if not impossible, to analytically evaluate the performance of the aforementioned methods. For $\overline{\boldsymbol{\mu}}$ in equations (7) and (8), both weighted and un-weighted versions are calculated as follows: $\overline{\boldsymbol{\mu}}^{w} = (n_1\boldsymbol{\mu}_1 + n_2\boldsymbol{\mu}_2 + n_3\boldsymbol{\mu}_3)/(n_1 + n_2 + n_3)$ and $\overline{\boldsymbol{\mu}}^{uw} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2 + \boldsymbol{\mu}_3)/3$. Overall, we empirically investigate the performance of eight approaches, namely, the scaled stochastic distance method with $\overline{\boldsymbol{\mu}}^{w}$ (SSD1), the scaled stochastic distance method with $\overline{\boldsymbol{\mu}}^{uw}$ (SSD2), the penalized stochastic distance method with $\overline{\boldsymbol{\mu}}^{w}$(PSD1), the penalized stochastic distance method with $\overline{\boldsymbol{\mu}}^{uw}$ (PSD2), the step-down procedure which proceeds from the marker with largest VUS to the one with smallest VUS (SW1), the step-up procedure which proceeds from the marker with smallest VUS to the one with largest VUS (SW2), the min-max approach extended to three diagnostic categories (Min-Max), and the linear combination coefficients from cumulative logistic regression (Cum-Logistic).

To investigate the performance of all eight approaches enpirically, six different settings of the joint distributions of five diagnostic tests ($p = 5$) are considered. For each setting, multivariate observations are generated from the underlying distributions with different sample sizes. The univariate composite scores S_{1i}, S_{2j} and S_{3k} are calculated by combining

the observed data using the estimated $\boldsymbol{c}$ from a specific combination method; and then VUS of the combined marker is estimated using the unbiased Mann-Whitney statistic in Equation (4). For each setting, 10,000 Monte Carlo repetitions are conducted. For each method, the mean VUS as well as the chance of obtaining the largest VUS across 10,000 Monte Carlo repetitions are reported in Tables 1–6.

4.1 Multivariate normal distributions with equal variance

Data from multivariate normal distributions with different mean vectors and equal variance matrices corresponding to three ordinal diagnostic categories are generated with

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \end{pmatrix}, \ \boldsymbol{\mu}_2 = \begin{pmatrix} 0.8 \\ 1.1 \\ 1.4 \\ 1.7 \\ 2.0 \end{pmatrix}, \ \boldsymbol{\mu}_3 = \begin{pmatrix} 1.6 \\ 2.2 \\ 2.8 \\ 3.4 \\ 4.0 \end{pmatrix},$$

and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = 0.7 \times \boldsymbol{I}_{5\times5} + 0.3 \times \boldsymbol{J}_{5\times5}$, $0.5 \times \boldsymbol{I}_{5\times5} + 0.5 \times \boldsymbol{J}_{5\times5}$, $0.3 \times \boldsymbol{I}_{5\times5} + 0.7 \times \boldsymbol{J}_{5\times5}$, where $\boldsymbol{I}$ and $\boldsymbol{J}$ stand for an identity matrix and a matrix with all elements equal to 1, respectively. These three different covariance matrices correspond to scenarios with low, medium, and high correlation respectively, and the corresponding simulation results are presented in Tables 1–3.

Overall speaking, the simulation results presented in Tables 1–3 show that SW1, SSD1, SSD2, and Cum-Logistic have better performance than the other approaches. The performance of each method somewhat depends on the correlation. When correlation is relatively small, Table 1 ($\rho = 0.3$) shows that SW1 performs much better than SSD1, SSD2 or Cum-Logistic. As correlation increases from small to large, Table 3 ($\rho = 0.7$) shows the performance of SW1 is slightly inferior to that of SSD1 or Cum-Logistic in view of the mean VUS. Under the setting with correlation $\rho = 0.5$ (Table 2), all of SW1, SSD1, SSD2, Cum-Logistic have comparable good performance.

Although the method using cumulative logistic regression might work well for certain scenarios, there exist some numerical difficulties with fitting cumulative logistic regression model. The iterative algorithms for maximum likelihood estimates of the model parameters can easily fail to converge, especially when the sample sizes are small. For fair comparisons, those ill-posed Monte Carlo samples are marked and excluded for calculating the mean VUS corresponding to Cum-Logistic.

