Published in final edited form as:
J Biopharm Stat. 2015 ; 25(6): 1215–1233. doi:10.1080/10543406.2014.971173.

CONFIDENCE INTERVAL ESTIMATION FOR SENSITIVITY TO THE EARLY DISEASED STAGE BASED ON EMPIRICAL LIKELIHOOD

Tuochuan Dong[1] and **Lili Tian**[1,*]

[1]Department of Biostatistics, University at Buffalo, Buffalo, NY 14214, USA

Abstract

Many disease processes can be divided into three stages: i.e. the non-diseased stage, the early diseased stage and the fully diseased stage. To assess the accuracy of diagnostic tests for such diseases, various summary indexes have been proposed, such as volume under the surface (VUS), partial volume under the surface (PVUS), and the sensitivity to the early diseased stage given specificity and the sensitivity to the fully diseased stage (P_2). This paper focuses on confidence interval estimation for P_2 based on empirical likelihood. Simulation studies are carried out to assess the performance of the new methods compared to the existing parametric and non-parametric ones. A real data set from Alzheimer's Disease Neuroimaging Initiative (ANDI)[2] is analyzed. Key Words: Empirical Likelihood; Diagnostic tests; The sensitivity to the early diseased stage.

Keywords

Empirical Likelihood; Diagnostic tests; The sensitivity to the early diseased stage

1. INTRODUCTION

Disease process is usually divided into two stages: the non-diseased and the diseased, and diagnostic tests are utilized to classify the subjects into different stages. The probability that a non-diseased subject is correctly classified is defined as the specificity, and the probability that a diseased subject is correctly identified is called sensitivity. When the outcome of diagnostic test is continuous, both sensitivity and specificity are functions of the cut-off value. As the cut-off value changes, sensitivity and specificity vary inversely to each other. The Receiver Operating Characteristic (ROC) curve, a plot of sensitivity versus (1-specificity) as the cut-off value runs through the whole range of all possible outcome values, is a popular graphical assessment of the diagnostic accuracy for a diagnostic test. For detailed review of statistical methods in ROC analysis, please see Shapiro (1999), Zhou et al. (2002), Pepe (2003) and Zou et al. (2010).

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

[1*]Correspondence to: Lili Tian, Department of Biostatistics, 717 Kimball Tower, 3435 Main St. Bldg. 26 Buffalo, NY 14214-3000 U.S.A. ltian@bu_alo.edu.

To assess the diagnostic accuracy of a binary-scale test, there exist many diagnostic accuracy measures such as the area under the curve (AUC). The AUC indicates the overall performance of a diagnostic test for all the cut-off values. However, in medical practice, a cut-off value is often chosen by medical practitioners so that a fixed value of specificity is achieved (typically 80, 90, or 95 per cent). Hence, the sensitivity given the specificity serves as a meaningful diagnostic measure. Towards this end, several papers discussed the issues of estimation of sensitivity given specificity. For example, Greenhouse and Mantel (1950) presented the inference procedures for a diagnostic test with continuous range, either with or without normal distribution assumptions; McNeil and Hanley (1984) estimated the point-wise confidence interval for sensitivity at a fixed specificity in the bi-normal model; Linnet (1987) took into account the sampling variation of the discrimination limits and proposed both parametric and non-parametric methods to construct the confidence interval; Platt et al. (2000) recommended a confidence interval by using Efron's bias-corrected acceleration (BCa) bootstrap; and Zhou and Qin (2005) introduced two non-parametric confidence intervals. Most recently, Qin et al. (2011) presented empirical likelihood-based confidence intervals for the sensitivity at a fixed level of specificity.

In practice, a disease process might involve three ordinal diagnostic stages: the normal healthy stage without even the earliest subtle disease symptoms, the early stage of the disease, and the stage of full-blown development of the disease. 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. Recently, the traditional ROC analysis has already been extended to three-stage cases, see e.g., Mossman (1999), Dreiseitl et al. (2000), Heckerling (2001), Nakas and Yiannoutsos (2004), Xiong et al. (2006), He and Frey (2008), Li and Zhou (2009), Nakas et al. (2010), Tian et al. (2010), He et al. (2010), Dong et al. (2011) and Li et al. (2012). For diseases such as AD, early detection is critical since it often means optimal time window for therapeutic treatment due to the fact that no pharmaceutical treatments to-date are effective for the late stage AD. However, it is far more challenging to diagnose subjects at the earliest disease stage for clinicians because of the subtle clinical symptoms in the early stage of many complex disease processes. Hence, the probability associated with the detection of early diseased stage is critical in medical science and serves as a very important diagnostic accuracy measure for diseases with three ordinal stages.

To be more specific, let Y_1, Y_2 and Y_3 denote the test results for the non-diseased, the early diseased, and the fully diseased group of a diagnostic test respectively, F_1, F_2 and F_3 denote corresponding cumulative distribution functions, and n_1, n_2 and n_3 denote sample sizes. Assume that the test results are measured on a continuous scale and that higher values indicate greater severity of the disease. Given a pair of threshold values c_1 and c_2 ($c_1 < c_2$), the subject is identified as non-diseased if the test result is smaller than c_1, as fully diseased if the test result is larger than c_2, and as early diseased if the test result is between c_1 and c_2. The specificity P_1, which is the correct classification rate for the non-diseased stage, sensitivity to the early diseased stage P_2, and the sensitivity to the fully diseased stage P_3 are defined as

$$\begin{aligned}
&P_1=F_1(c_1)\\
&P_2=F_2(c_2)-F_2(c_1)=F_2[F_3^{-1}(1-P_3)]-F_2[F_1^{-1}(P_1)]\\
&P_3=1-F_3(c_2).
\end{aligned} \tag{1}$$

Given P_1 and P_3, c_1 and c_2 can be determined. Consequently, P_2, the sensitivity to the early diseased stage given the specificity P_1 and the sensitivity to the fully diseased stage P_3, can be formulated as a function of P_1 and P_3, i.e. $P_2 = P_2(P_1, P_3)$ which also defines a surface in the three-dimensional space (P_1, P_3, P_2), namely, the ROC surface. The point (P_1, P_3, P_2) = (1, 1, 1) indicates the perfect discrimination ability.

To evaluate the diagnostic accuracy of the biomarkers for three-class diseases, various summary measures of the ROC surface have been proposed. Among them, the volume under the ROC surface (VUS), considered as the extension of AUC in the three-class disease paradigm, is a very popular one. The VUS denotes the probability that a randomly chosen subject from the non-diseased group, that from early diseased group and that from fully diseased group follow simple order, i.e., $VUS = P(Y_1 < Y_2 < Y_3)$. More details about VUS can be found in Nakas and Yiannoutsos (2004), Xiong et al. (2006), He and Frey (2008), Wan (2012) and Kang and Tian (2013).

In addition to the overall performance of a biomarker measured by VUS, an accurate estimate of P_2 helps clinicians to identify the best disease markers for early diagnosis and therefore the inference procedures for P_2 are very useful. Dong et al. (2011) first attempted to provide parametric and non-parametric confidence interval estimation methods for P_2. However, the most recommended methods depend on either normality assumption or Box-Cox transformation to normality. It is well known that not all of the non-normal distributions can be transformed to normal via Box-Cox transformation. Therefore, some alternative approaches for estimating the confidence interval of P_2 which do not depend on distributional assumption and also provide good coverage probabilities are worth exploring.

The goal of this paper is to present empirical likelihood-based confidence intervals for P_2, i.e. the sensitivity to the early diseased stage given specificity and the sensitivity to the fully diseased stage. Empirical likelihood is introduced by Owen (1990, 2001) and has many advantages over normal approximation-based methods. For instance, empirical likelihood-based confidence regions are range preserving and transformation respecting, the regularity conditions for empirical likelihood-based methods are weak and natural, and it utilizes the power of likelihood-based approaches to solve complex statistical problems. The empirical likelihood has been used widely in many applied areas including diagnostic tests with binary outcomes, e.g., Claeskens et al. (2003) suggested a smoothed empirical likelihood-based method (SEL) to estimate the sensitivity, and Qin et al. (2011) proposed two empirical likelihood-based confidence intervals for the sensitivity at a fixed level of specificity. The rest of this paper is organized as follows. Section 2 presents a review of existing methods. In Section 3, the large sample properties of P_2 and the empirical likelihood approaches are proposed. In Section 4, simulation studies are conducted to evaluate the proposed methods. In Section 5, a real data set from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database is analyzed. Section 6 is the discussion. The proofs for the formula of the variance for an estimator of P_2 and the empirical likelihood theorem are given in the Appendix.

