Published in final edited form as:

Sci China Math. 2012 August ; 55(8): 1583–1595. doi:10.1007/s11425-012-4462-3.

A regression approach to ROC surface, with applications to Alzheimer's disease

JiaLiang LI^{1,2,3}, Andrew XiaoHua ZHOU³, and Jason P. FINE⁴

JiaLiang LI: stalj@nus.edu.sg

¹Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Singapore

²Duke-NUS Graduate Medical School, Singapore

³Singapore Eye Research Institute, Singapore

⁴Department of Biostatistics, University of Washington, USA

⁵Department of Biostatistics and Department of Statistics, University of North Carolina, Chapel Hill, USA

Abstract

We consider the estimation of three-dimensional ROC surfaces for continuous tests given covariates. Three way ROC analysis is important in our motivating example where patients with Alzheimer's disease are usually classified into three categories and should receive different category-specific medical treatment. There has been no discussion on how covariates affect the three way ROC analysis. We propose a regression framework induced from the relationship between test results and covariates. We consider several practical cases and the corresponding inference procedures. Simulations are conducted to validate our methodology. The application on the motivating example illustrates clearly the age and sex effects on the accuracy for Mini-Mental State Examination of Alzheimer's disease.

Keywords

receiver operating characteristic surface; volume under ROC surface; rank regression; transformation model; maximum likelihood estimation; bootstrap

1 Introduction

Research on statistical methods for the design and analysis of diagnostic accuracy investigations advanced in the recent decades in Zhou et al. [39] and Pepe [24]. The typical focus of these studies is the probability of correctly identifying patients in the right disease categories (e.g. diseased or non-diseased). Investigators seek to identify various factors that affect the test results and subsequently the accuracy measures related to the correct classification probability.

For test results defined on continuous scales, receiver operating characteristic (ROC) curves and areas under the ROC curves (AUC) are standard summaries of test accuracy. Very recently it has been recognized in literature that such a technique is generalizable to three-

category classification problem where disease status may involve more than two dimensions (e.g. mild, moderate and severe disease conditions). Mossman [20] introduced the three way ROC analysis and formulated various decision rules for medical diagnosis. Heckerling [8] discussed parametric estimation of three-dimensional ROC measures. Nakas and Yiannoutsos [22] considered the estimation of the volume under the ROC surface (VUS) and provided useful variance formula, using theories from U-statistics. Li and Fine [15] further explored the probabilistic interpretation of three-class ROC surface and VUS and applied VUS for gene selection. Li and Zhou [18] conducted a theoretical investigation of nonparametric and semiparametric estimators of the three-way ROC surfaces and provided the weak convergence results. Zhang and Li [37] further considered optimizing the VUS with a nonparametric approach.

An overall AUC or VUS is useful to describe the test accuracy for a homogeneous population. However, when the study population includes subjects with different characteristics, the ability of the test to classify the subject into the correct category may vary substantially for different types of subjects. In a recent study of Alzheimer's disease in the U.S., a diagnostic test to detect three stages of the disease has been applied to the patients. Diagnosticians expect the accuracy of the diagnostic test may differ for men and women and also for subjects with different ages. In practice, it may be more sensible to report VUS measures for subjects with common covariate values. In population health studies, age- or sex-adjusted parameters have been widely acknowledged for their importance. To our knowledge, there has been no statistical procedure available to incorporate covariates for high dimensional ROC analysis with more than two disease categories.

We contribute to the methodology of estimating covariate-adjusted VUS in this paper. A variety of regression models were proposed to study how covariates affect the diagnostic accuracy for two-category classification problems in Pepe [24]. Few methods have been generalized thereafter to address the diagnostic task where three different classes need to be distinguished. In this paper, we start by modeling the relationship between test results from the three categories and a set of explanatory variables and then derive the regression model for correct classification probability. The covariate-specific ROC surface and VUS are obtained subsequently. Note that because covariate effects are specified in the models for the underlying test results, interpretation of the coefficients in those models may not be straightforward for the ROC and VUS. The functional relationship described in Section 2 is highly nonlinear and does not lend itself to an intuitive explanation of those effects. The primary contribution of this paper is estimation of ROC and VUS for particular values of the covariates, as opposed to inference for how ROC and VUS change as a function of covariates. The problem is that covariates can have large effects on the test results but no effect on the ROC and VUS.

The paper is organized as follows. In Section 2, we first present a regression model for a three-category diagnostic test. We show that the regression model on test results induces the covariate-specific ROC surface. In Section 3, we propose estimation methods to fit the regression models under three cases. Depending on how strong assumptions we want to make for the model, we propose to use statistical methods which have been used in other applied fields. In Section 4, we conduct simulation studies to validate our estimation procedure. In Section 5, we study real examples of Alzheimer's disease. In Section 6, we conclude by suggesting future research directions.

2 Covariate-specific ROC surface model

Let T_c stand for the test results for an individual randomly selected from the population of the c -th class ($c = 1, 2, 3$). We consider the following transformation-location-scale regression model to describe the relationship between p -dimensional covariates $\mathbf{Z} = (Z_1, Z_2, \dots, Z_p)^T$ and the test results.

$$g(T_c) = \mathbf{Z}^T \boldsymbol{\beta}_c + \sigma_c \varepsilon, \quad (2.1)$$

where $g(\cdot)$ is a strictly increasing function; $\boldsymbol{\beta}_c = (\beta_{c1}, \beta_{c2}, \dots, \beta_{cp})^T$ is the p -dimensional regression coefficient specific to the c -th class; σ_c is a scale parameter specific to the c -th class; ε follows a distribution on $(-\infty, \infty)$ with a density function f common for all three classes. We note that in this model the parameters $\boldsymbol{\beta}_c$ describe the dependence of the test results on the covariates for a particular category of population. Their interpretation follows the usual interpretation for regression coefficients in regression analysis.

In model (2.1), the effects of covariates on the test results are different for three classes. The variation of the test results conditional on covariates also exhibits heterogeneity across the classes. We assume that the three classes are ordered such that $g(T_1)$, $g(T_2)$ and $g(T_3)$ tend to take high, intermediate and low values, respectively.

