HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2014 December 11.

Published in final edited form as:

Stat Med. 2011 December 30; 30(30): 3532–3545. doi:10.1002/sim.4401.

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

Tuochuan Dong[1], **Lili Tian**[1,*], **Alan Hutson**[1], and **Chengjie Xiong**[2]

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

[2]Division of Biostatistics, Washington University at St. Louis, Buffalo, St. Louis, MO 63110, USA

SUMMARY

In practice, there exist many disease processes with three ordinal disease classes; i.e. the non-diseased stage, the early disease stage and the fully diseased stage. Since early disease stage is likely the best time window for treatment interventions, it is important to have diagnostic tests which have good diagnostic ability to discriminate the early disease stage from the other two stages. In this paper, we present both parametric and non-parametric approaches for confidence interval estimation of probability of detecting early disease stage given the true classification rates for non-diseased group and diseased group, namely, the specificity and sensitivity to full disease. A data set on the clinical diagnosis of early stage Alzheimers disease (AD) from the neuropsychological database at the Washington University Alzheimers Disease Research Center (WU ADRC) is analyzed using the proposed approaches.

Keywords

Alzheimers disease (AD); generalized inference; Box-Cox transformation; bootstrap method

1. INTRODUCTION

The methods pertaining to statistical inferences involving diagnostic accuracy in the literature have largely focused on the cases when subjects are categorized in a binary fashion, i.e., the non-diseased and the diseased. The primary quantities of interest are the probabilities of an incorrect decision in the healthy population (1- specificity) and of a correct decision in the diseased population (sensitivity), respectively. When a diagnostic test is based on an observed variable that lies on a continuous or graded scale, an assessment of the test can be made through the use of a Receiver Operating Characteristic (ROC) curve, which is a plot of sensitivity against 1-specificity. For excellent reviews of statistical methods involving ROC curves; see Shapiro [1], Zhou et al. [2] and Pepe [3].

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

In reality, there exists a transitional stage (early disease stage) in many disease processes. In other words, a disease process might involves three ordinal diagnostic stages: the normal healthy stage without even the earliest subtle disease symptoms, the early stage of the disease, and stage of full-blown development of the disease. For example, mild cognitive impairment (MCI) and/or early stage Alzheimers disease (AD) is a transitional stage between the cognitive changes of normal aging and the more serious problems. More details can be seen from Xiong et al. [4]. To be specific, let Y_1, Y_2 and Y_3 denote the results of a diagnostic test and let F_1, F_2 and F_3 denote the corresponding cumulative distribution functions for healthy subjects, the subjects with early stage disease, and fully diseased subjects, respectively. Assume the results are measured on a continuous scale and that higher values indicate greater severity of the disease. Let $P_1 = F_1(c_1)$, $P_3 = 1 - F_3(c_3)$, where c_1 and c_3 are threshold values ($c_1 < c_3$) for classifying a subject into the non-diseased stage group and the fully diseased stage group, given that the subject is from these corresponding groups, respectively. Therefore, P_1 is specificity and P_3 is sensitivity to full disease. Then the probability that a randomly selected subject from the early disease stage group has a test result between c_1 and c_3, i.e. being correctly classified, is

$$P_2 = F_2(c_3) - F_2(c_1) = F_2[F_3^{-1}(1 - P_3)] - F_2[F_1^{-1}(P_1)]. \quad (1)$$

The P_2 can also be called sensitivity to early disease. As a function of P_1 and P_3, $P_2 = P_2(P_1, P_3)$ defines a surface in the three-dimensional space (P_1, P_3, P_2), i.e., the ROC surface. The point (P_1, P_3, P_2) = (1, 1, 1) indicates the perfect discrimination ability of the marker between three ordinal disease groups. The volume under the ROC surface (VUS) and the partial volume under surface (PVUS) have been widely used as quantitative indexes of discriminating ability of a biomarker measured on a continuous scale; e.g., see Mossman [5], Dreiseitl [6] and Heckerling [7]. Furthermore, Nakas and Yiannoutsos [8] proposed distribution-free approaches for hypothesis testing for a single VUS and paired VUSs; Xiong et. al. [4] developed an asymptotic approach for confidence interval estimation of VUS and PVUS for normally distributed data; Nakas and Alonzo [9], and Alonzo and Nakas [10] proposed nonparametric inference procedures for diagnostic accuracy with three disease classes under umbrella ordering; and Xiong et. al. [11] developed a large sample approach for comparing several VUSs for normally distributed data. Most recently, Tian et al. [12] proposed an approach based on generalized inference for confidence interval estimation of the difference between paired VUSs and PVUSs.