4.2 Multivariate normal distributions with unequal variance

Now we consider multivariate normal distributions with different mean vectors and unequal variance matrices corresponding to three ordinal diagnostic categories. The mean vectors are the same as in Section 4.1, with variance matrices set as follows,

$\boldsymbol{\Sigma}_1 = 0.7 \times \boldsymbol{I}_{5\times5} + 0.3 \times \boldsymbol{J}_{5\times5}$
$\boldsymbol{\Sigma}_2 = 0.5 \times \boldsymbol{I}_{5\times5} + 0.5 \times \boldsymbol{J}_{5\times5}$
$\boldsymbol{\Sigma}_3 = 0.3 \times \boldsymbol{I}_{5\times5} + 0.7 \times \boldsymbol{J}_{5\times5}$

As shown in Table 4, the performances of SSD1, SSD2, SW1 and Cum-Logistic under this setting have good and comparable performance.

4.3 Multivariate log-normal distributions with unequal variance-covariance

In this section, we would like to investigate the diagnostic accuracy of the combined marker from different methods, assuming that multiple biomarkers follow multivariate log-normal distributions, that is, the log-transformed scores are multivariate normally distributed. Data

are first generated from the multivariate normal setting in Section 4.2 and then exponentiated to get the multivariate log-normal observations.

In this case, the normality assumption does not hold and the normal-based approaches such as SSD1 do not work at all, which is expected, as sample means and variance matrices under this setting cannot measure the location and variation correctly for non-normal data. From Table 5, it is suggested that SW1 proceeding from the marker with largest VUS to the marker with smallest VUS dominate the other methods.

4.4 Multivariate normal-χ^2-lognormal-exponential-gamma distributions via normal copula

We further investigate the performances of different methods assuming the p-variate scores follow multivariate normal-χ^2-lognormal-exponential-gamma distributions coupled together via normal copula Kojadinovic and Yan (2010) with exchangeable correlations ρ being 0.3, 0.5 and 0.7 for non-diseased, intermediate and diseased category, respectively. The marginal distributions of p biomarkers for non-diseased, intermediate and diseased subjects are chosen as follows, respectively,

$$\begin{pmatrix} N(0.1,1) \\ \chi^2_{0.1} \\ LN(-2.80,1) \\ \exp(0.1) \\ \Gamma(0.1,1) \end{pmatrix}, \quad \begin{pmatrix} N(0.8,1) \\ \chi^2_{1.1} \\ LN(-0.16,1) \\ \exp(1.7) \\ \Gamma(2.0,1) \end{pmatrix}, \quad \begin{pmatrix} N(1.6,1) \\ \chi^2_{2.2} \\ LN(0.53,1) \\ \exp(3.4) \\ \Gamma(4.0,1) \end{pmatrix}.$$

Under this setting, the mean structures are exactly the same as in Section 4.1. From Table 6, we can see the step-down procedure (SW1) proceeding from the marker with largest VUS to the one with smallest VUS is far more superior than all the other methods.

In summary, out of all the methods considered, the step-down procedure (SW1) is a good choice for combining multiple biomarkers, followed by the scaled stochastic distance method (SSD1 and SSD2), the cumulative logistic regression method, SW2, the penalized stochastic distance method (PSD1 and PSD2), and Min-Max. While SW1 is not based on normality, it requires $p-1$ nonparametric searching steps. On the other hand, SSD1 (or SSD2) requires normal assumption, but they are more efficient with large numbers of biomarkers.

5. Analysis of Data Example

In this section, all eight approaches investigated in simulation studies are applied to a real data set of 118 subjects from a cohort study in early stage Alzheimer's disease (AD) from the Washington University Knight Alzheimer's Disease Research Center to combine several psychometric tests for larger discriminating ability, i.e., larger VUS, than any individual psychometric test scores.