For two threshold values $c_1 > c_2$, the decision for classifying a subject into the c -th class ($c = 1, 2, 3$) is made based on whether $g(T)$ belongs to interval (c_1, ∞) , $(c_2, c_1]$ or $(-\infty, c_2]$. The correct classification rates (CCR) for the three classes can then be written as

$$\begin{aligned} \text{CCR}_1 &= P_{\mathbf{Z}}(g(T_1) > c_1) = 1 - F((c_1 - \mathbf{Z}^T \boldsymbol{\beta}_1)/\sigma_1), \\ \text{CCR}_2 &= P_{\mathbf{Z}}(g(T_2) \in (c_2, c_1]) = F((c_1 - \mathbf{Z}^T \boldsymbol{\beta}_2)/\sigma_2) - F((c_2 - \mathbf{Z}^T \boldsymbol{\beta}_2)/\sigma_2), \\ \text{CCR}_3 &= P_{\mathbf{Z}}(g(T_3) \leq c_2) = F((c_2 - \mathbf{Z}^T \boldsymbol{\beta}_3)/\sigma_3), \end{aligned}$$

where $F(t) = \int_{-\infty}^t f(x) dx$ is the cumulative distribution function of ε . Model (2.1) thus induces a regression model for the covariate effects on the three CCR's. Each CCR is the probability of correctly identifying one subject with covariate $\mathbf{Z}$ into his or her correct class for a fixed pair of thresholds (c_1, c_2) . If we move the threshold in the support of $g(T)$ and plot all the corresponding $(\text{CCR}_1, \text{CCR}_2, \text{CCR}_3)$ in the three-dimensional space, we then obtain the graph of an ROC surface (see Li and Zhou [18]).

Mathematically, we usually derive the ROC surface for the diagnostic test by expressing CCR_2 as a function of CCR_1 and CCR_3 in the following way:

$$\begin{aligned} \text{ROC}_{\mathbf{Z}}(u, v) &= F(\mathbf{Z}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)/\sigma_2 + F^{-1}(1-u)\sigma_1/\sigma_2) - F(\mathbf{Z}^T(\boldsymbol{\beta}_3 - \boldsymbol{\beta}_2)/\sigma_2 + F^{-1}(v)\sigma_3/\sigma_2) \\ &= F(\mathbf{Z}^T(\boldsymbol{\beta}_1^* + a_1 F^{-1}(1-u)) - F(\mathbf{Z}^T(\boldsymbol{\beta}_2^* + a_2 F^{-1}(v))), \end{aligned} \quad (2.2)$$

where

$$\boldsymbol{\beta}_1^* = (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)/\sigma_2, \quad a_1 = \sigma_1/\sigma_2, \quad \boldsymbol{\beta}_2^* = (\boldsymbol{\beta}_3 - \boldsymbol{\beta}_2)/\sigma_2, \quad a_2 = \sigma_3/\sigma_2,$$

and $(u, v) \in [0, 1] \times [0, 1]$ such that $\text{ROC}(u, v) \geq 0$. This ROC surface can be interpreted as the probability that the test correctly classifies a subject with covariate $\mathbf{Z}$ from Class II given that its CCRs for subjects with the same covariate $\mathbf{Z}$ but from Classes I and III are u and v (see Li and Fine [15]).

Model (2.2) provides a bridge to connect test accuracy with a set of explanatory variables. Covariate-specific ROC surfaces can be constructed based on such a model. We notice in (2.2) that we assume $\mathbf{Z}$ for the three classes to be identical. In principle, we may allow the covariates to be category-specific and define

$$\text{ROC}_{\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3}(u, v) = F((\mathbf{Z}_1^T \beta_1 - \mathbf{Z}_2^T \beta_2)/\sigma_2 + F^{-1}(1-u)\sigma_1/\sigma_2) - F((\mathbf{Z}_3^T \beta_3 - \mathbf{Z}_2^T \beta_2)/\sigma_2 + F^{-1}(v)\sigma_3/\sigma_2).$$

The ROC surface is then interpreted as the probability that the test correctly classifies a subject with covariate $\mathbf{Z}_2$ from Class II given that its CCRs for subjects with the covariate $\mathbf{Z}_1$ and $\mathbf{Z}_3$ from Classes I and III are u and v . For the sake of simplicity in the presentation, we only consider identical covariates for three classes in this paper. The methods for non-identical covariates can be easily extended.

Finally an overall summary measure of diagnostic accuracy from the graphical display of ROC surface can be obtained as the following covariate-specific volume under the ROC surface (VUS)

$$\text{VUS}_{\mathbf{Z}} = \int \int \text{ROC}_{\mathbf{Z}}(u, v) \, du \, dv, \quad (2.3)$$

which can be interpreted as the probability of correctly ranking three subjects with covariates $\mathbf{Z}$ each from one of the three categories (see Li and Fine [15]). A test with large VUS value is preferred in practice.

3 Estimation methodology

We consider estimation methods for covariates-adjusted CCR, ROC surface, and VUS for three scenarios in this section. It is difficult to make inference for covariate-adjusted diagnostic parameters without further assumptions for the model. In this paper, we consider the following three scenarios separately:

Case I. g and f are both completely known.

Case II. g is completely unknown and f is completely known.

Case III. g is known with a parametric form and f is completely unknown.

In practice, Case III might be the most flexible and general model since it places little restriction on the distribution form of diagnostic tests. As we notice from the previous section, the exact functional form of g is not important for accuracy studies since it is not used for the definition of the diagnostic accuracy measures. Therefore Case II, without estimating the unknown g , might be appealing when we only make a specific distribution assumption for f . Case I may be least realistic and is used only to facilitate understanding the more complicated cases. Under each scenario, we first consider estimating unknown elements involved in model (2.1). Subsequently we estimate diagnostic measures introduced in the previous section by replacing the unknown parameters or functions with their estimates. For the sake of completeness, we also present the estimation of g and f when they are unknown. The estimate of g might be convenient to study covariate dependence and help interpreting the results more clearly.

In the following, suppose we collect data from the three classes and denote the observations as $\{T_{ci}, \mathbf{Z}_{ci} : c = 1, 2, 3; i = 1, 2, \dots, n_c\}$. We assume observations are all independent.