The probability associated with the detection of early disease stage, P_2, is especially critical in medical sense. First, in many disease processes such as AD, early detection often means optimum time window for therapeutic treatment due to the fact that no pharmaceutical treatments to-date are effective for the late stage AD. Therefore, estimating the probability that a person is at the early disease stage has direct treatment implications. Second, there are well established and accepted criteria for differentiating normal aging (i.e., P_1) and fully developed AD (i.e., P_3). 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. An accurate estimate of P_2 therefore helps clinicians to identify the best disease markers for early diagnosis. Finally, it is already a standard practice

for clinicians to diagnose subjects into 3 groups: normal aging, early stage/very mild AD, and fully developed AD. For disease processes with three disease stage, the specificity (P_1), the sensitivity to early disease (P_2), and the sensitivity to full disease (P_3) of a test or a biomarker depend on the cut-points c_1 and c_3 which can be chosen to be some quantile (typically the 80th, 90th or etc) of the distribution of the test values of non-diseased subjects and fully diseased subjects providing a fixed specificity (for example, the 80th percentile provides a specificity of 0.8) and a fixed sensitivity to full disease. In other words, the specification of P_1 and P_3 only serves to set up the cutoffs c_1 and c_3 for a disease marker that can be used to diagnose subjects into three groups. Therefore, the sensitivity to early disease (P_2) at a given specificity (P_1) and a give sensitivity to full disease (P_3) provides a measure of the ability of a biomarker for early disease detection and can be used as another diagnostic measure in addition to *VUS* and *PVUS*. Hence it is of paramount theoretical and practical importance to develop inference procedures for P_2. However, to the best of our knowledge, the problem of making inference about P_2 given P_1 and P_3 has not been addressed in the literatures.

When the disease status is binary, the diagnostic accuracy of a test is usually described by its sensitivity and specificity. For a continuous-scale test or biomarker, it is often of interest to construct a confidence interval for the sensitivity at the cut-off that yields a predetermined level of specificity. Towards this end, some works for estimation of sensitivity given specificity have been done. Linnet [13] proposed both parametric and non-parametric methods for constructing confidence intervals for the sensitivity of a test at a fixed value of specificity, accounting for the random variation associated with the estimated cut-off point. Platt [14] pointed out several shortcomings in Linnet [13] methods and then proposed to use Efron's bias-corrected acceleration (BCa) bootstrap interval. Zhou and Qin [15] proposed two new intervals for the sensitivity of a diagnostic test at a fixed level of specificity.

The purpose of this paper is two-fold: 1) For disease processes with three disease categories, we propose to use the sensitivity to early disease (P_2) given specificity (P_1) and sensitivity to full disease (P_3) as a diagnostic measure which focuses on ability of early disease detection; 2) we examine the performance of several parametric and nonparametric approaches for confidence interval estimation of P_2 given P_1 and P_3, and then make recommendations about what procedures are most appropriate to use under different scenarios. This paper is organized as follows. In Section 2, the motivating example from Washington University (WU) Alzheimers Disease Research Center (ADRC) is described. In Section 3, the parametric confidence interval estimations for P_2 under either normality or normality of transformed data are discussed. In Section 4, nonparametric confidence intervals for P_2 are presented. In Section 5, we conduct simulation studies to assess the finite sample performance of the proposed confidence intervals. In Section 6, we analyze the data from the Alzheimer's disease study. In Section 7, we give a summary and discussion.

2. THE DATA

Alzhermer's Disease(AD) is one of the most common degenerative dementias for aged people. As "baby boomers" reach retirement, AD is becoming even more prevalent, thus resulting a major health care crisis in the United States. Because AD is irreversible, a major

challenge is to identify individuals in the early phase of it. The sample studied here is from the longitudinal cohort of Washington University (WU) Alzheimers Disease Research Center (ADRC). Only individuals with dementia of Alzheimer type (DAT) were included in the demented sample. For each subject, the severity of dementia was staged by the Clinical Dementia Rating (CDR) according to published rules [16]. This data set includes three diagnostic groups: non-demented (CDR 0=group $D-$), very mildly demented (CDR 0.5=group D_0), and mildly demented (CDR 1=group $D+$). There are 45, 44, and 29 subjects in groups $D-$, D_0, $D+$ respectively. Besides clinical evaluation, participants also completed standard psychometric tests. Episodic memory was assessed by the Logic Memory, Digit Span (both forward and backward), Associate Learning sub-tests of the Wechsler Memory Scale (WMS), and the Visual Retention Test. Three measures of semantic memory included the Information subset of the Wechsler Adult Intelligence Scale (WAIS), the Boston Naming Test, and word fluency for S and P. The other two tests in the psychometrics battery were an attentional measure (WMS Mental Control) and an un-timed visuospatial measure (Visual Retention Test). The factor scores including primary factor (called global factor), the mental control/frontal factor, the memory-verbal/temporal factor, and the visuospatial/parietal factor were also computed from the database. These composite factor scores reflect the brain regions thought to contribute to performance on the measures that loaded highly on the factors. For more details about this data set and the description of the related psychometric tests, see Xiong et. al. [4].