2. EXISTING METHODS

This section presents a brief review of the existing methods including the generalized inference method and bootstrap approaches for confidence interval estimation of sensitivity to the early diseased stage by Dong et al. (2011).

2.1. Generalized Inference Method

Assume Y_i follows normal distributions with mean μ_i and variance σ_i^2 for $i = 1, 2, 3$, the generalized pivotal quantity for P_2 as given in (1) can be written as

$$R_{P_2}=\Phi\left[\frac{R_{\mu_3}-R_{\mu_2}+\Phi^{-1}(1-P_3)R_{\sigma_3}}{R_{\sigma_2}}\right]-\Phi[\frac{R_{\mu_1}-R_{\mu_2}+\Phi^{-1}(P_1)R_{\sigma_1}}{R_{\sigma_2}}$$

where $R_{\mu_i}=\overline{y}_i-Z_i\sqrt{R_{\sigma_i^2}/n_i}$, $Z_i \sim N(0, 1)$ and $R_{\sigma_i}=\sqrt{\frac{(n_i-1)s_i^2}{V_i}}$ where $V_i{\sim}\chi_{n_i-1}^2$ for $i = 1, 2, 3$. By generating V_i and Z_i repeatedly, an array of R_{P_2}'s can be obtained. A two-sided $100(1 - \alpha)$% generalized inference confidence interval for P_2, **GI**, is ($R_{P_2}(\alpha/2)$, $R_{P_2}(1 - \alpha/2)$) where $R_{P_2}(\alpha)$ denotes the 100αth percentile of R_{P_2}.

When the normality assumptions are violated, the Box-Cox transformation is utilized as P_2 is invariant under monotonic transformations. Assume the data after transformation does follow the normality assumptions, then the **GI** method can be applied. Such confidence interval is noted as **BCGI** hereafter.

2.2. Non-parametric Approaches

The P_2 as given in (1) can be non-parametrically estimated as

$$\widehat{\overline{P}}_2=\frac{\sum_{i=1}^{n_2} I_{[\hat{F}_1^{-1}(P_1)\leq Y_i\leq \hat{F}_3^{-1}(1-P_3)]}}{n_2} \qquad (2)$$

With a bootstrap sample $\widehat{\overline{P}}_2^b$ ($b = 1$ to 500), the $100(1 - \alpha)$% bootstrap percentile confidence interval (**BTP**) can be obtained as

$$(\widehat{\overline{P}}_2^b(\alpha/2),\quad \widehat{\overline{P}}_2^b(1-\alpha/2))$$

where $\widehat{\overline{P}}_2^b(\alpha)$ is the 100α% percentile. An adjusted estimator of P_2 proposed by Agresti and Coull (1998) is

$$\widehat{\tilde{P}}_2 = \frac{\sum_{i=1}^{n_2} I_{[\hat{F}_1^{-1}(P_1) \le Y_i \le \hat{F}_3^{-1}(1-P_3)]} + z_{1-\alpha/2}^2/2}{n_2 + z_{1-\alpha/2}^2} \qquad (3)$$

where $z_{1-\alpha/2}$ stands for $100(1 - \alpha/2)\%$ percentile for standard normal distribution. The $100(1 - \alpha)\%$ **BTI** confidence interval is

$$\left(\widehat{\tilde{P}}_2 - z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}^{\mathrm{boot}}(\widehat{\tilde{P}}_2)},\ \ \widehat{\tilde{P}}_2 + z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}^{\mathrm{boot}}(\widehat{\tilde{P}}_2)}\right)$$

where $\widehat{\mathrm{Var}}^{\mathrm{boot}}(\widehat{\tilde{P}}_2)$ is the bootstrap estimate for the variance of $\widehat{\tilde{P}}_2$ (more details can be found in Dong et al. (2011)). Replacing $\widehat{\tilde{P}}_2$ with the mean $\overline{\widehat{\tilde{P}}}{}_2^b$ obtained from the bootstrap sample, the $100(1 - \alpha)\%$ **BTII** confidence interval is given as

$$\left(\overline{\widehat{\tilde{P}}}{}_2^b - z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}^{\mathrm{boot}}(\widehat{\tilde{P}}_2)},\ \ \overline{\widehat{\tilde{P}}}{}_2^b + z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}^{\mathrm{boot}}(\widehat{\tilde{P}}_2)}\right).$$

In Dong et al. (2011), through a simulation study, **GI** and **BCGI** were shown to provide accurate confidence intervals, given the corresponding normality assumptions were satisfied. Otherwise, **BTII** was recommended except in the scenarios with large P_2 and small sample sizes where **BTP** was preferred.

3. TWO NEW APPROACHES

In this section, two new methods for confidence interval estimation of P_2 are presented. Section 3.1 presents a method based on asymptotic normality and Section 3.2 presents two confidence intervals based on empirical likelihood.

3.1. Normal Approximation-Based Confidence Interval

For the diagnostic tests with binary diagnostic outcomes, Linnet (1987) provided the parametric formula for the variance of estimated sensitivity given the specificity, based on which normal approximation-based confidence interval was constructed. Further details can also be found in Zhou and Qin (2005) and Qin et al. (2011). Following the same vein, the variance of $\widehat{\overline{P}}_2$ can be proven as (see Appendix 1)

$$\sigma_{\widehat{\overline{P}}_2}^2 = \frac{P_2(1-P_2)}{n_2} + \frac{P_1(1-P_1)}{n_1} \cdot \frac{f_2^2[F_1^{-1}(P_1)]}{f_1^2[F_1^{-1}(P_1)]} + \frac{P_3(1-P_3)}{n_3} \cdot \frac{f_2^2[F_3^{-1}(1-P_3)]}{f_3^2[F_3^{-1}(1-P_3)]} \qquad (4)$$

where f_1, f_2 and f_3 are the probability density functions for Y_1, Y_2 and Y_3 respectively. It can be shown that when n_1, n_2 and n_3 are large, $\hat{\hat{P}}_2$ has an approximately normal distribution with mean P_2 and variance $\sigma^2_{\hat{\hat{P}}_2}$. The $\sigma^2_{\hat{\hat{P}}_2}$ can be estimated as

$$\widehat{\sigma^2_{\hat{\hat{P}}_2}} = \frac{\hat{\hat{P}}_2(1-\hat{\hat{P}}_2)}{n_2} + \frac{P_1(1-P_1)}{n_1} \cdot \frac{\hat{f}_2^2[\hat{F}_1^{-1}(P_1)]}{\hat{f}_1^2[\hat{F}_1^{-1}(P_1)]} + \frac{P_3(1-P_3)}{n_3} \cdot \frac{\hat{f}_2^2[\hat{F}_3^{-1}(1-P_3)]}{\hat{f}_3^2[\hat{F}_3^{-1}(1-P_3)]} \quad (5)$$

where $\hat{F}_1^{-1}(P_1)$ is the P_1th sample quantile of Y_1s, $\hat{F}_3^{-1}(1-P_3)$ is the $(1-P_3)$th sample quantile of Y_3s, and $\hat{f}_i$ is the kernel density estimate of f_i, $i = 1, 2, 3$. We use the "over-smoothed bandwidth selector" by Wand and Jones (1995) to select the bandwidth for the Gaussian kernel function. The $(1-\alpha)100\%$ normal approximation-based confidence interval

$$\left(\hat{\hat{P}}_2 - z_{1-\alpha/2}\sqrt{\widehat{\sigma^2_{\hat{\hat{P}}_2}}},\ \hat{\hat{P}}_2 + z_{1-\alpha/2}\sqrt{\widehat{\sigma^2_{\hat{\hat{P}}_2}}}\right)$$

is referred as asymptotic parametric variance confidence interval (**APV**) hereafter.