Case I

In Case I, if $g(\cdot) = \log(\cdot)$ and f is a known parametric distribution, model (2.1) is the familiar accelerated failure time model (see Wei [34]). The logarithm transformation may be reasonable when test results are positive-valued, such as the lifespan or numerical counts. When test results can take any value in $\mathbb{R}$, it may also be convenient to use the identity function for g .

There are some popular choices for the distribution of ε :

$$\begin{aligned} \text{Normal distribution} \quad f(\varepsilon) &= (2\pi)^{-1/2} e^{-\varepsilon^2/2}, \\ \text{Logistic distribution} \quad f(\varepsilon) &= e^\varepsilon / (1 + e^\varepsilon)^2, \\ \text{Gumbel distribution} \quad f(\varepsilon) &= \exp(\varepsilon - e^\varepsilon). \end{aligned}$$

The three distributions are all well-known members of the location-scale family (see Shao [28]) and have wide applications in lifetime data analysis (Lawless (2003)).

In this parametric case, the only unknown parameters β_c and σ_c may be obtained from the standard maximum likelihood estimation (MLE). It is well known that MLE produces consistent and asymptotic normal parameter estimates $\hat{\beta}_c$ and $\hat{\sigma}_c$.

The estimates for the three CCRs, the ROC surface (2.2) and covariate-adjusted VUS (2.3) can be constructed by the plug-in methods where we replace all unknown parameters with their ML estimators. Since all estimated diagnostic measures are smooth functions of the ML estimators, it is easy to show their consistency and asymptotic normality with the delta method under this case. We are particularly interested in the covariate-adjusted VUS values since it is a crucial summary of the performance of the test. Denote

$\theta = (\beta_1^T, \beta_2^T, \dots, \beta_3^T, \sigma_1, \sigma_2, \dots, \sigma_3)^T$. We can show that the variance of a particular VUS is

$$\Psi^T \text{cov}(\hat{\theta}) \Psi, \quad (3.1)$$

where $\text{cov}(\hat{\theta})$ is the covariance matrix of the MLE $\hat{\theta}$ and $\Psi = \{\text{VUS}\} \{\hat{\theta}\}^{-1}$. The partial derivative Ψ may assume complicated forms with a double integral given in the Appendix. The integral has no closed analytic form even if the error distribution is normal. To use this method one has to use numerical integration techniques (see Pinheiro and Bates [26]).

To make a coherent comparison across all cases, we choose to use the bootstrap (see Efron and Tibshirani (1993), Li et al. [16], Li et al. [17]) to compute the sampling variances of the estimates of VUS in the following numerical studies. Statistical inference may follow then.

Case II

In Case II, the model is usually termed as the linear transformation model (see Kabfleisch [12], Dabrowska and Doksum [2]) in survival analysis. There is a rich literature in the study of censored lifetime under this kind of model. See Fine, Ying and Wei [5] and Fine [3] for the comprehensive review on this topic. We focus on one approach introduced by Fine and Bosch [4] in this section.

We notice that the estimation of ROC surface and VUS only require plugging in the estimators of the finite-dimensional model parameters $\hat{\beta}_c$ and $\hat{\sigma}_c$. The estimation of $\hat{g}$ is only useful when we want to study the exact distribution of the test and its dependence on covariates. Since we just make a comparison among the test results from three categories in

diagnostic accuracy studies, we do not require any knowledge of the form of this unknown function. Semiparametric estimation of linear transformation model without estimating the unknown transformation has been studied in Fine and Bosch [4]. We briefly sketch this method in the following.

Suppose the error distribution to be standard normal for the simplicity of presentation. Results for other parametric distribution can be similarly derived. We may write model (2.1) in an integrated form as

$$g(T_i) = \sum_{c=1}^3 \mathbf{Z}_i^T \boldsymbol{\beta}_c I\{C_i=c\} + \varepsilon, \quad \varepsilon_i \sim N(0, 1), \quad i=1, 2, \dots, n_1+n_2+n_3,$$

where C_i indicates the category for the i th subject. We may estimate the unknown parameters $\boldsymbol{\beta}_c$ by a quasi-likelihood-type estimating equation,

$$\sum_{i < j} w(\mathbf{Z}_{ij}^T \alpha) \mathbf{Z}_{ij} \{I(T_i > T_j) - \Phi(\mathbf{Z}_{ij}^T \alpha)\} = 0,$$

where Φ is the distribution function of the standard normal, $w = \psi\{\Phi(1 - \Phi)\}^{-1}$, $\psi = d\{\Phi(s)\}/ds$, $\mathbf{Z}_{ij} = ((\mathbf{Z}_i I\{C_i = 1\} - \mathbf{Z}_j I\{C_j = 1\})^T, (\mathbf{Z}_i I\{C_i = 2\} - \mathbf{Z}_j I\{C_j = 2\})^T, (\mathbf{Z}_i I\{C_i = 3\} - \mathbf{Z}_j I\{C_j = 3\})^T)^T$ and $\alpha = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \boldsymbol{\beta}_3^T)^T / \sqrt{2}$. Many familiar statistical packages with the function for probit regression can be adapted to solve the estimating equation and obtain the parameters $\boldsymbol{\beta}_c$. Fine and Bosch [4] have supplied inference details for this method. The estimated regression coefficients may then be used to construct ROC surfaces and VUS.

The variance for the estimated VUS may be evaluated by (3.1), where $\text{cov}(\hat{\theta})$ is the covariance of the estimates $\hat{\theta}$ resulted from Fine and Bosch's method and Ψ is the same partial derivative. For the same sake of computation ease and feasibility as in Case I, we suggest using a straightforward bootstrap resampling method instead to make inference (see Li et al. [16] and Li et al. [17]).

Case III

In Case III, the transformation function $g(\cdot)$ may take a parametric form, such as the well-known Box-Cox power transformation

$$g(t; \lambda) = \begin{cases} (t^\lambda - 1)/\lambda, & \lambda \neq 0; \\ \log(t), & \lambda = 0. \end{cases}$$

In general, λ is considered to be from a compact index set where the true parameter resides. Foster et al. [6] suggest a practically reasonable range $\{\lambda : |\lambda| \leq 2\}$. There seems to be no more practical appeal to consider a much wider range for the power parameter as the interpretation would become difficult. We note that besides Box-Cox transformation, there are also other possible parameterized transformation methods. They can be similarly worked out. We only restrict our attentions to Box-Cox transformation in this paper.