The data set has been analyzed by Xiong et. al. [4] for confidence interval estimations of VUS and PVUS. The table with summary statistics of neuropsychometric tests from this sample is reproduced as Table 1. For Alzheimer's disease, a major challenge lies in identifying affected, but not yet fully demented individuals in the earliest phases of illness when treatment can have a more profound impact on functional status and rate of cognitive decline. Therefore, the goal of this paper is to to examine the accuracy of neuropsychological tests in the diagnosis of early stage AD given the sensitivity to full disease and specificity.

3. PARAMETRIC CONFIDENCE INTERVAL ESTIMATION OF P_2

In this section, we first examine a generalized inference approach for confidence interval estimation of P_2 for normally distributed data. For non-normal data, we propose to apply a Box-Cox type power transformation to the data followed by a generalized inference approach. The generalized variables and generalized pivots were introduced by Tsui and Weerahandi [17] and Weerahandi [18]; see the book by Weerahandi [19] for a detailed discussion. A brief summary of the core concepts is included in the Appendix. The concepts of generalized confidence interval and generalized P-value have been successfully applied to a variety of practical settings where standard exact solutions do not exist for confidence intervals and hypothesis testing. It has been shown that generalized inference approaches typically have good performance, even at small sample sizes; e.g. Weerahandi [20], Weerahandi and Berger [21], Krishnamoorthy and Lu [22], Tian and Cappelleri [23], Tian [24], Li, Liao and Liu [25] and Li, Liao and Liu [26].

3.1 Under the Normal Assumptions

Let Y_{1j} (j = 1, 2, …, n_1), Y_{2j} (j = 1, 2, …, n_2), and Y_{3j} (j = 1, 2, …, n_3) denote the n_1, n_2, n_3 observations for the non-diseased, early stage, and diseased groups respectively. Assume Y_{ij}(j = 1, 2, …, n_i) follows normal distributions with mean μ_i and variance σ_i^2 for i = 1, 2, 3. Then P_2 defined in (1) can be expressed as follows:

$$P_2=\Phi\left[\frac{\mu_3-\mu_2+\Phi^{-1}(1-P_3)\sigma_3}{\sigma_2}\right]-\Phi\left[\frac{\mu_1-\mu_2+\Phi^{-1}(P_1)\sigma_1}{\sigma_2}\right] \quad (2)$$

where Φ denotes the cumulative distribution function for the standard normal variable. For the ith group, let $\bar{Y}_i$ and S_i^2 be the sample mean and the sample variance, and let $\bar{y}_i$ and s_i^2 denote the corresponding observed values. The P_2 can be estimated as follows:

$$\hat{P}_2=\Phi\left[\frac{\bar{Y}_3-\bar{Y}_2+\Phi^{-1}(1-P_3)s_3}{s_2}\right]-\Phi\left[\frac{\bar{Y}_1-\bar{Y}_2+\Phi^{-1}(P_1)s_1}{s_2}\right]. \quad (3)$$

It is well-known that

$$V_i=(n_i-1)S_i^2/\sigma_i^2\sim\chi^2_{n_i-1}.$$

Therefore, the generalized pivotal quantity for σ_i^2 is

$$R_{\sigma_i^2}=\frac{(n_i-1)s_i^2}{V_i}\sim\frac{(n_i-1)s_i^2}{\chi^2_{n_i-1}}, i=1,2,3. \quad (4)$$

Furthermore,

$$Z_i=\frac{\bar{Y}_i-\mu_i}{\sqrt{\sigma_i^2/n_i}}\sim N(0,1), i=1,2,3.$$

The generalized pivotal quantity of μ_i is

$$R_{\mu_i}=\bar{y}_i-Z_i\sqrt{R_{\sigma_i^2}/n_i}, i=1,2,3. \quad (5)$$

The generalized pivotal quantities for normal mean and variance were first proposed in Krishnamoorthy and Mathew [27]. Finally, the generalized pivotal quantity for $P2$ is

$$R_{P_2}=\Phi\left[\frac{R_{\mu_3}-R_{\mu_2}+\Phi^{-1}(1-P_3)R_{\sigma_3}}{R_{\sigma_2}}\right]-\Phi\left[\frac{R_{\mu_1}-R_{\mu_2}+\Phi^{-1}(P_1)R_{\sigma_1}}{R_{\sigma_2}}\right] \quad (6)$$

where $R_{\sigma_i} = \sqrt{R_{\sigma_i^2}}$ for i = 1, 2, 3. One can easily check that R_{P_2} is a bona fide generalized pivot given the following holds: 1) the distributions of R_{P_2} is independent of any unknown parameters; and 2) the observed value of R_{P_2} equals to P_2 as defined in equation (2) for given $\bar{y}_i$ and s_i^2 (i = 1, 2, 3).

