Computational Statistics and Data Analysis 68 (2013) 326–338

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Estimating confidence intervals for the difference in diagnostic accuracy with three ordinal diagnostic categories without a gold standard

Le Kang [a], Chengjie Xiong [b], Lili Tian [c,*]

[a] Center for Devices and Radiological Health, US Food and Drug Administration, Silver Spring, MD 20993, United States
[b] Division of Biostatistics, Washington University in St. Louis, St. Louis, MO 63110, United States
[c] Department of Biostatistics, University at Buffalo, Buffalo, NY 14214, United States

ARTICLE INFO

Article history:
Received 23 September 2012
Received in revised form 3 July 2013
Accepted 4 July 2013
Available online 15 July 2013

Keywords:
EM algorithm
Generalized pivot
Gold standard
Parametric bootstrap
Volume under the ROC surface

ABSTRACT

With three ordinal diagnostic categories, the most commonly used measures for the overall diagnostic accuracy are the volume under the ROC surface (VUS) and partial volume under the ROC surface (PVUS), which are the extensions of the area under the ROC curve (AUC) and partial area under the ROC curve (PAUC), respectively. A gold standard (GS) test on the true disease status is required to estimate the VUS and PVUS. However, oftentimes it may be difficult, inappropriate, or impossible to have a GS because of misclassification error, risk to the subjects or ethical concerns. Therefore, in many medical research studies, the true disease status may remain unobservable. Under the normality assumption, a maximum likelihood (ML) based approach using the expectation–maximization (EM) algorithm for parameter estimation is proposed. Three methods using the concepts of generalized pivot and parametric/nonparametric bootstrap for confidence interval estimation of the difference in paired VUSs and PVUSs without a GS are compared. The coverage probabilities of the investigated approaches are numerically studied. The proposed approaches are then 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 compare the overall diagnostic accuracy of early stage AD between two different pairs of neuropsychological tests.

1. Introduction

Diagnostic testing is an extremely important aspect of medical care. Medical diagnosis involves the classification of patients into two or more categories. These categories may imply the presence or absence of a particular medical condition. Signs, symptoms or clinical tests are used to determine the classification. The evaluation of a diagnostic test procedure involves the estimation of parameters that describe the accuracy of diagnostic test relative to true classification and it is of paramount importance to compare the accuracies of diagnostic tests to decide on the best test for certain disease. For instance, one of the common indices used for overall diagnostic accuracy on the case when subjects are categorized in a binary fashion, i.e., non-diseased and diseased, is AUC (Zhou et al., 2002; Pepe, 2003; Shapiro, 1999). The comparison of the overall diagnostic accuracy between two diagnostic tests is frequently addressed by measuring the difference in the paired AUCs.

In practice, the diagnostic decision is not limited to a binary choice in many situations. For example, a clinical assessment, NPZ-8, of the presence of HIV-related cognitive dysfunction (AIDS Dementia Complex—ADC) would discriminate between

* Correspondence to: Department of Biostatistics, University at Buffalo, 706 Kimball Tower, 3435 Main Street, Buffalo, NY 14214-3000, United States. Tel.: +1 716 829 2715; fax: +1 716 829 2200.
E-mail address: ltian@buffalo.edu (L. Tian).

0167-9473/$ – see front matter
http://dx.doi.org/10.1016/j.csda.2013.07.007

patients exhibiting clinical symptoms of ADC (combined stages 1–3), subjects exhibiting minor neurological symptoms (ADC stage 0.5) and neurologically unimpaired individuals (ADC stage 0) (Nakas and Yiannoutsos, 2004). Another example provided by Xiong et al. (2006) concerns mild cognitive impairment (MCI) or early stage Alzheimer's disease (AD) being a transitional stage between the cognitive changes from normal aging and the more severe problems caused by the AD. Thereafter, we refer to the disease status between "non-diseased" and "diseased" as "intermediate", in other words, transitional status.

Given that an independent gold standard (GS) test on the disease status is available, Scurfield (1996) and Xiong et al. (2006) extended binary statistical tools such as the ROC curve and AUC and developed the volume and partial volume under the ROC surface (VUS and PVUS) to summarize the diagnostic accuracy with three ordinal diagnostic categories. Furthermore, Nakas and Yiannoutsos (2004) proposed a nonparametric estimation of a single VUS; Xiong et al. (2007) proposed a large sample approach for comparing several VUSs for normally distributed data. Most recently, Tian et al. (2011) addressed exact confidence interval estimation for the difference in paired VUSs and PVUSs based on the concepts of generalized pivot and showed that their approach generally can provide confidence intervals with reasonable coverage probabilities even at small sample sizes.