Under such a specification for $g(\cdot)$, when the distribution for ε is unknown, estimation methods of model (2.1) have been studied by Han [7], Newey [23], Robinson [27], Wang and Ruppert [33] and Foster et al. [6].

When the error density is completely unknown, it may be more flexible to estimate such a density (absorbing the nuisance σ_c) for each class separately. We thus slightly change the model form to be

$$g(T_c) = \mathbf{Z}^T \boldsymbol{\beta}_c + \varepsilon_c, \quad (3.2)$$

where $\varepsilon_c \sim f_c$. We consider the two-stage method introduced by Foster et al. [6] to estimate λ , $\boldsymbol{\beta}_c$ and f_c . In fact, to compute the accuracy measures such as the ROC surface and VUS, it suffices to have the distribution function F_c instead of the underlying density f_c which is usually more difficult to estimate.

Step 1

For a fixed λ , the regression coefficients are obtained by the least squared estimation as

$$\hat{\boldsymbol{\beta}}_c(\lambda) = \left(\sum_{i=1}^{n_c} \mathbf{Z}_{ci} \mathbf{Z}_{ci}^T \right)^{-1} \sum_{i=1}^{n_c} \mathbf{Z}_{ci} g(T_{ci}; \lambda).$$

Step 2

Evaluate the minimum distance statistics defined as

$$S(\lambda) = (n_1 + n_2 + n_3)^{-1} \sum_{c=1}^3 \sum_{i=1}^{n_c} \int_0^\infty [I(T_{ci} < t) - \hat{F}_c \{g(t; \lambda) - \mathbf{Z}_{ci}^T \hat{\boldsymbol{\beta}}_c; \hat{\boldsymbol{\beta}}_c, \lambda\}]^2 dW(t),$$

where

$$\hat{F}_c(t; \boldsymbol{\beta}_c, \lambda) = n_c^{-1} \sum_{i=1}^{n_c} I \{g(T_{ci}; \lambda) - \mathbf{Z}_{ci}^T \boldsymbol{\beta}_c < t\}$$

and $W(t)$ is a user-specified weight function. In theory $W(t)$ could any bounded, differentiable, positive, and strictly increasing distribution function. We adopt a normal distribution by following the recommendation in Foster et al. [6].

Step 3

Repeat Steps 1 and 2 for a grid of λ values within a compact interval Λ . Choose $\hat{\lambda}$ to be the value minimizing $S(\lambda)$.

Step 4

The estimator for $\boldsymbol{\beta}_c$ is obtained as $\hat{\boldsymbol{\beta}}_c(\hat{\lambda})$.

Large sample properties for similar estimators have been studied carefully by Foster et al. [6]. The technical difference to adapt our methods to Foster et al. [6] is just that we need to specify three distributions for the three different classes. The rest of the arguments are in the same manner. The estimators from the above procedure thus enjoy the consistency and asymptotic normality as described in Foster et al. [6]. Subsequently we estimate the ROC surface by

$$\widehat{\text{ROC}}_{\mathbf{Z}}(u, v) = \hat{F}_2(\mathbf{Z}^T(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2) + \hat{F}_1^{-1}(1 - u)) - \hat{F}_2(\mathbf{Z}^T(\hat{\boldsymbol{\beta}}_3 - \hat{\boldsymbol{\beta}}_2) + \hat{F}_3^{-1}(v)), \quad (3.3)$$

and VUS by integrating (3.3). We notice that $\hat{\lambda}$ is also not used in the final inference for diagnostic accuracy measures. For the variance estimates for $\widehat{VUS}$, we recommend to use the bootstrap methods introduced by Foster et al. [6] and Jin et al. [11]. Such re-sampling procedures have been proved theoretically and practically to yield accurate variance estimates and facilitate statistical inference (in Li et al. [16], Li et al. [17]).

4 Simulation study

We conduct a simulation experiment in this section to assess the performance of the proposed estimators. In fact the consistency and asymptotic distribution properties of the estimators involved in three cases have been well established in the literature. In this paper we are concerned with the performance of adapting such estimation methods in the inference for the diagnostic accuracy measures.

We generate two dimensional covariates $\mathbf{Z} = (Z_1, Z_2)$ from two-independent uniform distributions on $[0, 1]$. The regression coefficients for the three classes are $\beta_1 = (3, 2.5)^T$, $\beta_2 = (1.5, 2)^T$, and $\beta_3 = (1, 1)^T$. We fix $\sigma_c = 0.5$ and generate the model error ε from normal distributions. Two sample size scenarios were considered: (i) class sample size equal to 40 plus a binomial variate with size 40 and success probability 0.5; (ii) class sample size equal to 50 plus a binomial variate with size 20 and success probability 0.5. Case (i) tends to give more unequal sample proportions of three classes since it involves a larger variance than Case (ii). We choose $g(x) = x^3$ to generate the response T_c for the three classes according to (2.1).

For each generated data set, we use the estimation methods under three assumed cases. In Case I, we use the known transformation g and known distribution functions for ε . In Case II, we only have the known distribution of ε and need to estimate β_c semiparametrically. In Case III, we estimate λ for a the power transformation and also estimate the error distributions F . For all cases, we estimate the coefficients β_c for three classes and the overall accuracy measures VUS. In this paper we use the bootstrap method to resample from the data set with the bootstrap sample size $B = 800$ and compute the bootstrap standard errors for all the parameter estimates.

We report the estimated coefficients β over 1000 simulations in Tables 1 and 2. Eyeballing the results in the tables, we notice that all three methods can achieve consistent estimates of the regression parameters. Parameter estimates under Case I have the smallest bias while the bias for Case III is relatively larger. However, Case I is least realistic since we can know the true function g and error distribution f exactly only in a simulation setting. The simulation results thus indicate the estimation methods under Case III approximate closely the results under Case I. All methods appear to have smallest finite sample bias and variance when the error distribution is normal (see Table 1). The estimation results for σ_c and f_c are similarly satisfactory and not reported in the tables.

We also examine the performance of the estimated VUS parameters. We report VUS values for two covariates Z_1 and Z_2 at their three quartiles over their distributions. The true VUS values are computed by using a numerical integration for (3). All methods yield consistent estimation of the true covariate-specific VUS values. Such results affirm the applicability of our proposed statistical procedure for computing diagnostic accuracy measures.