Computing Algorithm—Given a normally distributed data set with n_1 non-diseased subject, n_2 subjects at early stage of the disease, and n_3 diseased subjects, the confidence interval for P_2 using generalized inference approach can be obtained via the following steps:

1. For i = 1, 2, 3, calculate $\bar{y}_i$ and s_i^2.
2. For i = 1, 2, 3, generate independent random numbers V_i from $\chi^2_{n_i-1}$, then calculate $R_{\sigma_i^2}$.
3. For i = 1, 2, 3, generate independent random numbers Z_i from standard normal distributions $N(0, 1)$, and V_i from $\chi^2_{n_i-1}$, then calculate R_{μ_i}.
4. Calculate R_{P_2} as in equation (6).
5. Repeat Steps 2–4 for a total B = 2500 times to obtain a set of values of R_{P_2}.

Denote $R_{P_2}(\alpha)$ as the 100αth percentile of R_{P_2} 's. Then $(R_{P_2}(\alpha/2); R_{P_2}(1-\alpha/2))$ is a two-sided 100$(1-\alpha)$% generalized confidence interval of P_2.

3.2 Without the Normal Assumptions

Most of the time the normal assumptions as given in Section 3.1 are not satisfied. For such data, we will examine the use of the generalized inference approach to non-normal data by first applying a Box-Cox transformation to the data and then applying the generalized inference procedure proposed in Section 3.1. Due to the fact that ROC is invariant under monotonic transformation, this type of approach has been found useful in the context of ROC analysis for a wide variety of situations (e.g. Zou et al. [28]; Zou and Hall [30]; Faraggi and Reiser [29]; Fluss et al. [31]; Schisterman et al. [32], [33]; Molodianovitch et al. [34]). By employing a similar technique, we can also show that the P_2 is invariant under monotonic transformations. Let

$$Y_i^{(\lambda)} = \begin{cases} \frac{Y_i^{\lambda}-1}{\lambda} & \lambda, \neq 0 \\ log(Y_i), & \lambda = 0 \end{cases} \quad (7)$$

for i = 1, 2, 3, where it is assumed that $Y_i^{(\lambda)} \sim N(\mu_i, \sigma_i^2)$. Based on the observations of three groups, the loglikelihood function can be readily obtained as follows:

$$\sum_i^3 \sum_j^{n_i} [-\frac{1}{2} log(2\pi) - \frac{Y_{ij}^{\lambda} - \mu_i}{2\sigma_i^2} + (\lambda - 1) log\, Y_{ij}]. \quad (8)$$

The maximum likelihood estimate (MLE) of λ can be obtained by maximizing the above function. After applying the Box-Cox transformation, the generalized inference approach proposed in Section 3.1 is applied directly to the transformed data.

4. NON-PARAMETRIC CONFIDENCE INTERVAL ESTIMATION OF P_2

The parametric approaches either rely on normality assumption or require solving an equation for Box-Cox transformation. Therefore, it is of also interest to examine the performance of non-parametric approaches for confidence interval estimation of P_2. Three nonparametric methods using bootstrap samples will be considered. The first is bootstrap percentile confidence interval, and the other two are based on the intervals proposed by Agresti and Coull [36] with variance estimated from bootstrap samples.

Assume the distributions for the non-diseased group (i.e. F_1) and the fully diseased group (i.e. F_3) are known. Define $A_j = I_{[F_1^{-1}(P_1) \le Y_{2j} \le F_3^{-1}(1-P_3)]}$ for (j = 1, 2, …, n_2). Therefore A_j's are Bernoulli random variables with the successful rate $P_2 = P[F_1^{-1}(P_1) \le Y_{2j} \le F_3^{-1}(1-P_3)]$. Let $\bar{P}_2 = \sum_{j=1}^{n_2} A_j/n_2$, the standard (1 − α)100% Wald interval for P_2 is

$$(\bar{P}_2 - z_{1-\alpha/2}\sqrt{\bar{P}_2(1-\bar{P}_2)/n_2}, \bar{P}_2 + z_{1-\alpha/2}\sqrt{\bar{P}_2(1-\bar{P}_2)/n_2})$$

where $z_{1-\alpha/2}$ stands for 100(1 − α/2)% percentile for standard normal distribution.

In reality, the true distributions F_1 and F_3 are unknown and therefore $F_1^{-1}(P_1)$ and $F_3^{-1}(1-P_3)$ need to be replaced by their sample estimates $\hat{F}_1^{-1}(P_1)$ and $\hat{F}_3^{-1}(1-P_3)$. The estimated $\hat{P}_2$ is given by

$$\hat{P}_2 = \frac{\sum_{i=1}^{n_2} I_{[\hat{F}_1^{-1}(P_1) \le Y_i \le \hat{F}_3^{-1}(1-P_3)]}}{n_2}. \quad (9)$$

The estimated 100(1 − α)% Wald interval for P_2 is

$$(\hat{P}_2 - z_{1-\alpha/2}\sqrt{\hat{P}_2(1-\hat{P}_2)/n_2}, \hat{P}_2 + z_{1-\alpha/2}\sqrt{\hat{P}_2(1-\hat{P}_2)/n_2}).$$

The Wald interval is known to have poor performance, especially for small sample sizes [35].