Each individual was assessed by experienced clinicians. The diagnosis of AD was based on the Clinical Dementia Rating (CDR) according to published rules Morris (1993). In this application, we are concentrating on three diagnostic categories, non-demented (CDR 0, 45 individuals), very mildly demented (CDR 0.5, 44 individuals), and mildly demented (CDR 1, 29 individuals). Approximately 2 weeks after the clinical evaluation, subjects also completed a battery of psychometric tests. Episodic memory, which involves the recollection of specific events, situations and experiences, e.g., first day of school or graduation, was assessed by 5 of those psychometric tests, the Logical Memory (LM), Digit Span Forward (DSF), Digit Span Backward (DSB), Associate Learning subtests of the

Wechsler Memory Scale (WMS) Wechsler and Stone (1973) and the Visual Retention Test (Form C, 10-s exposure) (VRT) Benton (1963). Xiong *et al.* Xiong et al. (2006) reported the estimated VUS for these 5 psychometric tests: 0.724 (LM), 0.522 (DSF), 0.599 (DSB), 0.630 (WMS), and 0.587 (VRT).

The linear combinations with associated VUS from the SSD1, SSD2, PSD1, PSD2, Cum-Logistic, SW1 and SW2 methods are provided in the following, where the combination coefficient corresponding to LM is set to 1 to guarantee a unique solution.

	LM	DSF	DSB	VRT	WMS	(VUS)
SSD1	1.0000	0.1533	0.2272	0.3915	0.0765	(0.8077)
SSD2	1.0000	0.1513	0.2219	0.3924	0.0747	(0.8066)
PSD1	1.0000	0.4863	0.6464	0.7902	0.7121	(0.8106)
PSD2	1.0000	0.4742	0.6233	0.7810	0.6957	(0.8108)
Cum-Logistic	1.0000	0.1610	0.4396	0.2934	0.1654	(0.8138)
SW1	1.0000	0.1162	0.4830	0.1290	0.3558	(0.8296)
SW2	1.0000	0.0729	0.1553	0.0924	0.3360	(0.8235)

The min-max approach provides the following combination

$$1.0000\times\max\{LM, DSF, DSB, VRT, WMS\}+1.1956\times\min\{LM, DSF, DSB, VRT, WMS\}$$

with an estimated VUS of 0.7724 for the combined marker. The Shapiro-Wilk test for multivariate normality Royston (1982) returns p-values of < 0.0001, < 0.0001 and 0.0184 for non-diseased, intermediate and diseased category, respectively. Therefore, the results using the procedures based on normality (SSD1, SSD2, PSD1, PSD2) should not be interpreted. All 8 methods provide a linearly combined marker that yields a larger VUS than any of the original test and the step-down method (SW1) provides a linear combination with the largest VUS.

6. Discussion

In this article, we extend two existing combination approaches to deal with three ordinal diagnostic categories. We also propose two new types of linear combination methods to combine diagnostic tests or biomarkers to improve diagnostic accuracy measure, VUS. The first proposed normal-based approach requires only the estimated means and variance-covariances of multiple diagnostic tests for each diagnostic category to calculate the linear combination coefficients. Therefore, it is efficient with large numbers of biomarkers, which is quite common nowadays with high-throughput bioinformatics tools, for instance, microarray technologies. Under the normality assumption with moderate to large correlations, our simulations show the normal-based approach, especially SSD1, has relatively good performance in terms of obtaining a combined marker with the largest VUS. Recently, Zhang Zhang (2010) proposed to directly maximize the accuracy index VUS with three diagnostic categories under normality assumption. Although appealing, the mathematical equations for finding the derivatives are formidable. The author stated that the analytic solution to directly maximizing the VUS is not generally attainable. For this reason, the proposed normal-based approach may offer investigators an opportunity to combine the diagnostic tests and biomarkers for the disease processes with more than three ordinal categories. The second proposed approach is a stepwise approach which is distribution-free in nature and hence is robust with non-normal data. The computing effort and cost in obtaining the combination coefficient is significantly less than the empirical search in p-dimensional space Pepe and Thompson (2000). Our simulations show, for either non-normal data or normal

data with small correlations, the step-down procedure (SW1) proceeding from the marker with largest VUS to the marker with smallest VUS is a reasonable choice for biomarker combination. It is worthwhile to point out that both the stepwise approach and the normal-based approach could be easily generalized to diseases with more than three diagnostic categories. The cumulative logistic regression approach (Cum-Logistic) has great chance to produce a combined marker with largest VUS under normality assumption with large sample sizes. Note that one of the assumptions underlying cumulative logistic regression model is the proportional odds. This is to say, the coefficients that describe the relationship between, say, the lowest versus all higher categories of the response variable are the same as those that describe the relationship between the next lowest category versus all higher categories, etc. We recommend to test this assumption before applying this approach to the combination of markers. The min-max combination method is a fast one, although the performance is not as good as simulations indicated. It is interesting to explore if adding some other order statistics will improve the combination while maintaining its computational efficiency in the future research.