3.2. Empirical Likelihood Confidence Interval

Define an indicator function ϕ as

$$\phi(X,Y,Z) = \begin{cases} 1, & X<Y<Z \\ \frac{1}{2}, & X=Y<Z \ \text{ or } X<Y=Z \\ \frac{1}{6}, & X=Y=Z \\ 0, & \text{otherwise.} \end{cases}$$

Given P_1 and P_3, for a test result Y of a subject from the early diseased group, define a random variable

$$U = \phi[F_1^{-1}(P_1), Y, F_3^{-1}(1-P_3)].$$

It is evident that

$$\begin{aligned} E(U) &= E\{\phi[F_1^{-1}(P_1), Y, F_3^{-1}(1-P_3)]\} \\ &= P[F_1^{-1}(P_1) < Y < F_3^{-1}(1-P_3)] \\ &= P[F_1^{-1}(P_1) < Y \le F_3^{-1}(1-P_3)] \\ &= P_2. \end{aligned}$$

Based on this relationship between P_2 and U, we can develop an empirical likelihood procedure for making inference about P_2. Let $\mathbf{p} = (p_1, \ldots, p_{n_2})$ be a probability vector for the early diseased group, and $\sum_{i=1}^{n_2} p_i = 1$ and $p_i \ge 0$ for all i. The empirical likelihood for P_2 can be defined as

$$\tilde{L}(P_2)=\sup\left\{\prod_{i=1}^{n_2}p_i:\sum_{i=1}^{n_2}p_i=1,\sum_{i=1}^{n_2}p_i(U_i-P_2)=0\right\}$$

where $U_i=\phi[F_1^{-1}(P_1),Y_i,F_3^{-1}(1-P_3)]$, i = 1,2,…,n_2. Since U_i's depend on the unknown distribution functions F_1 and F_3, we replace them by their empirical distributions $\hat{F}_1$ and $\hat{F}_3$, and obtain a profile empirical likelihood for P_2

$$L(P_2)=\sup\left\{\prod_{i=1}^{n_2}p_i:\sum_{i=1}^{n_2}p_i=1,\sum_{i=1}^{n_2}p_i(\hat{U}_i-P_2)=0\right\}$$

where $\hat{U}_i=\phi[\hat{F}_1^{-1}(P_1),Y_i,\hat{F}_3^{-1}(1-P_3)]$, i = 1, 2,…, n_2. By the Lagrange multiplier method, we can easily obtain the following expression for p_i

$$\tilde{p}_i=\frac{1}{n_2}\left\{1+\tilde{\lambda}(\hat{U}_i-P_2)\right\}^{-1}$$

where $\tilde{\lambda}$ is the solution of

$$\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{\hat{U}_i-P_2}{1+\tilde{\lambda}(\hat{U}_i-P_2)}=0. \qquad (6)$$

Note that $\prod_{i=1}^{n_2}p_i$, subject to $\sum_{i=1}^{n_2}p_i=1$, attains its maximum $n_2^{-n_2}$ at $p_i=n_2^{-1}$. The profile empirical likelihood ratio for P_2 is defined as

$$r(P_2)=\prod_{i=1}^{n_2}(n_2\tilde{p}_i)=\prod_{i=1}^{n_2}\left\{1+\tilde{\lambda}(\hat{U}_i-P_2)\right\}^{-1}.$$

Hence the corresponding profile empirical log-likelihood ratio is

$$l(P_2)\equiv-2\ \log\ r(P_2)=2\sum_{i=1}^{n_2}\log\ \{1+\tilde{\lambda}(\hat{U}_i-P_2)\} \qquad (7)$$

where $\tilde{\lambda}$ is the solution of (6).

Since the profile empirical log-likelihood ratio $l(P_2)$ is a sum of dependent variables, its asymptotic distribution is no longer a standard chi-square distribution. In the Appendix 2, it is proven that $l(P_2)$ follows a scaled χ^2 distribution. The asymptotic distribution of $l(P_2)$ is summarized in the following theorem.

Theorem—Assume that F_1, F_2 and F_3 are continuous distribution functions, and the density functions f_1, f_2 and f_3 are positive and continuous at c_1 and c_2. If $0 < \rho_1 = \lim_{n_1, n_2 \to \infty} n_1/n_2 < \infty$, $0 < \rho_2 = \lim_{n_2, n_3 \to \infty} n_3/n_2 < \infty$, and P_2 is the true value of the sensitivity to the early diseased stage given specificity and the sensitivity to the fully diseased stage, the limiting distribution of $l(P_2)$, defined by (7), is a scaled chi-square distribution with one degree of freedom. That is,

$$r_{P_1,P_2,P_3} \cdot l(P_2) \xrightarrow{\mathscr{L}} \chi_1^2$$

where the scale constant r_{P_1,P_2,P_3} is

$$r_{P_1,P_2,P_3} = \frac{\sigma^2_{\hat{U}_i}}{n_2 \cdot \sigma^2_{\widehat{\widehat{P}}_2}}$$

with $\sigma^2_{\hat{U}_i} = P_2(1 - P_2)$ and $\sigma^2_{\widehat{\widehat{P}}_2}$ as given in (4).

In order to construct confidence interval for P_2 based on the above Theorem, we need to estimate $\sigma^2_{\hat{U}_i}$ and $\sigma^2_{\widehat{\widehat{P}}_2}$. The $\sigma^2_{\hat{U}_i}$ can be estimated as $\widehat{\widehat{P}}_2 1(-\widehat{\widehat{P}}_2)$ and a Gaussian kernel was used to obtain a parametric estimation of $\sigma^2_{\widehat{\widehat{P}}_2}$, as shown in (5). The $100(1 - \alpha)\%$ **ELP** confidence interval for P_2 is

$$\mathrm{CI}_\alpha(P_2) = \left\{ P_2 : r^*_{P_1,P_2,P_3} \cdot l(P_2) \le \chi_1^2(1-\alpha) \right\}$$

where $r^*_{P_1,P_2,P_3} = \frac{\widehat{\widehat{P}}_2(1 - \widehat{\widehat{P}}_2)}{n_2 \cdot \widehat{\sigma^2_{\widehat{\widehat{P}}_2}}}$ and $\chi_1^2(1-\alpha)$ is the $(1 - \alpha)$th quantile of χ_1^2. The performance of this **ELP** method highly depends on the density estimates from the Gaussian kernel, whose bandwidth is chosen without a well recognized standard. Therefore, the following bootstrap approach is proposed to estimate $\sigma^2_{\widehat{\widehat{P}}_2}$ instead:

For $b = 1$ to $B = 500$ bootstrap iterations,

Step 1: Draw re-samples of sizes n_1, n_2, and n_3 with replacement from the non-diseased sample Y_{1j}'s, the early diseased sample Y_{2j}'s, and the fully diseased sample Y_{3j}'s respectively. Denote the bootstrap samples as $\{Y^b_{\mathrm{ij}}\}$, $i = 1, 2, 3$, $j = 1, 2, \ldots, n_i$.

Step 2: Calculate the bootstrap version of $\widehat{\widehat{P}}_2^b$ according to (2).

Step 3: The proposed bootstrap variance estimator for $\widehat{\widehat{P}}_2$ is defined as

$$\widehat{\sigma^2_{\widehat{\overline{P}}_2}}^{\,b} = \frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\overline{P}}_2^{\,b} - \overline{\widehat{\overline{P}}}_2^{\,b}\right)^2$$

where $\widehat{\overline{P}}_2$ is defined in (2).

This leads to the second 100(1 − α)% empirical likelihood confidence interval (**ELB**) for P_2

$$\mathrm{CI}_\alpha(P_2) = \left\{P_2 : r^*_{P_1,P_2,P_3} \cdot l(P_2) \le \chi^2_1(1-\alpha)\right\}$$

where $r^*_{P_1,P_2,P_3} = \dfrac{\widehat{\overline{P}}_2(1-\widehat{\overline{P}}_2)}{n_2 \cdot \widehat{\sigma^2_{\widehat{\overline{P}}_2}}^{\,b}}$ and $\chi^2_1(1-\alpha)$ is the (1 − α)th quantile of χ^2_1.

4. SIMULATION STUDIES

Simulation studies are carried out to compare the performance of the proposed empirical likelihood confidence intervals **ELP** and **ELB**, as well as the asymptotic confidence interval **APV**, with the existing ones, i.e. **GI, BCGI, BTP, BTII** proposed in Dong et al. (2011). As **BTI** is always inferior than **BTII**, it is not included in the tables.