The bootstrap percentile confidence interval (BTP) use bootstrap samples to compute $\hat{\bar{P}}_2^b$ for b = 1 to 500 bootstrap iterations as follows:

$$(\hat{\bar{P}}_2^b(\alpha), \hat{\bar{P}}_2^b(1-\alpha)) \quad (10)$$

where $\hat{\bar{P}}_2^b(\alpha)$ is the 100α% percentile of the bootstrap distribution of $\hat{\bar{P}}_2$.

The AC interval, proposed by Agresti and Coull [36], is know to have good performance for binomial proportions. Applying it to our settings, the 100(1 − α)% AC interval for P_2 is:

$$(\tilde{P}_2 - z_{1-\alpha/2}\sqrt{\hat{Var}_{AC}(\tilde{P}_2)}, \tilde{P}_2 + z_{1-\alpha/2}\sqrt{\hat{Var}_{AC}(\tilde{P}_2)}), \quad (11)$$

where

$$\tilde{P}_2 = \frac{\sum_{i=1}^{n_2} I_{[F_1^{-1}(P_1) \le Y_{2i} \le F_3^{-1}(1-P_3)]} + z_{1-\alpha/2}^2/2}{n_2 + z_{1-\alpha/2}^2/2} \quad (12)$$

and

$$Var_{AC}(\tilde{P}_2) = \frac{\tilde{P}_2(1-\tilde{P}_2)}{n_2 + z_{1-\alpha/2}^2/2}. \quad (13)$$

The estimated $\tilde{P_2}$ is given by

$$\hat{\tilde{P}}_2 = \frac{\sum_{i=1}^{n_2} I_{[\hat{F}_1^{-1}(P_1) \le Y_i \le \hat{F}_3^{-1}(1-P_3)]} + z_{1-\alpha/2}^2/2}{n_2 + z_{1-\alpha/2}^2/2}. \quad (14)$$

The estimated variance $Var_{AC}(\tilde{P_2})$ can be obtained directly by substituting $\tilde{P_2}$ with $\hat{\tilde{P}}_2$ in equation (13).

Zhou and Qin [15] considered the non-parametric solution for estimating confidence intervals for the sensitivity at a fixed level of specificity of a diagnostic test with binary disease status, and proposed to use bootstrap methods to estimate the variance. This idea can be extended to estimate non-parametric confidence interval of P_2 given P_3 and P_1. Follow the same vein, we will use bootstrap methods to estimate the variance of $\hat{\tilde{P}}_2$. With the estimated variance, we then apply Agresti and Coull's idea to derive confidence intervals for P_2.

Computing Algorithms

Given a data set with n_1 non-diseased subjects, n_2 subjects at early stage of the disease, and n_3 diseased subjects, three nonparametric confidence intervals for P_2 discussed in this section can be obtained by the following algorithm.

For b = 1 to B (it is recommended that $B \ge 200$, e.g. [15]. In this paper we use 500) bootstrap iterations,

- Draw resamples of sizes n_1, n_2, and n_3 with replacements from the non-diseased sample Y_{1j}'s, the early stage sample Y_{2j}'s, and the diseased sample Y_{3j}'s, respectively. Denote the bootstrap samples $\{Y_{ij}^b\}$, i = 1, 2, 3, j = 1, 2, …, n_i.
- Calculate the bootstrap version of $\hat{\tilde{P}}_2^b$ and $\hat{\tilde{P}}_b^2$ according to (9) and (14) respectively.

The bootstrap percentile confidence interval (BTP) in (10) can be obtained by using the array $\hat{\tilde{P}}_2^b$ (b = 1, …, 500).

The proposed bootstrap variance estimator of $\hat{\tilde{P}}_2$ is defined as:

$$\hat{Var}^{boot}(\hat{\tilde{P}}_2)=\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\tilde{P}}_2^b-\bar{\hat{\tilde{P}}}_2^b)^2, \quad (15)$$

where $\bar{\hat{\tilde{P}}}_2^b=(1/B)\sum_{b=1}^{B}\hat{\tilde{P}}_2^b$. Similarly as Zhou and Qin [15], we propose two Agresti and Coull's confidence intervals for P_2 given P_1 and P_3. The first (1 − α)100% level interval, called BTI interval, is defined as

$$(\hat{\tilde{P}}_2-z_{1-\alpha}\sqrt{\hat{Var}^{boot}(\hat{\tilde{P}}_2)},\hat{\tilde{P}}_2+z_{1-\alpha}\sqrt{\hat{Var}^{boot}(\hat{\tilde{P}}_2)}) \quad (16)$$

where $\hat{\tilde{P}}_2$ is defined in (14). The second 100(1 − α)% level interval, called BTII interval, for P_2 is defined by

$$(\bar{\hat{\tilde{P}}}_2^b-z_{1-\alpha}\sqrt{\hat{Var}^{boot}(\hat{\tilde{P}}_2)},\bar{\hat{\tilde{P}}}_2^b+z_{1-\alpha}\sqrt{\hat{Var}^{boot}(\hat{\tilde{P}}_2)}). \quad (17)$$