5 Analysis of the Alzheimer's disease data

In this section, we consider applying our proposed statistical methods to a practical example in a study of Alzheimer's disease (AD). All the subjects in this example were 65 years or older enrolled at the 30 NIH funded Alzheimer's Research Centers before July 2002

throughout the U.S. who died and underwent autopsy. All had taken the Mini-Mental State Examination (MMSE) within 2 years before death. The definitive examination of AD was based on the extent of neuritic plaques and neurofibrillary tangles, the hallmarks of AD, at brain autopsy. The NIA/Reagan Institute criteria, based on the frequency of both plaques and tangles in the neocortex, classify patients into one of the three different diseases classes: Class I include subjects with a high likelihood of dementia being due to AD; Class II include subjects with intermediate likelihood of dementia being due to AD; and Class III include subjects with low likelihood of dementia being due to AD. Each patient underwent the test and a continuous test result was recorded thereafter. A total of 3728 results were included in this analysis where the sample size for the three individual classes are 2283, 850 and 595, respectively. See Zhou and Castelluccio [38] and Koepsell et al. [13] for more details about this data.

We were interested in determining the diagnostic accuracy of the continuous tests MMSE in terms of classifying patients into the three categories correctly. Previously, Li and Zhou [18] considered nonparametric and semiparametric estimation procedures to obtain an overall estimation of the diagnostic performance of the test. The authors therein reported the ROC surface and VUS for this data set. Their methods, however, did not incorporate the covariate factors in the data and were thus limited for practical use. In the sequel we used the estimation methods introduced in this paper to expand the previous findings with an adjustment of the covariates.

We thus considered model (2.1) for MMSE test where the covariates include the sex and age of the patients. We started with the methods described in Case III since we intended to make as few assumptions as possible on the model forms. We computed $S(\lambda)$ for a grid of λ values and plotted the results in Figure 1. In this case, the lowest score $S(\lambda)$ corresponds to $\lambda = -1/4$. Hence we chose $g(T, \lambda) = T^{-1/4}$ as the Box-Cox transformation to the original MMSE score. The estimated regression coefficients were reported in Table 3. Since the response was transformed, the interpretation of these coefficients might not be easy. We notice that the coefficients and standard errors were different for the three classes. Significant relationship between the transformed response and the covariates was found for Class II.

Using the methods in this paper, we then proceeded to compute ROC surfaces and VUS values for specific covariate values based on the fitted model. Figure 2 displays two estimated ROC surfaces for males and females with age 70. It appears that the volumes under the two surfaces are quite large, compared to the ROC surface for the overall population obtained in Li and Zhou [18]. Diagnostic accuracy for this particular group of subjects is thus relatively better than the population average. Such figures are helpful to select appropriate cut off values for the MMSE scores when the decision making is subject to desired level of correct classification probabilities for the three classes. When parametric likelihood assumptions are plausible, we might choose the estimation strategies in Case I or II and thus produce smooth ROC surfaces based on the assumed distribution functions. It is not difficult to modify our program to obtain such figures.

In Table 4 we report the estimated VUS values for a range of covariate values. For example, for female subjects who are 80 years old, the probability that they can be correctly classified into the three disease categories of AD is about 0.7993. Such a probability of correct classification among three disease categories for males at the same age is only 0.7332. We can empirically observe a pattern in Table 4. It seems that in general the MMSE test is more accurate for women than for men at any given age. Also the accuracy of the test increases as the subject becomes older. Such findings for the test accuracy of MMSE have never been reported in the literature. Since adjusting covariate effects such as age and sex effects is an

important issue in epidemiological studies, our results might potentially benefit the clinical diagnosis and treatment of AD and other similar problems.

Li and Zhou [18] have examined the estimation of VUS under parametric tri-normal assumption and found very similar estimates of VUS values to those obtained with nonparametric methods. This suggests that we may apply the methods in Case II for this particular data set. We thus re-analyze the data by using Fine and Bosch's methods for Case II. The ease of implementation of this method may also be appealing to practitioners. The ROC surfaces and VUS values resulted from this method are close to what we have obtained using the methods under Case III. We included the VUS results in Table 4 for a comparison.

6 Covariate-adjusted ROC measures

After one constructs covariate-specific ROC surfaces at different covariate values, it is sensible to develop a summary index which adjust all these covariate-dependent surfaces across the distribution of covariates.

Specifically, we may define a covariate-adjusted ROC surface as

$$\text{ROC}(u, v) = \int \text{ROC}_z(u, v) dF_Z(z),$$

where $\text{ROC}_z(u, v)$ is a surface for covariate z and F_Z is the distribution of the covariates in the population. The covariate-adjusted VUS may be similarly defined as a weighted average of the covariate-specific VUS across the distribution F_Z .

In practice, the distribution F_Z can be replaced by the empirical distribution for $\mathbf{Z}$ in the sample. The inference for covariate-adjusted ROC measures can be drawn easily by following a bootstrap resampling procedure. That is, we generate bootstrap samples from the original sample, evaluate the covariate-specific ROC measures for all distinct covariate values based on the particular bootstrap sample, evaluate the covariate-adjusted ROC measures then by averaging across the bootstrap sample, and finally join all the bootstrap estimates to yield the inference.

For the AD data analysis, we have computed the age-and-sex-adjusted VUS to be 0.76 (95% confidence interval [0.73, 0.79]) which is relatively larger than the unadjusted overall VUS 0.39 ([0.37, 0.41]) reported in Li and Zhou [18]. The adjusted VUS indicates that for three subjects with the same age and sex each from the three categories the MMSE test can correctly classify them simultaneously with a probability 0.76. The correct classification probability would be lower when the three subjects have different covariate values sampled from the general population. In practice the covariate-adjusted ROC measures may be more useful when the samples are obtained according to designed covariate values. Janes and Pepe [10] have provided helpful interpretation for the two-dimensional covariate-adjusted ROC curve.