We evaluate these approaches under the normal and beta distribution scenarios proposed in Dong et al. (2011), to check whether the new approaches can give comparable performance as the recommended **GI/BCGI** parametric approach where the normality assupmtions are satisfied with or without Box-Cox transformation. In addition, we also investigated the combined scenario where the normality assumptions cannot be met; that is, gamma for the non-diseased, log-normal for the early diseased and Weibull for the fully diseased group. The density functions for the combined distribution scenario are plotted in Figure 1. Sample sizes (n_1, n_2, n_3) are set as (10, 10, 10), (30, 30, 30), (50, 30, 30), (50, 50, 50), (100, 100, 100), (100, 50, 50) and (100, 100, 50). With a fixed 80% specificity and a fixed 80% sensitivity to the fully diseased stage, the parameters for the distributions are chosen correspondingly so that P_2 equals to 50% or 90%. Under each setting, 5,000 random samples are generated. The simulation results are presented in Tables 1–3.

Table 1 presents simulation results under the normal distributions. The performance of the newly proposed empirical likelihood confidence interval **ELB** is satisfactory in terms of coverage probability although the **ELB** tends to be slightly conservative for the small sample sizes. **ELP** performs well for P_2 = 0.5 except at the sample size (10, 10, 10), but becomes conservative when P_2 = 0.9. **BTII** gives good estimates at P_2 = 0.5, but when P_2 increases to 0.9, **BTII** obtains a 0.8956 coverage probability at the sample size 11 (10, 10, 10), which is much lower than the 95% nominal level. In addition, as the sample size increases, **BTII** grows conservative. The **BTP** interval is generally conservative. The normal approximation-based confidence interval **APV** is slightly conservative at small sample sizes. The generalized inference method **GI** performs the best in the closeness of coverage probability to the nominal level and the length of the confidence interval.

Table 2 presents simulation results for the beta distribution. The coverage probability of **ELB** remains conservative for the small sample sizes at $P_2 = 0.5$, however, when $P_2 = 0.9$, for the small sample size (10, 10, 10), **ELB** attains coverage probability which is very close to the nominal level, and is even better than the **BCGI** approach. The other empirical likelihood method **ELP**, yields satisfactory coverage probabilities when $P_2 = 0.5$ except at the sample size (10, 10, 10), while it is conservative for medium sample sizes when $P_2 = 0.9$. The non-parametric method **BTII** is satisfactory at $P_2 = 0.5$; while at $P_2 = 0.9$, it changes from being liberal to being conservative as sample sizes increase. The large sample method **APV** is generally liberal when sample sizes are small. The generalized inference approach with Box-Cox transformation is usually satisfactory, but it can be worse than **ELB** for a few scenarios, such as (100, 100, 50) at $P_2 = 0.5$ or (10, 10, 10) at $P_2 = 0.9$.

In Table 3, the simulation results for the combined distribution are presented. For such cases, the Box-Cox transformation fails to transform the data to the normal distributions. Therefore, as expected, the performance of **BCGI** is unsatisfactory. Generally speaking, the **ELB** method is close to the 95% nominal level except being slightly conservative at the sample size (10, 10, 10). The **ELP** method provide reasonable coverage at $P_2 = 0.5$ except for the sample size (10, 10, 10). however, it becomes conservative for $P_2 = 0.9$. **BTII** maintains the nominal level for most cases except for the sample size (10, 10, 10), where the coverage probability can be as low as 0.7848. In addition, for scenarios such as (100, 50, 50) and (100, 100, 50), **BTII** becomes more conservative than **ELB**. The **BTP** method is generally conservative except at the sample size (10, 10, 10) when $P_2 = 0.9$. The asymptotic approach **APV** remains liberal for most of the cases; 12 however, as the sample size increases to (100, 100, 100), the coverage probability is very close to the 95% nominal level.

In summary, the **GI** method or the **BCGI** method work well for normal and beta distributions, but becomes unusable for the combined distributions case, where the Box- Cox transformation fails to work. The performance of **APV** is very unstable as it is slightly conservative for the normal case and is generally liberal for the non-normal ones. The **BTII**, for large P_2's, is conservative under large unbalanced sample sizes and gives very liberal estimates under small sample sizes. The **BTP** produces conservative confidence intervals for most of the cases. The **ELP** performs well for scenarios with smaller P_2, but it turns out to be conservative for the cases with higher P_2. Finally, the proposed **ELB** method gives stable confidence interval estimation with coverage probability close to the nominal level for almost all cases, except that it can be slightly conservative under small sample sizes. Therefore, overall speaking, the **ELB** method is highly recommended, especially for the cases when normality assumptions are violated and Box-Cox transformation fails to work.

5. EXAMPLE

Alzheimer's disease (AD) is the most common form of dementia, and it is one of the most costly diseases for society in Europe and the United States. According to Wimo et al. (2013), the total estimated worldwide costs of dementia were US$604 billion in 2010. About 70% of the costs occurred in western Europe and North America. The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a research project that is designed to validate the use of biomarkers including blood tests, tests of cerebrospinal fluid, and MRI/PET imaging for Alzheimer's disease clinical trials and diagnosis. It aims to define the rate of progress of mild cognitive impairment (MCI) and AD, to develop improved methods for clinical trials, and to provide a large database which will improve design of clinical treatment trials.

In the ADNI database, there are many biomarkers to measure the disease progress of AD. Here we use a small subset which includes ratio of levels of protein Tau and protein $A\beta_{42}$ (TAU/ABETA), Fluoro Deoxy Glucose (FDG) and Alzheimer's Disease Assessment Scale (ADAS11) at the 24th month visit. The clinical dementia rating (CDR) denotes the severity of dementia and a global CDR is derived from individual ratings in multiple domains by an experienced clinician. CDR 0 indicates no dementia and CDR 0.5, 1, 2 and 3 represent very mild, mild, moderate, and severe dementia, respectively. Since patients with large CDR such as 2 or 3 are rarely available, patients with CDR greater than or equal to 1 are referred as the fully diseased group. CDR 0 and 0.5 refer to the non-diseased group and the early diseased group respectively. This subset contains 194, 290 and 183 subjects for the non-diseased, the early diseased, and the fully diseased group respectively. Due to missing values, the actual sample sizes for each variable may vary, as reported in Table 4. Figures 2 presents the estimated kernel densities of the three disease groups for TAU/ABETA, FDG and ADAS11 respectively. By utilizing the Shapiro-Wilk's normality test, TAU/ABETA is found to satisfy the normality assumptions after the Box-Cox transformation; for FDG, the original data meets the normality assumptions; and for ADAS11, the data either with or without the Box-Cox transformation cannot achieve the normality assumptions for all three groups simultaneously. Since the parametric assumptions are not met, **GI/BCGI** cannot be rationally applied. Therefore, only the other methods are used to analyze this variable. Table 5 presents the estimated confidence intervals of P_2 for each variable. Under the recommended **ELB** approach, ADAS11 achieves (0.4660, 0.6657) as its 95% confidence interval for P_2, suggesting it gives a mediocre performance to diagnose the early stage AD patients.

6. SUMMARY AND DISCUSSION

For disease processes with three ordinal stages, the sensitivity to the early diseased stage given specificity and sensitivity to the fully diseased stage, P_2, is considered as an important diagnostic accuracy index, especially for early disease detection. The higher P_2, the better the diagnostic ability of the diagnostic test or biomarker for identifying the early diseased stage. Therefore, an accurate estimation of the confidence interval for P_2 will facilitate investigators to identify the good biomarkers. This article proposes the **ELB** approach and compares it with the existing confidence intervals. Simulation studies show that **ELB** not only is more robust than parametric methods which heavily rely on the normality assumptions, but also generally gives more accurate confidence intervals than non-parametric methods, especially for unbalanced data sets. Therefore, the **ELB** method is highly recommended in practice.

For future work, following the same vein of Dong et al. (2014), we would like to develop the semi-parametric inference procedure for the difference of two correlated P_2's, based on the empirical likelihood technique.

Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimers Association; Alzheimers Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. The ADNI research was also supported by NIH grants P30 AG010129 and K01 AG030514.

APPENDIX 1: PROOF OF VARIANCE of P¯^2 in (4)

The asymptotic variance of $\widehat{\overline{P}}_2$ is shown in (4). The following is the proof.