Some related research topics are currently under investigation. The methods explored here implicitly assume that the scaling metric of each of the biomarkers is linear. However, for some cognitive tests, this may not be the case, see Crane *et al.* Crane et al. (2008). It will be of great interest to determine whether some approaches that first produce a linear scaling metric for each biomarker and then apply the proposed methods may provide additional ability to distinguish among disease severity categories. Furthermore, it is also of interest to explore the performance of a generalized version of the cumulative logistic regression approach discussed in Section 2.2 without the proportional odds assumption.

References

Benton, A. L. (1963). *The Revised Visual Retention Test: Clinical and Experimental Applications.* Psychological Corporation: New York.

Crane, P. K., Narasimhalu, K., Gibbons, L. E., Mungas, D. M., Haneuse, S., Larson, E. B., Kuller, L., Hall, K., and van Belle, G. (2008). Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. *Journal of Clinical Epidemiology*, 61(10):1018–1027.

Etzioni, R., Kooperberg, C., Pepe, M., Smith, R., and Gann, P. H. (2003). Combining biomarkers to detect disease with application to prostate cancer. *Biostatistics*, 4:523–538.

Golub, G. H. and van der Vorst, H. A. (2000). Eigenvalue computation in the 20th century. *Journal of Computational and Applied Mathematics*, 123:35–65.

He, X. and Frey, E. C. (2008). The meaning and use of the Volume Under a Three-Class ROC Surface (VUS). *IEEE Transactions on Medical Imaging*, 27:577–588.

Jin, H. and Lu, Y. (2009). The optimal linear combination of multiple predictors under the generalized linear models. *Statistics & Probability Letters*, 79:2321–2327.

Johnson, R. and Wichern, D. (2002). *Applied Multivariate Statistical Analysis.* Prentice Hall Upper Saddle River, NJ.

Kojadinovic, I. and Yan, J. (2010). Modeling multivariate distributions with continuous margins using the copula R package. *Journal of Statistical Software*, 34:1–20.

Liu, C., Liu, A., and Halabi, S. (2011). A min-max combination of biomarkers to improve diagnostic accuracy. *Statistics in Medicine*, 30(16):2005–2014.

Morris, J. C. (1993). The clinical dementia rating (CDR): current version and scoring rules. *Neurology*, 43:1412–1414.

Pepe, M. S. and Thompson, M. L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics*, 1:123–140.

Robertson, T., Wright, F., Dykstra, R., and Robertson, T. (1988). *Order Restricted Statistical Inference.* Wiley New York.

Royston, J. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics*, 31:115–124.

Seber, G. (2008). *A Matrix Handbook for Statisticians.* Wiley-Interscience.

Sidransky, D. (2002). Emerging molecular markers of cancer. *Nature Reviews Cancer*, 2:210–219.

Su, J. Q. and Liu, J. S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association*, 88:1350–1355.

Thompson, M. L. and Zucchini, W. (1989). On the statistical analysis of ROC curves. *Statistics in Medicine*, 8(10):1277–1290.

Wechsler, D. and Stone, C. P. (1973). *Wechsler Memory Scale Manual.* Psychological Corporation: New York.

Xiong, C. J., 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.

Xiong, C. J., van Belle, G., Miller, J. P., Yan, Y., Gao, F., Yu, K., and Morris, J. C. (2007). A parametric comparison of diagnostic accuracy with three ordinal diagnostic groups. *Biometrical Journal*, 49:682–693.

Zhang, Y. Y. (2010). *ROC analysis in diagnostic medicine.* PhD thesis, National University of Singapore.