7 Concluding remarks

We present a statistical method to address the estimation of covariate-adjusted ROC surfaces and the volumes under such surfaces in this article. The main idea is to relate the diagnostic accuracy of the test to the covariate values through the regression model (2.1). We explore various model assumptions and discuss appropriate statistical procedures. We follow the analytic methods in Case III mainly in this paper as it depends on less model assumptions. Even when investigators have strong belief about the test distribution and decide to consider

Cases I or II, it may still be worthwhile to produce the results following the methods for Case III to assess the model goodness-of-fit.

We stress that we are using regression models primarily to get an estimate of diagnostic measures such as ROC surface and VUS for particular covariate values. Sometimes it is unnecessary to estimate the entire set of model components, e.g. the unknown transformation function in Case II. The interpretation of the regression coefficients β_c may be linked to the association between ROC (or VUS) and covariates. However, such a link is indirectly through the test value. Theoretical development between VUS and covariates directly is needed for further research.

The interpretation of the results from the ROC-related regression analysis must be cautious. When evaluating conditional accuracy measures specific to certain fixed covariate value (such as fixed age or gender), the available information from the sample becomes limited and the resulting estimates are not as efficient as those based on the full sample. We realize that the regression procedure is essentially equivalent (though not literally carried out in this way) to divide the whole sample into small strata according to covariate values and evaluate the diagnostic accuracy within each stratum. The problem is even more severe for continuous covariates since they can only attain zero mass. Another potential weakness for covariate adjustment is the inference: when too many accuracy measures are computed with the same data the problem of multiple comparison would arise and any significance finding has to be re-considered after more sophisticated calculation. Introducing regression into accuracy analysis may not be an easily entertainable idea after all.

Acknowledgments

We are thankful to the data and scientific support provided by the National Alzheimer's Coordinating Center (NACC). This work was supported by National University of Singapore Academic Research Funding (Grant No. R-155-000-109-112), the National Medical Research Council in Singapore, NACC (Grant No. U01 AG16976), the National Institute of Health (Grant No. R01 EB005829) and National Natural Science Foundation of China (Grant No. 30728019).

Appendix

Expression of Ψ

We provide the expression for Ψ . In practice it may be implemented with the numerical integration method with a given distribution F .

$$\Psi = \begin{pmatrix} \mathbf{Z} \sigma_2^{-1} \int f(\mathbf{Z}^T(\beta_1 - \beta_2)/\sigma_2 + F^{-1}(1-u)\sigma_1/\sigma_2) du \\ \mathbf{Z} \sigma_2^{-1} \int \int f\left(\frac{\mathbf{Z}^T(\beta_3 - \beta_2)}{\sigma_2} + \frac{F^{-1}(v)\sigma_3}{\sigma_2}\right) - f\left(\frac{\mathbf{Z}^T(\beta_1 - \beta_2)}{\sigma_2} + \frac{F^{-1}(1-u)\sigma_1}{\sigma_2}\right) dudv \\ \mathbf{Z} \sigma_2^{-1} \int f(\mathbf{Z}^T(\beta_3 - \beta_2)/\sigma_2 + F^{-1}(v)\sigma_3/\sigma_2) dv \\ \sigma_2^{-1} \int f(\mathbf{Z}^T(\beta_1 - \beta_2)/\sigma_2 + F^{-1}(1-u)\sigma_1/\sigma_2) F^{-1}(1-u) du \\ \sigma_2^{-2} \int \int f\left(\frac{\mathbf{Z}^T(\beta_3 - \beta_2)}{\sigma_2} + \frac{F^{-1}(v)\sigma_3}{\sigma_2}\right) (\mathbf{Z}^T(\beta_3 - \beta_2) + F^{-1}(v)\sigma_3) \\ - f\left(\frac{\mathbf{Z}^T(\beta_1 - \beta_2)}{\sigma_2} + \frac{F^{-1}(1-u)\sigma_1}{\sigma_2}\right) (\mathbf{Z}^T(\beta_1 - \beta_2) + F^{-1}(1-u)\sigma_1) dudv \\ \sigma_2^{-1} \int f(\mathbf{Z}^T(\beta_3 - \beta_2)/\sigma_2 + F^{-1}(v)\sigma_3/\sigma_2) F^{-1}(v) dv \end{pmatrix}.$$

References

1. Birnbaum, ZW. On a use of the Mann-Whitney statistics. In: Neyman, J., editor. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability I; 1956. p. 13-17.

2. Dabrowska DM, Doksum KA. Partial likelihood in transformation models with censored data. *Scand J Statist.* 1988; 15:1–24.
3. Fine JP. Analysing competing risks data with transformation models. *JRSS Ser B.* 1999; 61:817–830.
4. Fine JP, Bosch J. Risk assessment via a robust probit model, with application to toxicology. *J Amer Statist Assoc.* 2000; 95:375–382.
5. Fine JP, Ying Z, Wei LJ. On the linear transformation model with censored data. *Biometrika.* 1998; 85:980–986.
6. Foster AM, Tian L, Wei LJ. Estimation for the Box-Cox transformation model without assuming parametric error distribution. *J Amer Statist Assoc.* 2001; 96:1097–1101.
7. Han AK. A non-parametric analysis of transformations. *J Econom.* 1987; 35:191–209.
8. Heckerling PS. Parametric three-way receiver operating characteristic surface analysis using Mathematica. *Medical Decision Making.* 2001; 20:409–417. [PubMed: 11575490]
9. Hoeffding W. A class of statistics with asymptotically normal distribution. *J Math Statist.* 1948; 19:293–325.
10. Janes H, Pepe MS. Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. *Biometrika.* 2009; 96:371–382. [PubMed: 22822245]
11. Jin Z, Ying Z, Wei LJ. A simple resampling method by perturbing the minimand. *Biometrika.* 2001; 90:341–353.
12. Kalbfleisch JD. Likelihood methods and nonparametric tests. *J Amer Statist Assoc.* 1978; 73:167–170.
13. Koepsell TD, Chi YY, Zhou XH, et al. An alternative method for estimating efficacy of the AN1792 Vaccine for Alzheimer's disease. *Neurology.* 2007; 69:1868–1872. [PubMed: 17984455]
14. Kruskal WH, Wallis WA. The use of ranks in one-criterion variance analysis. *J Amer Statist Assoc.* 1952; 47:583–621.
15. Li J, Fine JP. ROC analysis for multiple classes and multiple categories and its application in microarray study. *Biostatistics.* 2008; 9:566–576. [PubMed: 18304996]
16. Li J, Tai BC, Nott DJ. Confidence interval for the bootstrap P-value and sample size calculation of the bootstrap test. *J Nonparametr Stat.* 2009; 21:649–661.
17. Li J, Zhang CM, Doksum KA, et al. Simultaneous confidence intervals for semiparametric logistic regression and confidence regions for the multi-dimensional effective dose. *Statistica Sinica.* 2010; 20:637–659.
18. Li J, Zhou XH. Nonparametric and semiparametric estimation of the three way receiver operating characteristic surface. *J Statist Plann Inference.* 2009; 139:4133–4142.
19. Lindstrom MJ, Bates DM. Nonlinear mixed-effects models for repeated measures data. *Biometric.* 1990; 46:673–687.
20. Mossman D. Three-way ROCs. *Medical Decision Making.* 1999; 19:78–89. [PubMed: 9917023]
21. Mukarjee H, Stern S. Feasible nonparametric estimation of multiargument monotone functions. *J Amer Statist Assoc.* 1994; 89:77–80.
22. Nakas CT, Yiannoutsos CT. Ordered multiple-class ROC analysis with continuous measurements. *Statistics in Medicine.* 2004; 23:3437–3449. [PubMed: 15505886]
23. Newey WK. Efficient instrumental variables estimation of nonlinear models. *Econometrica.* 1990; 58:809–837.
24. Pepe, MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford: Oxford University Express; 2003.
25. Pinheiro JC, Bates DM. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *J Comput Graph Statist.* 1995; 4:12–35.
26. Pinheiro, JC.; Bates, DM. *Mixed-Effects Models in S and S-PLUS.* New York: Springer;
27. Robinson PM. Best nonlinear three-stage least squares estimation of certain econometric models. *Econometrica.* 1991; 59:755–786.
28. Shao, J. *Mathematical Statistics.* New York: Springer; 1999.