Proof:

$$
\begin{aligned}
\sigma^2_{\widehat{\overline{P}}_2} &= E_{\widehat{C}_1,\widehat{C}_2}\{\mathrm{Var}_{\hat{F}_2}[\hat{F}_2(\hat{c}_1 \le Y \le \hat{c}_2)]\} + \mathrm{Var}_{\widehat{C}_1,\widehat{C}_2}\{E_{\hat{F}_2}[\hat{F}_2(\hat{c}_1 \le Y \le \hat{c}_2)]\} \\
&= E_{\widehat{C}_1,\widehat{C}_2}\left\{\frac{P_2(\hat{c}_1 \le Y \le \hat{c}_2)[1 - P_2(\hat{c}_1 \le Y \le \hat{c}_2)]}{n_2}\right\} + \mathrm{Var}_{\widehat{C}_1,\widehat{C}_2}[P_2(\hat{c}_1 \le Y \le \hat{c}_2)]\cdot
\end{aligned}
$$

As $\hat{c}_1 \xrightarrow{P} c_1$ and $\hat{c}_2 \xrightarrow{P} c_2$, and we assume P_2 is continuous, so

$$
E_{\widehat{C}_1,\widehat{C}_2}\left\{\frac{P_2(\hat{c}_1 \le Y \le \hat{c}_2)[1 - P_2(\hat{c}_1 \le Y \le \hat{c}_2)]}{n_2}\right\} \xrightarrow{P} \frac{P_2(c_1 \le Y \le c_2)[1 - P_2(c_1 \le Y \le c_2)]}{n_2} = \frac{P_2(1 - P_2)}{n_2}.
$$

Furthermore, since $\hat{c}_1 \perp \hat{c}_2$, we have

$$
\begin{aligned}
\mathrm{Var}_{\widehat{c_1},\widehat{c_2}}[P_2(\hat{c}_1 \le Y \le \hat{c}_2)] &= \mathrm{Var}_{\widehat{c_1},\widehat{c_2}}[F_2(\hat{c}_2) - F_2(\hat{c}_1)] \\
&= \mathrm{Var}_{\widehat{c_2}}[F_2(\hat{c}_2)] + \mathrm{Var}_{\widehat{c_1}}[F_2(\hat{c}_1)] \\
&= f_2^2(\hat{c}_2)\cdot \mathrm{Var}(\hat{c}_2) + f_2^2(\hat{c}_1)\cdot \mathrm{Var}(\hat{c}_1) \\
&= \frac{P_3(1-P_3)}{n_3\cdot f_3^2(\hat{c}_2)}\cdot f_2^2(\hat{c}_2) + \frac{P_1(1-P_1)}{n_1\cdot f_1^2(\hat{c}_1)}\cdot f_2^2(\hat{c}_1) \\
&\xrightarrow{P} \frac{P_3(1-P_3)}{n_3\cdot f_3^2(c_2)}\cdot f_2^2(c_2) + \frac{P_1(1-P_1)}{n_1\cdot f_1^2(c_1)}\cdot f_2^2(c_1).
\end{aligned}
$$

Hence

$$
\sigma^2_{\widehat{\overline{P}}_2} = \frac{P_2(1-P_2)}{n_2} + \frac{P_1(1-P_1)}{n_1}\cdot\frac{f_2^2[F_1^{-1}(P_1)]}{f_1^2[F_1^{-1}(P_1)]} + \frac{P_3(1-P_3)}{n_3}\cdot\frac{f_2^2[F_3^{-1}(1-P_3)]}{f_3^2[F_3^{-1}(1-P_3)]}.
$$

APPENDIX 2: PROOF OF THEOREM IN SECTION 3

Proof:

By similar arguments used in Owen (1990), we can easily show that $|\lambda| = O_p(n_2^{-1/2})$ and $max_{1 \le j \le n_2} |\hat{U} - P_2| = O(1)$ a.s.. Then we have

$$\begin{aligned} l(P_2) &= 2\sum_{j=1}^{n_2} \log\left\{1+\tilde{\lambda}(\hat{U}_j - P_2)\right\} \\ &= 2\sum_{j=1}^{n_2} \{\lambda(\hat{U}_j - P_2) - \tfrac{1}{2}\lambda^2(\hat{U}_j - P_2)^2\} + r_{n_2} \end{aligned}$$

where $\left| r_{n_2} \right| \le c\sum_{j=1}^{n_2} \left|\lambda(\hat{U}_j - P_2)\right|^3 \le c\left|\lambda^3\right| n_2 = O_p(n_2^{-1/2})$.

From (6),

$$\begin{aligned} \lambda &= \frac{\sum_{j=1}^{n_2}(\hat{U}_j - P_2)}{\sum_{j=1}^{n_2}(\hat{U}_j - P_2)^2} + o_p(n_2^{-1/2}), \\ \sum_{j=1}^{n_2} \lambda(\hat{U}_j - P_2) &= \sum_{j=1}^{n_2} \lambda(\hat{U}_j - P_2) + o_p(1). \end{aligned}$$

Therefore,

$$\begin{aligned} l(P_2) &= \sum_{j=1}^{n_2} \lambda(\hat{U}_j - P_2) + o_p(1) \\ &= \frac{[\sum_{j=1}^{n_2}(\hat{U}_j - P_2)]^2}{\sum_{j=1}^{n_2}(\hat{U}_j - P_2)^2} + o_p(1) \\ &= \frac{[\sqrt{n_2}(\hat{P}_2 - P_2)]^2}{\frac{1}{n_2}\sum_{j=1}^{n_2}(\hat{U}_j - P_2)^2} + o_p(1) \end{aligned}$$

where ϕ is defined in (6) and $\hat{P}_2$ is a three-sample statistic and

$$\hat{P}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} \phi\left[\hat{F}_1^{-1}(P_1), y_j, \hat{F}_3^{-1}(1 - P_3)\right].$$

From the previous proof and the central limit theorem, we know that $\left[\sqrt{n_2}(\hat{P}_2 - P_2)\right]^2$ is asymptotically normal with the variance $n_2 \cdot \sigma_{\widehat{P}_2}^2$. From the law of large numbers, we have

$$\frac{1}{n_2}\sum_{j=1}^{n_2}(U_j - P_2)^2 \xrightarrow{P} \mathrm{Var}(U_j).$$

It is easy to check

$$\left|\frac{1}{n_2}\sum_{j=1}^{n_2}(\hat{U}_j-P_2)^2-\frac{1}{n_2}\sum_{j=1}^{n_2}(U_j-P_2)^2\right|\xrightarrow{P}0.$$

Therefore, by the Slutsky Theorem,

$$r_{P_1,P_2,P_3}\cdot l(P_2)\xrightarrow{\mathscr{L}}\chi_1^2$$

where the scale constant r_{P_1,P_2,P_3} is

$$r_{P_1,P_2,P_3}=\frac{\sigma^2_{\hat{U}_i}}{n_2\cdot\sigma^2_{\widehat{\overline{P}}_2}}.$$

References

Agresti A, Coull BA. Approximate is better than ”exact” for interval estimation of Binomial proportions. The American Statistician. 1998; 52:119–126.

Claeskens G, Jing BY, Peng L, Zhou W. An empirical likelihood confidence interval for an ROC curve. The Canadian Journal of Statistics. 2003; 31:173–190.

Dong T, Tian L, Hutson A, Xiong CJ. Parametric and non-parametric confidence intervals of the probability of identifying early disease stage given sensitivity to full disease and specificity with three ordinal diagnostic groups. Statistics in Medicine. 2011; 30:3532–3545. [PubMed: 22139763]

Dong T, Kang L, Hutson A, Xiong CJ, Tian L. Confidence interval estimation of the difference between two sensitivities to the early disease stage. Biometrical Journal. 2014; 56:270–286. [PubMed: 24265123]

Dreiseitl S, Ohno-Machado L, Binder M. Comparing three-class diagnostic tests by three-way ROC analysis. Medical Decision Making. 2000; 20:323–331. [PubMed: 10929855]

Greenhouse SW, Mantel N. The evaluation of diagnostic tests. Biometrics. 1950; 6:399–412. [PubMed: 14791576]

Heckerling PS. Parametric three-way receiver operating characteristic surface analysis using Mathematica. Medical Decision Making. 2001; 21:409–417. [PubMed: 11575490]

He X, Frey EC. The meaning and use of the volume under a three-class ROC surface (VUS). IEEE Transactions on Medical Imaging. 2008; 27:577–588. [PubMed: 18450532]

He X, Gallas BD, Frey EC. Three-Class ROC analysis—toward a general decision theoretic solution. IEEE Transactions on Medical Imaging. 2010; 29:206–215. [PubMed: 19884079]

Kang L, Tian L. Estimation of the volume under the ROC surface with three ordinal diagnostic categories. Computational Statistics & Data Analysis. 2013; 62:39–51.