Table 1: Mean VUS and chance of obtaining largest VUS (under in parenthesis)

(n_1, n_2, n_3)	SSD1	SSD2	PSD1	PSD2	Cum-Logistic	SW1	SW2	Min-Max
$(20, 20, 20)$	0.9135	—	0.8916	—	0.9113	0.9216	0.9100	0.8566
	(0.194)	—	(0.010)	—	(0.183)	(0.511)	(0.085)	(0.015)
$(20, 30, 50)$	0.9095	0.9095	0.8883	0.8894	0.9079	0.9147	0.9051	0.8519
	(0.084)	(0.080)	(0.002)	(0.002)	(0.158)	(0.601)	(0.066)	(0.008)
$(30, 40, 50)$	0.9074	0.9074	0.8885	0.8889	0.9088	0.9111	0.9027	0.8504
	(0.084)	(0.084)	(0.001)	(0.001)	(0.201)	(0.576)	(0.050)	(0.004)
$(50, 50, 50)$	0.9057	—	0.8880	—	0.9072	0.9082	0.9006	0.8483
	(0.176)	—	(0.002)	—	(0.242)	(0.541)	(0.039)	(0.001)

Simulation setting: normal data with equal variance
$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = 0.7 \times \boldsymbol{I}_{5\times5} + 0.3 \times \boldsymbol{J}_{5\times5}$

Table 2: Mean VUS and chance of obtaining largest VUS (under in parenthesis)

(n_1, n_2, n_3)	SSD1	SSD2	PSD1	PSD2	Cum-Logistic	SW1	SW2	Min-Max
$(20, 20, 20)$	0.8951	—	0.8441	—	0.8944	0.8977	0.8759	0.8239
	(0.332)	—	(0.000)	—	(0.289)	(0.348)	(0.021)	(0.009)
$(20, 30, 50)$	0.8905	0.8905	0.8383	0.8412	0.8896	0.8913	0.8702	0.8188
	(0.156)	(0.154)	(0.000)	(0.000)	(0.290)	(0.386)	(0.008)	(0.007)
$(30, 40, 50)$	0.8879	0.8879	0.8396	0.8406	0.8896	0.8879	0.8677	0.8168
	(0.163)	(0.162)	(0.000)	(0.000)	(0.356)	(0.313)	(0.003)	(0.002)
$(50, 50, 50)$	0.8860	—	0.8395	—	0.8875	0.8849	0.8655	0.8147
	(0.342)	—	(0.000)	—	(0.411)	(0.245)	(0.001)	(0.000)

Simulation setting: normal data with equal variance
$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = 0.5 \times \boldsymbol{I}_{5\times5} + 0.5 \times \boldsymbol{J}_{5\times5}$

Table 3: Mean VUS and chance of obtaining largest VUS (under in parenthesis)

(n_1, n_2, n_3)	SSD1	SSD2	PSD1	PSD2	Cum-Logistic	SW1	SW2	Min-Max
$(20, 20, 20)$	0.9131	—	0.8104	—	0.9107	0.8931	0.8546	0.8301
	(0.530)	—	(0.000)	—	(0.369)	(0.091)	(0.002)	(0.007)
$(20, 30, 50)$	0.9090	0.9090	0.8011	0.8074	0.9070	0.8894	0.8485	0.8250
	(0.268)	(0.258)	(0.000)	(0.000)	(0.408)	(0.062)	(0.000)	(0.004)
$(30, 40, 50)$	0.9066	0.9066	0.8044	0.8065	0.9080	0.8872	0.8463	0.8228
	(0.235)	(0.235)	(0.000)	(0.000)	(0.502)	(0.027)	(0.000)	(0.001)
$(50, 50, 50)$	0.9049	—	0.8053	—	0.9064	0.8845	0.8440	0.8207
	(0.442)	—	(0.000)	—	(0.549)	(0.010)	(0.000)	(0.000)

Simulation setting: normal data with equal variance
$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = 0.3 \times \boldsymbol{I}_{5\times5} + 0.7 \times \boldsymbol{J}_{5\times5}$

SSD1: scaled stochastic distance method with $\overline{\boldsymbol{\mu}}$ accounting for unbalanced sample size
SSD2: scaled stochastic distance method with $\overline{\boldsymbol{\mu}}$ accounting no unbalanced information
PSD1: penalized stochastic distance method with $\overline{\boldsymbol{\mu}}$ accounting for unbalanced info
PSD2: penalized stochastic distance method with $\overline{\boldsymbol{\mu}}$ accounting no unbalanced info
SW1: step-down procedure (stepwise method proceeding from marker with largest VUS to smallest VUS)
SW2: step-up procedure (stepwise method proceeding from marker with smallest VUS to largest VUS)
Min-Max: min-max approach implemented for three diagnostic categories
Cum-Logistic: linear combination coefficients from cumulative logistic regression

Table 4: Mean VUS and chance of obtaining largest VUS (under in parenthesis)