5. SIMULATION STUDIES

Simulation studies are carried out to assess the coverage probabilities of the proposed confidence interval estimations (the generalized inference approach, the generalized inference approach with Box-Cox transformation, the percentile bootstrap interval (BTP), and two intervals based on Agresti and Coull's paper, namely BTI and BTII) for P_2 under different distributional assumptions: normal, beta and gamma. The AC interval proposed in equ. (11) using estimated $\tilde{P}_2$ and variance $Var_{AC}(\tilde{P}_2)$ has poor coverage accuracy and therefore is not considered here. Beta and gamma distributions are used as representatives of non normal distributions because they are widely used in practical application and also because they come with a variety of shapes.

To represent a wide rage of sample size settings, (n_1, n_2, n_3) is set as (10, 10, 10), (30, 30, 30), (20, 10, 10), (30, 20, 10), (50, 30, 30) and (50, 50, 50). With a fixed 80% specificity and a fixed 80% sensitivity to full disease, the parameters are chosen correspondingly so that P_2 equals to 50%, 70%, 80% and 90% respectively. For each parameter setting, 5,000 random samples are generated and the parametric and non-parametric confidence intervals proposed in Sections 3 and 4 are obtained. The simulation results are presented in Tables 2–4 for the bootstrap percentile approach (BTP), and the BTII approach and the generalized inference approach (without or with Box-Cox transformation). The BTI approach is not presented due to the fact that it is constantly inferior to BTII approach. The coverage probabilities, the coverage errors for lower and upper tails, i.e. the proportion of runs in which the lower (or upper) limit of the confidence interval excluding the true P_2 at nominal level, and the average lengths of proposed confidence intervals are presented.

Table 2 presents simulation results under normal assumption. The bootstrap percentile approach tends to slightly overestimate the coverage probability except it underestimates the coverage probabilities for small sample unbalanced case as $P_2 = 0.9$; while the confidence intervals by the BTII method are reasonably close to the nominal level for most of the scenarios except sometimes they tend to be liberal. The generalized confidence intervals are the most accurate despite the fact they tend to slightly underestimate the coverage probabilities when sample sizes are small and P_2 is large. Regardless of sample sizes and true values of P_2, the bootstrap percentile confidence interval has the longest length.

In Tables 3 and 4, we present simulation results at the nominal level of 95% for the beta and gamma distributions respectively. For beta and gamma distributions, simulation study shows the Box-Cox transferred data generally satisfies normality. Generally speaking, the Box-Cox transformed generalized approach gives uniformly good coverage probabilities for all cases, except that it might be slightly conservative when the sample sizes are small. The bootstrap percentile confidence interval are generally conservative except for small sample scenarios when $P_2 = 0.9$. The BTII approach performs well except that it tends to be liberal for small sample scenarios when $P_2 = 0.9$. Similarly as normal cases, the bootstrap percentile confidence intervals have the longest length.

In summary, as the normality is satisfied for either original data or the transformed data, the parametric approaches, i.e. the generalized approach or the Box-Cox transformed generalized approach can generally provide confidence intervals with satisfactory coverage probabilities. Although the generalized approach is simple to use, the Box-Cox transformation involves solving equation. On the other hand, when the normality assumption can not be met, the BTII approach is a good choice except the scenarios with large P_2 and small sample sizes, for which the bootstrap percentile approach can provide reasonable confidence intervals.

6. EXAMPLE: REVISITED

In this section, the confidence intervals of the probabilities of detecting early AD (P_2) for all neuropsychometric tests in the data set of Alzheimer's disease from a study at the Washington University (WU) Alzheimer's Disease Research Center (ADRC) are estimated by the proposed parametric and nonparametric approaches. The details of the data set are presented in Section 2 and the summary statistics of neuropsychometric tests by three diagnostic groups are presented in Table 1.

A close look at the data using Shapiro-Wilk normality test shows that the *frontal factor* and *temporal factor* satisfy normality, while *parietal factor*, *associate learning* and *word fluency* satisfy normality after a Box-Cox transformation. For these variables, the generalized inference approach or the Box-Cox transformed generalized inference are recommended. For the rest of variables, the confidence intervals of P_2 can be obtained by both BTP and BTII methods. The specificity P_1 and sensitivity to full disease P_3 are fixed at 0.8. For comparison purpose, the confidence intervals by all the proposed approaches are presented in Table 5. Furthermore, for each variable, the nonparametric and parametric (with or

without Box-Cox transformation) point estimates in (3) and (9) are calculated, and the corresponding most appropriate point estimates are highlighted in Table 5.