29. Spizzichino, F. Subjective Probability Models for Lifetimes. Chapman Hall/CRC; 2001.
30. Stram DO, Lee JW. Variance components testing in the longitudinal mixed-effects models. *Biometrics*. 1994; 50:1171–1177. [PubMed: 7786999]
31. van der Vaart, AW. Asymptotic Statistics. Cambridge: Cambridge press; 1999.
32. Vonesh EF, Carter RL. Mixed-effects nonlinear regression for unbalanced repeated measures. *Biometrics*. 1992; 48:1–18. [PubMed: 1581479]
33. Wang N, Ruppert D. Nonparametric estimation of the transformation in the transform-both-sides regression models. *J Amer Statist Assoc*. 1995; 90:731–738.
34. Wei LJ. The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*. 1992; 11:1871–1879. [PubMed: 1480879]
35. Xiong C, van Belle G, Miller JP, et al. Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups. *Statistics in Medicine*. 2006; 25:1251–1273. [PubMed: 16345029]
36. Xiong C, van Belle G, Miller JP, et al. A parametric comparison of diagnostic accuracy with three ordinal diagnostic groups. *Biometrical Journal*. 2007; 49:682–693. [PubMed: 17763377]
37. Zhang Y, Li J. Combining multiple markers for multi-category classification: an ROC surface approach. *Australian and New Zealand Journal of Statistics*. 2011; 53:63–78.
38. Zhou XH, Castelluccio P. Adjusting for non-ignorable verification bias in clinical studies for Alzheimer's disease. *Statistics in Medicine*. 2004; 23:221–230. [PubMed: 14716724]
39. Zhou, XH.; Obuchowski, NA.; McClish, DK. *Statistical Methods in Diagnostic Medicine*. Wiley; New York: 2002.

Figure 1.
 $S(\lambda)$ for different λ

Figure 2. ROC surfaces for men (a) and women (b) at age 70. The three coordinates correspond to the correct classification probabilities for the three classes

Table 1

Simulation results for data generated from normal distribution with class sample size 40+Binomial(40, 0.5). Estimates are the average of estimated parameters in 1000 simulations. SE are the standard error of the estimated parameters in 1000 simulations. Bootstrap SE are the average of the estimated standard error based on the bootstrap methods

Parameters	Case	Estimates	SE	Bootstrap SE
$\beta_1 = \begin{pmatrix} 3 \\ 2.5 \end{pmatrix}$	I	$\begin{pmatrix} 2.89 \\ 2.46 \end{pmatrix}$	$\begin{pmatrix} 0.26 \\ 0.25 \end{pmatrix}$	$\begin{pmatrix} 0.24 \\ 0.24 \end{pmatrix}$
	II	$\begin{pmatrix} 2.80 \\ 2.45 \end{pmatrix}$	$\begin{pmatrix} 0.33 \\ 0.29 \end{pmatrix}$	$\begin{pmatrix} 0.35 \\ 0.35 \end{pmatrix}$
	III	$\begin{pmatrix} 3.13 \\ 2.64 \end{pmatrix}$	$\begin{pmatrix} 0.47 \\ 0.49 \end{pmatrix}$	$\begin{pmatrix} 0.44 \\ 0.49 \end{pmatrix}$
$\beta_2 = \begin{pmatrix} 1.5 \\ 2 \end{pmatrix}$	I	$\begin{pmatrix} 1.44 \\ 2.07 \end{pmatrix}$	$\begin{pmatrix} 0.29 \\ 0.26 \end{pmatrix}$	$\begin{pmatrix} 0.27 \\ 0.24 \end{pmatrix}$
	II	$\begin{pmatrix} 1.40 \\ 2.09 \end{pmatrix}$	$\begin{pmatrix} 0.36 \\ 0.32 \end{pmatrix}$	$\begin{pmatrix} 0.37 \\ 0.34 \end{pmatrix}$
	III	$\begin{pmatrix} 1.57 \\ 2.10 \end{pmatrix}$	$\begin{pmatrix} 0.39 \\ 0.38 \end{pmatrix}$	$\begin{pmatrix} 0.41 \\ 0.36 \end{pmatrix}$
$\beta_3 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$	I	$\begin{pmatrix} 1.04 \\ 0.97 \end{pmatrix}$	$\begin{pmatrix} 0.28 \\ 0.27 \end{pmatrix}$	$\begin{pmatrix} 0.26 \\ 0.24 \end{pmatrix}$
	II	$\begin{pmatrix} 1.07 \\ 1.05 \end{pmatrix}$	$\begin{pmatrix} 0.49 \\ 0.48 \end{pmatrix}$	$\begin{pmatrix} 0.46 \\ 0.48 \end{pmatrix}$
	III	$\begin{pmatrix} 1.10 \\ 1.05 \end{pmatrix}$	$\begin{pmatrix} 0.45 \\ 0.44 \end{pmatrix}$	$\begin{pmatrix} 0.48 \\ 0.47 \end{pmatrix}$
VUS($Z_1 = 0.25, Z_2 = 0.25$) = 0.48	I	0.48	0.065	0.064
	II	0.49	0.069	0.073
	III	0.49	0.072	0.069
VUS($Z_1 = 0.50, Z_2 = 0.50$) = 0.77	I	0.76	0.058	0.059
	II	0.78	0.057	0.059
	III	0.78	0.059	0.064
VUS($Z_1 = 0.75, Z_2 = 0.75$) = 0.92	I	0.92	0.037	0.039
	II	0.92	0.039	0.044
	III	0.93	0.038	0.043