Li J, Zhou XH. Nonparametric and semiparametric estimation of the three way receiver operating characteristic surface. Journal of Statistical Planning and Inference. 2009; 139:4133–4142.

Li J, Zhou XH, Fine JP. A regression approach to ROC surface, with applications to Alzheimer’s disease. Science China Mathematics. 2012; 55:1583–1595. [PubMed: 24459466]

Linnet K. Comparison of quantitative diagnostic tests: type I error, power, and sample size. Statistics in Medicine. 1987; 6:147–158. [PubMed: 3589244]

McNeil BJ, Hanley JA. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Medical Decision Making. 1984; 4:137–150. [PubMed: 6472062]

Mossman D. Three-way ROCs. Medical Decision Making. 1999; 19:78–89. [PubMed: 9917023]

Nakas CT, Alonzo TA, Yiannoutsos CT. Accuracy and cut-off point selection in three-class classification problems using a generalization of the Youden index. Statistics in Medicine. 2010; 29:2946–2955. [PubMed: 20809485]

Nakas CT, Yiannoutsos CT. Ordered multiple-class ROC analysis with continuous measurements. Statistics in Medicine. 2004; 23:3437–3449. [PubMed: 15505886]

Owen A. Empirical likelihood ratio confidence regions. Annals of Statistics. 1990; 18:90–120.

Owen, A. Empirical likelihood. New York: Chapman & Hall/CRC; 2001.

Pepe, MS. The statistical evaluation of medical tests for classification and prediction. Oxford Statistical Science Series; 2003. p. 28

Platt RW, Hanley JA, Yang H. Bootstrap confidence intervals for the sensitivity of a quantitative diagnostic test. Statistics in Medicine. 2000; 19:313–322. [PubMed: 10649298]

Qin GS, Davis AE, Jing BY. Empirical likelihood-based confidence intervals for the sensitivity of a continuous-scale diagnostic test at a fixed level of specificity. Statistical Methods in Medical Research. 2011; 20:217–231. [PubMed: 19654172]

Shapiro D. The interpretation of diagnostic tests. Statistical Methods in Medical Research. 1999; 8:113–134. [PubMed: 10501649]

Tian L, Xiong C, Lai C, Vexler A. Exact confidence interval estimation for the difference in diagnostic accuracy with three ordinal diagnostic groups. Journal of Statistical Planning and Inference. 2010; 141:549–558. [PubMed: 23538945]

Wan S. An empirical likelihood confidence interval for the volume under ROC surface. Statistics & Probability Letters. 2012; 82:1463–1467.

Wand, MP., Jones, MC. Kernel smoothing. New York: Chapman & Hall/CRC; 1995.

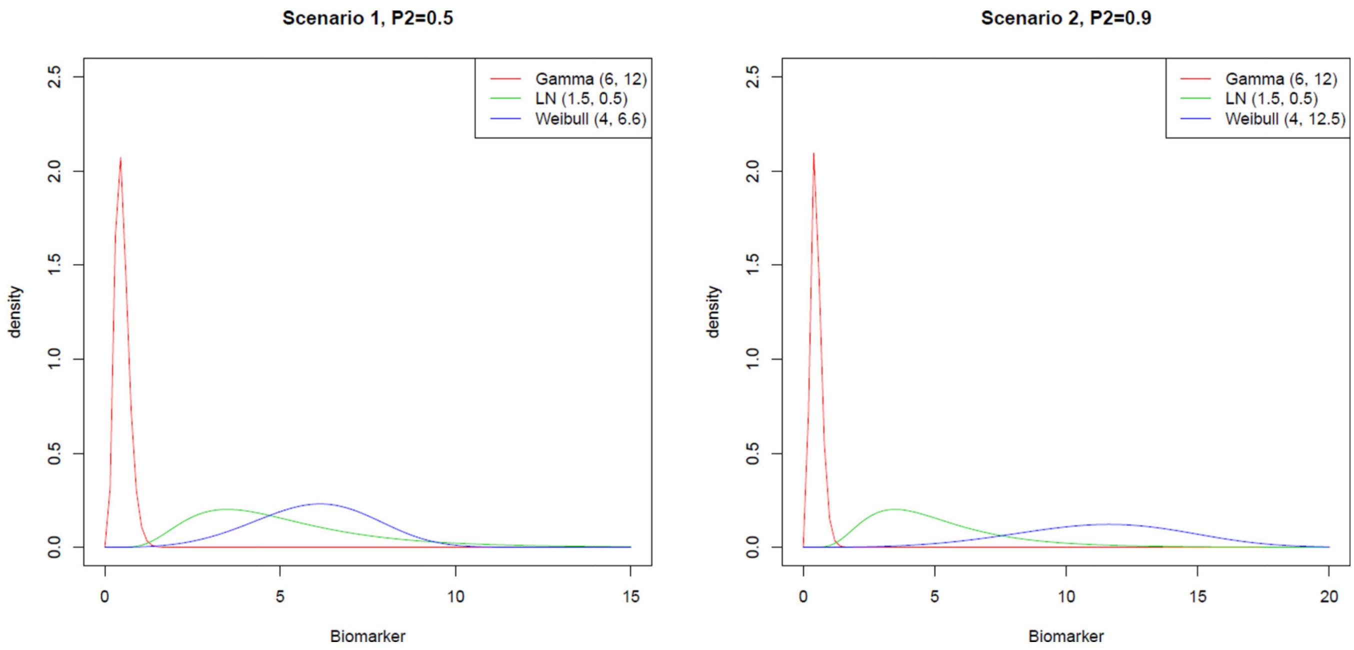
Wimo A, Jönsson L, Bond J, Prince M, Winblad B. The worldwide economic impact of dementia 2010. Alzheimer’s & dementia : the journal of the Alzheimer’s Association. 2013; 9:1–11.

Xiong CJ, Gerald VB, Philip M, John CM. Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups. Statistics in Medicine. 2006; 25:1251–1273. [PubMed: 16345029]

Zhou, XH., Obuchowski, N., McClish, D. Statistical methods in diagnostic medicine. Wiley; New York: 2002.

Zhou XH, Qin GS. Improved Confidence Intervals for the sensitivity to full disease at a fixed Level of specificity of a continuous-scale diagnostic test. Statistics in Medicine. 2005; 24:465–477. [PubMed: 15635678]

Zou, KH., Liu, A., Bandos, A., Ohno-Machado, L., Rockette, H. Statistical evaluation of diagnostic performance: topics in ROC analysis. CRC Press; 2010.


Scenario 1, P2=0.5

Scenario 2, P2=0.9

Figure 1.

Density functions for the non-diseased, early diseased and fully diseased group for the two simulation scenarios in Table 3.