From the point estimates and estimated confidence intervals, we can see that the *global factor* has the best diagnostic accuracy for identifying early stage dementia, followed by *logical memory* and *visual retention*, while *word fluency* and *parietal factor* have very poor ability for early stage diagnosis.

7. SUMMARY AND DISCUSSION

The probability of detecting early disease stage (P_2) when disease processes involve three ordinal disease stages given sensitivity to the full disease (P_3) and specificity (P_1) can serve as another diagnostic measure. This article aims to examine the performance of several parametric and nonparametric approaches for confidence interval estimation of P_2 given P_1 and P_3 and to make recommendations about what procedures are most appropriate to use under different scenarios. These methods can be applied to identify important makers for the detection of early stage disease (e.g. preclinical AD) which is usually the most important stage of the disease for intervention. As the simulations results indicate, the parametric approaches generally perform satisfactorily. Out of non-parametric approaches, the bootstrap percentile approach generally slightly overestimate the coverage probabilities, while the BTII method is a good choice except the scenarios with large P_2, for which the bootstrap percentile approach can provide reasonable confidence intervals.

Based on the simulation studies, the following recommendations are made. First, if normality is satisfied for either original or transformed data, we suggest the generalized inference approach. This approach is easy to use, and has good coverage probability even for small sample sizes and unbalanced sample sizes. Second, if the normality assumption is not met, the nonparametric BTII approach works well for most scenarios; however, if the estimated P_2 is large and the sample sizes are small, we recommended the use of the bootstrap percentile approach (BTP). Furthermore, as sample sizes are ≥ 50, the BTII approach is recommended for normal distribution and the generalized inference approach is recommended for Beta and Gamma distributions due to the following facts: 1) BTII intervals have the shortest length and satisfactory coverage probabilities for normal distribution; 2) the generalized inference approach has the best coverage probabilities and the shortest confidence interval for Beta and Gamma.

All of the proposed approaches are simulation-based approach. The generalized inference approach based on normality is an easy-to-use approach while the Box-Cox transformed generalized inference approach involves solving an equation. The nonparametric approaches are simple except that the variance is computed through bootstrap samples. A R-program is available upon request from ltian@buffalo.edu.

Acknowledgments

The work by Dr. Xiong was partly supported by grants NIH/NIA R01 AG029672, AG003991, AG005681, and AG026276 from the National Institute on Aging and grant NIRG-08-91082 from the Alzheimer's Association.

Appendix

Generalized Pivots and Generalized Test variables

In the following, the basic concepts for generalized inference developed by Tsui & Weerahandi [17] and Weerahandi [18] are described.

Suppose that $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)'$ form a random sample from a distribution which depends on the parameters $\theta = (\psi, \boldsymbol{\nu})$ where ψ is the parameter of interest and $\boldsymbol{\nu}'$ is a vector of nuisance parameters. A generalized pivot $R(\boldsymbol{Y}; \boldsymbol{y}, \psi, \boldsymbol{\nu})$, where $\boldsymbol{y}$ is a observed value of $\boldsymbol{Y}$, for interval estimation defined in Weerahandi [18], has the following two properties:

1. $R(\boldsymbol{Y}; \boldsymbol{y}, \psi, \boldsymbol{\nu})$ has a distribution free of unknown parameters.
2. The value of $R(\boldsymbol{y}; \boldsymbol{y}, \psi, \boldsymbol{\nu})$ is ψ.

Let that R_α be the 100αth percentile of R. Then R_α becomes the $100(1 - \alpha)\%$ lower bound for ψ and $(R_{\alpha/2}, R_{1-\alpha/2})$ becomes a $100(1 - \alpha)\%$ two-sided generalized confidence interval for ψ.

Now consider testing $H_0 : \psi = \psi_0$ vs. $H_1 : \psi > \psi_0$ where ψ_0 is a specified quantity. A generalized test variable of the form $T(\boldsymbol{Y}; \boldsymbol{y}, \psi, \boldsymbol{\nu})$, where $\boldsymbol{y}$ is an observed value of $\boldsymbol{Y}$, is chosen to satisfy the following three conditions (Tsui & Weerahandi [17]) :

1. For fixed $\boldsymbol{y}$, the distribution of $T(\boldsymbol{Y}; \boldsymbol{y}, \psi, \boldsymbol{\nu})$ is free of the vector of nuisance parameters $\boldsymbol{\nu}$.
2. The value of $T(\boldsymbol{Y}; \boldsymbol{y}, \psi, \boldsymbol{\nu})$ at $\boldsymbol{Y} = \boldsymbol{y}$ is free of any unknown parameters.
3. For fixed $\boldsymbol{y}$ and $\boldsymbol{\nu}$, and for all t, $Pr[T(\boldsymbol{Y}; \boldsymbol{y}, \psi, \boldsymbol{\nu}) > t]$ is either an increasing or a decreasing function of ψ.