(n_1, n_2, n_3)	SSD1	SSD2	PSD1	PSD2	Cum-Logistic	SW1	SW2	Min-Max
$(20, 20, 20)$	0.8954	—	0.8456	—	0.8943	0.8982	0.8766	0.7994
	(0.336)	—	(0.000)	—	(0.273)	(0.365)	(0.024)	(0.001)
$(20, 30, 50)$	0.8916	0.8916	0.8400	0.8428	0.8904	0.8925	0.8711	0.7937
	(0.167)	(0.159)	(0.000)	(0.000)	(0.278)	(0.388)	(0.008)	(0.000)
$(30, 40, 50)$	0.8884	0.8884	0.8412	0.8421	0.8900	0.8887	0.8684	0.7916
	(0.159)	(0.157)	(0.000)	(0.000)	(0.345)	(0.335)	(0.003)	(0.000)
$(50, 50, 50)$	0.8863	—	0.8411	—	0.8878	0.8855	0.8661	0.7892
	(0.336)	—	(0.000)	—	(0.398)	(0.265)	(0.001)	(0.000)

Simulation setting: normal data with unequal variance

Table 5: Mean VUS and chance of obtaining largest VUS (under in parenthesis)

(n_1, n_2, n_3)	SSD1	SSD2	PSD1	PSD2	Cum-Logistic	SW1	SW2	Min-Max
$(20, 20, 20)$	0.6727	—	0.5784	—	0.8378	0.8835	0.8678	0.7981
	(0.002)	—	(0.004)	—	(0.066)	(0.849)	(0.075)	(0.003)
$(20, 30, 50)$	0.7078	0.7066	0.4845	0.4208	0.8257	0.8769	0.8625	0.7931
	(0.000)	(0.000)	(0.001)	(0.001)	(0.033)	(0.932)	(0.031)	(0.001)
$(30, 40, 50)$	0.7082	0.7077	0.4303	0.4141	0.8323	0.8717	0.8590	0.7910
	(0.000)	(0.000)	(0.000)	(0.000)	(0.028)	(0.949)	(0.022)	(0.000)
$(50, 50, 50)$	0.7095	—	0.4099	—	0.8372	0.8674	0.8561	0.7887
	(0.000)	—	(0.000)	—	(0.027)	(0.957)	(0.014)	(0.000)

Simulation setting: multivariate log-normal data

Table 6: Mean VUS and chance of obtaining largest VUS (under in parenthesis)

(n_1, n_2, n_3)	SSD1	SSD2	PSD1	PSD2	Cum-Logistic	SW1	SW2	Min-Max
$(20, 20, 20)$	0.7808	—	0.7740	—	0.7829	0.8252	0.8116	0.6870
	(0.080)	—	(0.018)	—	(0.015)	(0.798)	(0.088)	(0.003)
$(20, 30, 50)$	0.7782	0.7780	0.7702	0.7733	0.7815	0.8115	0.8020	0.6775
	(0.019)	(0.017)	(0.002)	(0.004)	(0.018)	(0.895)	(0.044)	(0.000)
$(30, 40, 50)$	0.7771	0.7771	0.7723	0.7732	0.7879	0.8077	0.8000	0.6743
	(0.017)	(0.018)	(0.002)	(0.003)	(0.030)	(0.892)	(0.039)	(0.000)
$(50, 50, 50)$	0.7767	—	0.7724	—	0.7912	0.8050	0.7986	0.6709
	(0.024)	—	(0.006)	—	(0.043)	(0.886)	(0.041)	(0.000)

Simulation setting: normal-χ^2-lognormal-exponential-gamma copula data

SSD1: scaled stochastic distance method with $\overline{\boldsymbol{\mu}}$ accounting for unbalanced sample size
SSD2: scaled stochastic distance method with $\overline{\boldsymbol{\mu}}$ accounting no unbalanced information
PSD1: penalized stochastic distance method with $\overline{\boldsymbol{\mu}}$ accounting for unbalanced info
PSD2: penalized stochastic distance method with $\overline{\boldsymbol{\mu}}$ accounting no unbalanced info
SW1: step-down procedure (stepwise method proceeding from marker with largest VUS to smallest VUS)
SW2: step-up procedure (stepwise method proceeding from marker with smallest VUS to largest VUS)
Min-Max: min-max approach implemented for three diagnostic categories
Cum-Logistic: linear combination coefficients from cumulative logistic regression