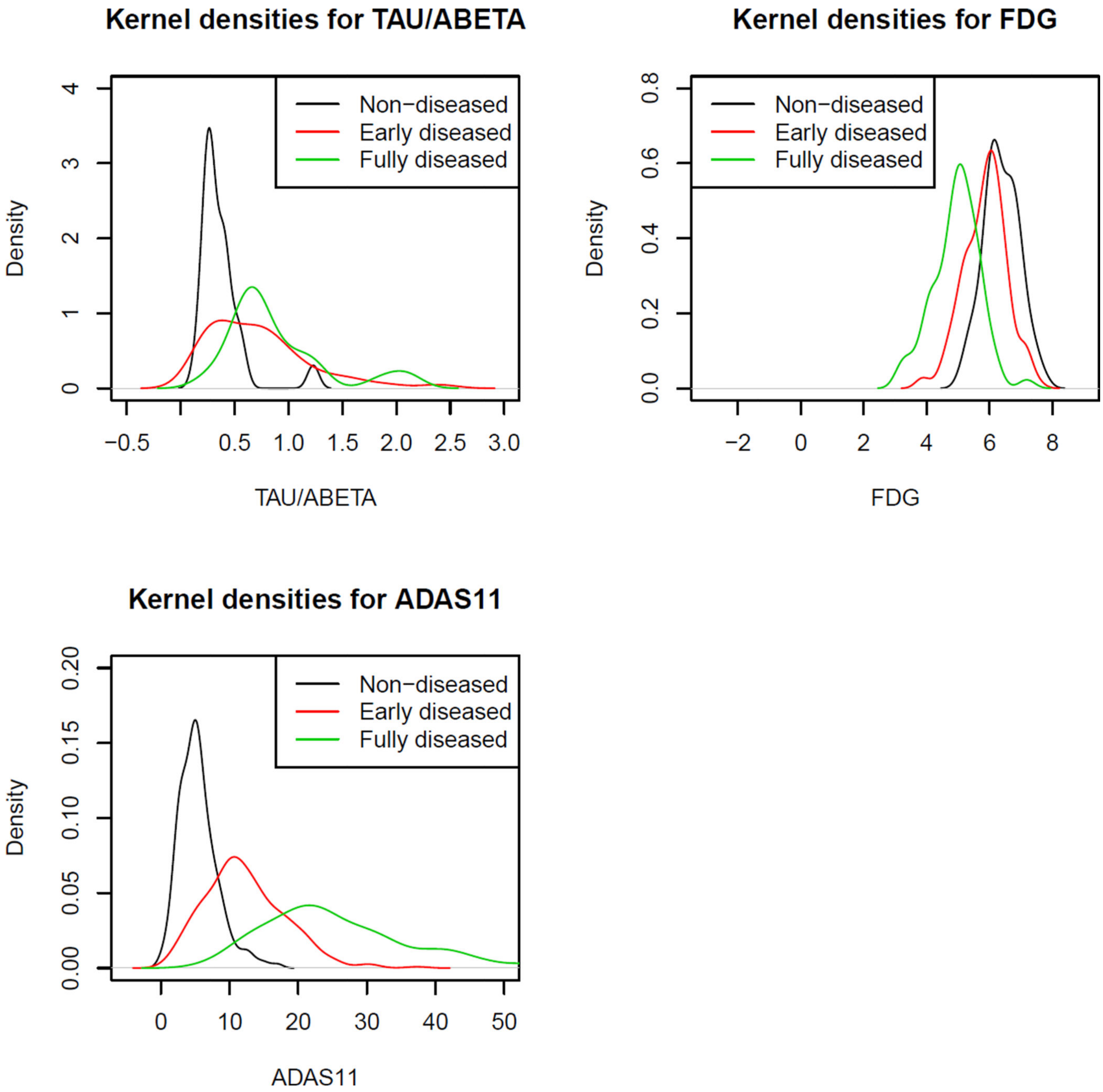


Figure 2.

Estimated kernel densities for TAU/ABETA, FDG and ADAS11 in the ADNI data.

Table 1

Summary of approximate 95% two-sided confidence bounds of BTII, BTP, ELB, ELP, GI and APV for P_2 under normal distributions (based on 5,000 simulations).

Three Independent Normal Distributions												
	Coverage Probability						**Length of Confidence Interval**					
$(\mu_1, \sigma_1) = (0, 1)'$, $(\mu_2, \sigma_2) = (2.5, 1.1)'$, $(\mu_3, \sigma_3) = (3.69, 1.2)'$, $P_2 = 0.5$												
	Non-parametric		**Empirical**		**Parametric**		**Non-parametric**		**Empirical**		**Parametric**	
Sample Sizes	**BTII**	**BTP**	**ELB**	**ELP**	**GI**	**APV**	**BTII**	**BTP**	**ELB**	**ELP**	**GI**	**APV**
(10, 10, 10)	0.9376	0.9774	0.9782	0.9976	0.9632	0.9782	0.6372	0.8109	0.6990	0.7072	0.6930	0.6990
(30, 30, 30)	0.9580	0.9756	0.9622	0.9468	0.9576	0.9622	0.5107	0.5571	0.5154	0.4927	0.4328	0.5154
(50, 30, 30)	0.9538	0.9728	0.9584	0.9478	0.9518	0.9584	0.5026	0.5487	0.5112	0.4878	0.4223	0.5112
(50, 50, 50)	0.9604	0.9724	0.9564	0.9440	0.9516	0.9564	0.4230	0.4441	0.4271	0.4035	0.3359	0.4271
(100, 100, 100)	0.9532	0.9642	0.9554	0.9490	0.9488	0.9554	0.3121	0.3168	0.3140	0.2982	0.2383	0.3140
(100, 50, 50)	0.9502	0.9710	0.9518	0.9414	0.9486	0.9518	0.4130	0.4346	0.4175	0.3963	0.3302	0.4175
(100, 100, 50)	0.9416	0.9656	0.9486	0.9392	0.9524	0.9486	0.3813	0.3880	0.3764	0.3610	0.3063	0.3764
$(\mu_1, \sigma_1)= (0, 1)'$, $(\mu_2, \sigma_2) = (2.5, 1.1)'$, $(\mu_3, \sigma_3) = (5.51, 1.2)'$, $P_2 = 0.9$												
	Non-parametric		Empirical		Parametric		Non-parametric		Empirical		Parametric	
Sample Sizes	BTII	BTP	ELB	ELP	GI	APV	BTII	BTP	ELB	ELP	GI	APV
(10, 10, 10)	0.8956	0.9460	0.9600	0.9588	0.9350	0.9600	0.3639	0.4577	0.5525	0.5853	0.5243	0.5525
(30, 30, 30)	0.9696	0.9836	0.9732	0.9794	0.9454	0.9732	0.2607	0.2763	0.3010	0.3258	0.2386	0.3010
(50, 30, 30)	0.9636	0.9868	0.9690	0.9748	0.9458	0.9690	0.2458	0.2611	0.2854	0.3110	0.2249	0.2854
(50, 50, 50)	0.9754	0.9816	0.9594	0.9798	0.9440	0.9594	0.2065	0.2160	0.2219	0.2341	0.1757	0.2219
(100, 100, 100)	0.9670	0.9774	0.9556	0.9608	0.9478	0.9556	0.1470	0.1497	0.1489	0.1522	0.1194	0.1489
(100, 50, 50)	0.9732	0.9806	0.9576	0.9758	0.9424	0.9576	0.1922	0.2011	0.2088	0.2211	0.1640	0.2088
(100, 100, 50)	0.9716	0.9812	0.9582	0.9638	0.9516	0.9582	0.1605	0.1625	0.1623	0.1651	0.1304	0.1623

BTII: Confidence interval is computed by the BTII approach.
BTP: Confidence interval is computed by the BTP approach.
ELB: Confidence interval is computed by the ELB approach.
ELP: Confidence interval is computed by the ELP approach.
GI: Confidence interval is computed by the GI approach.
APV: Confidence interval is computed by the APV approach.

Table 2

Summary of approximate 95% two-sided confidence bounds of BTII, BTP, ELB, ELP, BCGI and APV for P_2 under beta distributions (based on 5,000 simulations).