Table 2

Simulation results for data generated from normal distribution with class sample size 50+Binomial(20, 0.5). Estimates are the average of estimated parameters in 1000 simulations. SE are the standard error of the estimated parameters in 1000 simulations. Bootstrap SE are the average of the estimated standard error based on the bootstrap methods

Parameters	Case	Estimates	SE	Bootstrap SE
$\beta_1 = \begin{pmatrix} 3 \\ 2.5 \end{pmatrix}$	I	$\begin{pmatrix} 2.93 \\ 2.42 \end{pmatrix}$	$\begin{pmatrix} 0.25 \\ 0.28 \end{pmatrix}$	$\begin{pmatrix} 0.27 \\ 0.25 \end{pmatrix}$
	II	$\begin{pmatrix} 2.86 \\ 2.39 \end{pmatrix}$	$\begin{pmatrix} 0.34 \\ 0.30 \end{pmatrix}$	$\begin{pmatrix} 0.33 \\ 0.36 \end{pmatrix}$
	III	$\begin{pmatrix} 2.90 \\ 2.40 \end{pmatrix}$	$\begin{pmatrix} 0.44 \\ 0.46 \end{pmatrix}$	$\begin{pmatrix} 0.44 \\ 0.48 \end{pmatrix}$
$\beta_2 = \begin{pmatrix} 1.5 \\ 2 \end{pmatrix}$	I	$\begin{pmatrix} 1.52 \\ 2.03 \end{pmatrix}$	$\begin{pmatrix} 0.28 \\ 0.25 \end{pmatrix}$	$\begin{pmatrix} 0.25 \\ 0.22 \end{pmatrix}$
	II	$\begin{pmatrix} 1.56 \\ 2.05 \end{pmatrix}$	$\begin{pmatrix} 0.32 \\ 0.30 \end{pmatrix}$	$\begin{pmatrix} 0.30 \\ 0.31 \end{pmatrix}$
	III	$\begin{pmatrix} 1.55 \\ 2.06 \end{pmatrix}$	$\begin{pmatrix} 0.35 \\ 0.34 \end{pmatrix}$	$\begin{pmatrix} 0.33 \\ 0.35 \end{pmatrix}$
$\beta_3 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$	I	$\begin{pmatrix} 0.97 \\ 0.98 \end{pmatrix}$	$\begin{pmatrix} 0.26 \\ 0.25 \end{pmatrix}$	$\begin{pmatrix} 0.28 \\ 0.26 \end{pmatrix}$
	II	$\begin{pmatrix} 0.92 \\ 0.94 \end{pmatrix}$	$\begin{pmatrix} 0.38 \\ 0.35 \end{pmatrix}$	$\begin{pmatrix} 0.31 \\ 0.37 \end{pmatrix}$
	III	$\begin{pmatrix} 0.88 \\ 0.89 \end{pmatrix}$	$\begin{pmatrix} 0.43 \\ 0.37 \end{pmatrix}$	$\begin{pmatrix} 0.40 \\ 0.35 \end{pmatrix}$
VUS($Z_1 = 0.25, Z_2 = 0.25$) = 0.48	I	0.47	0.060	0.062
	II	0.49	0.053	0.057
	III	0.49	0.059	0.053
VUS($Z_1 = 0.50, Z_2 = 0.50$) = 0.77	I	0.77	0.056	0.059
	II	0.78	0.061	0.062
	III	0.78	0.067	0.063
VUS($Z_1 = 0.75, Z_2 = 0.75$) = 0.92	I	0.92	0.034	0.037
	II	0.91	0.034	0.036
	III	0.90	0.038	0.035

Table 3Fitted regression coefficients for $Y^{-1/4}$

Class	Covariates	Coefficients	SE	P-value
I	Intercept	0.4472	0.0082	< 0.0001
	Age	0.0001	0.0001	0.3210
	Sex	0.0006	0.0015	0.6850
II	Intercept	0.4293	0.0040	< 0.0001
	Age	0.0001	0.0001	0.0469
	Sex	0.0016	0.0007	0.0460
III	Intercept	0.4297	0.0039	< 0.0001
	Age	0.0001	0.0001	0.6300
	Sex	-0.0007	0.0005	0.1540

Table 4

Covariate-specific VUS and bootstrap confidence intervals (in the brackets) for different age and sex, using methods constructed under Cases III and II

Age	Sex	
	Male	Female
Case III		
60	0.6795 [0.6011, 0.7557]	0.7569 [0.6799, 0.8353]
70	0.7010 [0.6226, 0.7794]	0.7831 [0.7047, 0.8615]
80	0.7332 [0.6744, 0.7920]	0.7993 [0.7405, 0.8581]
90	0.7729 [0.7141, 0.8317]	0.8122 [0.7534, 0.8710]
Case II		
60	0.6985 [0.6123, 0.7632]	0.7498 [0.6631, 0.8239]
70	0.7144 [0.6158, 0.7625]	0.7900 [0.7155, 0.8726]
80	0.7395 [0.6780, 0.8041]	0.7867 [0.7359, 0.8402]
90	0.7697 [0.7015, 0.8273]	0.8108 [0.7498, 0.8679]