Notice that all the aforementioned methods assume the existence of a GS test. In other words, the true disease category is known. For instance, in the diagnosis of early stage AD (Xiong et al., 2006, 2007), the dementia severity of Alzheimer type was staged by the Clinical Dementia Rating (CDR) according to published rules (Morris, 1993), which is considered as a "GS" for evaluating different neuropsychological tests and biomarkers for early stage AD. The resulting diagnosis by clinical assessments such as CDR, although expected to be quite accurate, presumably was not totally free of misclassification errors and thus was not perfect. Such misclassifications are known to produce bias in estimating the diagnostic accuracy of disease markers, e.g., VUS. Further, such bias may prove to be detrimental when it comes to compare the diagnostic accuracy of multiple disease markers. It is therefore important to develop valid statistical methods of diagnostic comparison that do not rely on the existence of a perfect GS.

Some works involving estimating diagnostic accuracy without a GS have been done for binary diagnostic tests. For example, Henkelman et al. (1990) considered the estimation of ROC curves of continuous-scale tests in the absence of a GS test; Beiden et al. (2000) proposed maximum likelihood (ML) estimates of the ROC curves using the EM algorithm; Hsieh et al. (2009) proposed a ML based procedure for construction of confidence intervals for the difference in paired AUCs without a GS; Zhou et al. (2005) also developed a nonparametric ML method for estimating ROC curves in the absence of a GS test.

In this paper, we will focus on interval estimation for the difference in paired VUSs and PVUSs with three ordinal diagnostic categories without a GS by proposing a ML based approach using the EM algorithm in conjunction with the generalized variable approach as well as the parametric/nonparametric bootstrap methods. This paper is organized as follows. We first introduce some preliminaries about VUS and PVUS in Section 2. In Section 3, we will present the proposed methods. The performance of the proposed approaches including their robustness will be assessed by a numerical study in Section 4. In Section 5, our proposed methods will be applied to a real world study of very early stage AD diagnosis. We close with a broader discussion for evaluating diagnostic tests without a GS.

2. Preliminaries

The ROC surface, analogous to the ROC curve, has been proposed to assess the accuracy of tests with three ordinal diagnostic categories. Let Y_1, Y_2 and Y_3 denote the scores resulting from a diagnostic test 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 continuous scale and higher values indicate greater severity of the disease. Given a pair of threshold values c_1 and $c_3 (c_1 < c_3)$, let $\delta_1 = F_1(c_1)$, $\delta_3 = 1 - F_3(c_3)$ be the true classification rates for non-diseased and diseased categories, respectively. Then the probability that a randomly selected subject from an intermediate category has a score between c_1 and c_3 is

$$\delta_2 = F_2(c_3) - F_2(c_1) = F_2\left[F_3^{-1}(1-\delta_3)\right] - F_2\left[F_1^{-1}(\delta_1)\right]. \tag{1}$$

The triplet $(\delta_1, \delta_2, \delta_3)$, where $\delta_2 = \delta_2(\delta_1, \delta_3)$ is a function of (δ_1, δ_3), would produce an ROC surface in the three-dimensional space for all possible $(c_1, c_3) \in \mathbb{R}^2$. As the ROC curve for a binary diagnosis represents the trade-off between sensitivity and specificity, which are correct classification probabilities for the two categories (non-diseased and diseased), the ROC surface represents the three-way trade-off among the correct classification probabilities for the three categories.

In order to summarize the overall diagnostic accuracy for the diagnostic test, the volume under the ROC surface (VUS) has been considered. It is defined as

$$\mathrm{VUS} = \int_0^1 \int_0^{1-F_3\left[F_1^{-1}(\delta_1)\right]} F_2\left[F_3^{-1}(1-\delta_3)\right] - F_2\left[F_1^{-1}(\delta_1)\right] d\delta_3 d\delta_1. \tag{2}$$

This is a generalization of the AUC for a ROC curve under a binary classification. One could show that VUS is mathematically equivalent to the probability $P(Y_1 < Y_2 < Y_3)$ when Y_1, Y_2 and Y_3 are randomly selected from each diagnostic category, respectively. For a useless test (e.g., when Y_1, Y_2 and Y_3 have identical distributions), VUS is 1/6. Similar to the PAUC of a ROC curve in which investigators are only interested in a certain lower range of false positive rate, the partial volume under the