Three Independent Beta Distributions												
	Coverage Probability						**Length of Confidence Interval**					
$(\alpha_1, \beta_1) = (1, 6)'$, $(\alpha_2, \beta_2) = (6, 6)'$, $(\alpha_3, \beta_3) = (9.6, 6)'$, $P_2 = 0.5$												
	Non-parametric		**Empirical**		**Parametric**		**Non-parametric**		**Empirical**		**Parametric**	
Sample Sizes	**BTII**	**BTP**	**ELB**	**ELP**	**BCGI**	**APV**	**BTII**	**BTP**	**ELB**	**ELP**	**BCGI**	**APV**
(10, 10, 10)	0.9426	0.9752	0.9818	0.9988	0.9630	0.8980	0.6124	0.7938	0.6827	0.6688	0.6530	0.7104
(30, 30, 30)	0.9632	0.9720	0.9724	0.9554	0.9484	0.9268	0.4755	0.5212	0.4892	0.4562	0.3798	0.4896
(50, 30, 30)	0.9588	0.9692	0.9626	0.9468	0.9490	0.9192	0.4611	0.5086	0.4743	0.4479	0.3724	0.4808
(50, 50, 50)	0.9580	0.9732	0.9626	0.9520	0.9524	0.9320	0.3850	0.4081	0.3930	0.3676	0.2930	0.3853
(100, 100, 100)	0.9596	0.9692	0.9544	0.9442	0.9314	0.9334	0.2819	0.2881	0.2829	0.2686	0.2064	0.2748
(100, 50, 50)	0.9598	0.9640	0.9636	0.9516	0.9400	0.9226	0.3760	0.3961	0.3839	0.3621	0.2881	0.3780
(100, 100, 50)	0.9578	0.9586	0.9510	0.9412	0.9348	0.9328	0.3350	0.3403	0.3352	0.3173	0.2525	0.3281
$(\alpha_1, \beta_1) = (1, 6)'$, $(\alpha_2, \beta_2) = (6, 6)'$, $(\alpha_3, \beta_3) = (20.4, 6)'$, $P_2 = 0.9$												
	Non-parametric		Empirical		Parametric		Non-parametric		Empirical		Parametric	
Sample Sizes	BTII	BTP	ELB	ELP	BCGI	APV	BTII	BTP	ELB	ELP	BCGI	APV
(10, 10, 10)	0.8842	0.9398	0.9578	0.9528	0.9282	0.7494	0.3785	0.4839	0.5588	0.5596	0.4577	0.3129
(30, 30, 30)	0.9696	0.9726	0.9648	0.9722	0.9358	0.9262	0.2629	0.2832	0.3054	0.3120	0.2157	0.2267
(50, 30, 30)	0.9648	0.9742	0.9652	0.9712	0.9494	0.9246	0.2410	0.2594	0.2833	0.2983	0.2063	0.2185
(50, 50, 50)	0.9740	0.9744	0.9634	0.9768	0.9404	0.9408	0.2072	0.2165	0.2248	0.2241	0.1598	0.1912
(100, 100, 100)	0.9696	0.9706	0.9602	0.9582	0.9428	0.9438	0.1461	0.1488	0.1491	0.1461	0.1088	0.1419
(100, 50, 50)	0.9692	0.9656	0.9588	0.9762	0.9536	0.9320	0.1910	0.1979	0.2045	0.2119	0.1517	0.1813
(100, 100, 50)	0.9736	0.9752	0.9564	0.9582	0.9434	0.9372	0.1585	0.1615	0.1599	0.1560	0.1150	0.1519

BTII: Confidence interval is computed by the BTII approach.
BTP: Confidence interval is computed by the BTP approach.
ELB: Confidence interval is computed by the ELB approach.
ELP: Confidence interval is computed by the ELP approach.
BCGI: Confidence interval is computed by the BCGI approach.
APV: Confidence interval is computed by the APV approach.

Table 3

Summary of approximate 95% two-sided confidence bounds of BTII, BTP, ELB, ELP, BCGI and APV for P_2 under the combined distributions (based on 5,000 simulations).

Independent Gamma, Log-normal and Weibull Distributions												
	Coverage Probability						**Length of Confidence Interval**					
***Gamma*(α, β) = (6, 12)′, *LN*(μ, σ) = (1.5, 0.5)′, *Weibull*(a, b) = (4, 6.6)′, P_2 = 0.5**												
	Non-parametric		**Empirical**		**Parametric**		**Non-parametric**		**Empirical**		**Parametric**	
Sample Sizes	**BTII**	**BTP**	**ELB**	**ELP**	**BCGI**	**APV**	**BTII**	**BTP**	**ELB**	**ELP**	**BCGI**	**APV**
(10, 10, 10)	0.9242	0.9640	0.9716	0.9972	0.9374	0.8792	0.5512	0.7247	0.6301	0.6175	0.5895	0.6642
(30, 30, 30)	0.9538	0.9646	0.9596	0.9460	0.9120	0.9254	0.4254	0.4701	0.4468	0.4159	0.3524	0.4427
(50, 30, 30)	0.9570	0.9674	0.9568	0.9440	0.9098	0.9256	0.4281	0.4727	0.4468	0.4171	0.3528	0.4440
(50, 50, 50)	0.9562	0.9600	0.9564	0.9432	0.8984	0.9288	0.3513	0.3716	0.3600	0.3370	0.2741	0.3500
(100, 100, 100)	0.9586	0.9596	0.9530	0.9448	0.8702	0.9400	0.2591	0.2654	0.2619	0.2474	0.1944	0.2528
(100, 50, 50)	0.9536	0.9620	0.9516	0.9404	0.9028	0.9246	0.3528	0.3733	0.3584	0.3366	0.2742	0.3504
(100, 100, 50)	0.9530	0.9578	0.9436	0.9330	0.8698	0.9294	0.3085	0.3123	0.3080	0.2890	0.2255	0.2979
Gamma(α, β) = (6, 12)′, *LN*(μ, σ) = (1.5, 0.5)′, *Weibull*(a, b) = (4, 12.5)′, P_2 = 0.9												
	Nonparametric		Empirical		Parametric		Nonparametric		Empirical		Parametric	
Sample Sizes	BTII	BTP	ELB	ELP	BCGI	APV	BTII	BTP	ELB	ELP	BCGI	APV
(10, 10, 10)	0.7848	0.8628	0.9682	0.9702	0.9422	0.6998	0.3174	0.4037	0.5712	0.5724	0.3895	0.3043
(30, 30, 30)	0.9566	0.9628	0.9582	0.9824	0.9188	0.9238	0.2394	0.2534	0.2777	0.3180	0.2066	0.2266
(50, 30, 30)	0.9520	0.9638	0.9620	0.9862	0.9202	0.9284	0.2371	0.2520	0.2797	0.3178	0.2063	0.2260
(50, 50, 50)	0.9590	0.9620	0.9504	0.9822	0.9030	0.9312	0.1923	0.2002	0.2093	0.2276	0.1592	0.1910
(100, 100, 100)	0.9580	0.9606	0.9582	0.9642	0.8884	0.9436	0.1393	0.1413	0.1414	0.1474	0.1122	0.1447
(100, 50, 50)	0.9620	0.9606	0.9562	0.9876	0.9100	0.9248	0.1919	0.2004	0.2062	0.2265	0.1598	0.1916
(100, 100, 50)	0.9728	0.9612	0.9508	0.9658	0.8878	0.9332	0.1622	0.1645	0.1661	0.1728	0.1274	0.1661

BTII: Confidence interval is computed by the BTII approach.

BTP: Confidence interval is computed by the BTP approach.

ELB: Confidence interval is computed by the ELB approach.

ELP: Confidence interval is computed by the ELP approach.

BCGI: Confidence interval is computed by the BCGI approach.

APV: Confidence interval is computed by the APV approach.

Table 4

Summary Statistcs for ADNI data.

	CDR 0			CDR 0.5			CDR 1			
Biomarker	**N**	**Mean**	**Std**	**N**	**Mean**	**Std**	**N**	**Mean**	**Std**	**VUS**
TAU/ABETA	24	0.37	0.21	48	0.72	0.48	26	0.89	0.48	0.3890
FDG	82	6.37	0.56	130	5.86	0.68	70	4.95	0.74	0.5560
ADAS11	193	5.44	2.83	288	12.26	5.84	180	26.23	11.70	0.7575

Table 5

Estimated confidence intervals for the probability of detecting early diseased individuals using TAU/ABETA, FDG and ADAS11 of the ADNI data (sensitivity to fully diseased stage and specificity are assumed to equal to 0.8).

Confidence Intervals for the test covariates											
		BTII		BTP		ELB		GI		BCGI	
Biomarkers	$\hat{P}_2^{\mathrm{NP}}$	**lb**	**ub**	**lb**	**ub**	**lb**	**ub**	**lb**	**ub**	**lb**	**ub**
TAU/ABETA	0.1335	0.0052	0.2614	0.0371	0.2685	0.0073	0.3712	-	-	0.0000	0.2104
FDG	0.2011	0.0875	0.3388	0.1001	0.3620	0.0724	0.3716	0.0349	0.3152	-	-
ADAS11	0.5754	0.4806	0.6927	0.4829	0.6834	0.4660	0.6657	-	-	-	-

TAU/ABETA: Ratio of the CSF parameters: protein Tau and protein $A\beta 42$.

FDG: Fluoro Deoxy Glucose.

ADAS11: Alzheimer's Disease Assessment Scale.

BTII: Confidence interval is computed by the BTII approach.

BTII: Confidence interval is computed by the BTII approach.

ELB: Confidence interval is computed by the ELB approach.

ELP: Confidence interval is computed by the ELP approach.

APV: Confidence interval is computed by the APV approach.

$\hat{P}_2^{\mathrm{NP}}$: The nonparametric estimation of P_2 in equ (3).