A generalized extreme region is defined as $C = [\boldsymbol{Y} : T(\boldsymbol{Y}; \boldsymbol{y}, \psi, \boldsymbol{\nu}) \geq T(\boldsymbol{y}; \boldsymbol{y}, \psi, \boldsymbol{\nu})]$ if $T(\boldsymbol{Y}; \boldsymbol{y}, \psi, \boldsymbol{\nu})$ is stochastically increasing in ψ. If $T(\boldsymbol{Y}; \boldsymbol{y}, \psi, \boldsymbol{\nu})$ is stochastically decreasing in ψ, a generalized extreme region is defined as $C = [\boldsymbol{Y} : T(\boldsymbol{Y}; \boldsymbol{y}, \psi, \boldsymbol{\nu}) \leq T(\boldsymbol{y}; \boldsymbol{y}, \psi, \boldsymbol{\nu})]$. Then the generalized P-value is defined as $P(C|\psi_0)$.

References

1. Shapiro D. The interpretation of diagnostic tests. Statistical Methods in Medical Research. 1999; 8:113–134. [PubMed: 10501649]
2. Zhou, X-H.; Obuchowski, N.; McClish, D. Statistical Methods in Diagnostic Medicine. New York: Wiley; 2002.
3. Pepe, MS. The statistical evaluation of medical tests for classification and prediction. Vol. 28. Oxford Statistical Science Series; 2003.
4. 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]
5. Mossman D. Three-way ROCs. Medical Decision Making. 1999; 19:78–89. [PubMed: 9917023]
6. 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]

7. Heckerling PS. Parametric three-way receiver operating characteristic surface analysis using Mathematica. Medical Decision Making. 2001; 21:409–417. [PubMed: 11575490]
8. Nakas C, Yiannoutsos C. Ordered multiple-class ROC analysis with continuous measurement. Statistics in Medicine. 2004; 23:3437–3449. [PubMed: 15505886]
9. Nakas C, Alonzo T. ROC graphs for assessing the ability of a diagnostic marker to detect three disease classes with an umbrella ordering. Biometrics. 2007; 63:603–609. [PubMed: 17688513]
10. Alonzo T, Nakas C. Comparison of ROC umbrella volumes with an application to the assessment of lung cancer diagnostic markers. Biometrical Journal. 2007; 49:654–664. [PubMed: 17726716]
11. Xiong C, van Belle G, Miller JP, Yan Y, Gao F, Feng S, Yu K, Morris JC. A parametric comparison of diagnostic accuracy with three ordinal diagnostic groups. Biometrical Journal. 2007; 49(5):682–693. [PubMed: 17763377]
12. 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]
13. Linnet K. Comparison of quantitative diagnostic tests: type I error, power, and sample size. Statistics in Medicine. 1987; 6:147–158. [PubMed: 3589244]
14. 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]
15. Zhou X-H, 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]
16. Morris JC. The clinical dementia rating (CDR): current version and scoring rules. Neurology. 1993; 43:2412–2414. [PubMed: 8232972]
17. Tsui KW, Weerahandi S, et al. Generalized P-values in significance testing of hypotheses in the presence of nuisance parameters. Journal of American Statistical Association. 1989; 84:602–607. (1989).
18. Weerahandi S. Generalized confidence intervals. Journal of American Statistical Association. 1993; 88:899–905.
19. Weerahandi, S. Exact Statistical Methods for Data Analysis. New York: Springer; 2003.
20. Weerahandi S. ANOVA under unequal error variances. Biometrics. 1995; 51:589–599.
21. Weerahandi S, Berger VW. Exact inference for growth curves with intraclass correlation structure. Biometrics. 1999; 55:921–924. [PubMed: 11315029]
22. Krishnamoorthy K, Lu Y. Inference on the common means of several normal populations based on the generalized variable method. Biometrics. 2003; 59:237–247. [PubMed: 12926708]
23. Tian L, Cappelleri JC. A new approach for interval estimation and hypothesis testing of a certain intraclass correlation coefficient: the generalized variable method. Statistics in Medicine. 2004; 23:2125–2135. [PubMed: 15211607]
24. Tian L. Confidence intervals for $P(Y1 > Y2)$ with normal outcomes in linear models. Statistics in Medicine. 2008; 27:4221–4237. [PubMed: 18407578]
25. Li C, Liao C, Liu J. On the exact interval estimation for the difference in paired areas under the ROC curves. Statistics in Medicine. 2008; 27:224–242. [PubMed: 17139702]
26. Li C, Liao C, Liu J. A non-inferiority test for diagnostic accuracy based on the paired partial areas under ROC curves. Statistics in Medicine. 2008; 27:1762–1776. [PubMed: 17968858]
27. Krishnamoorthy K, Mathew T. Inferences on the means of lognormal distributions using generalized p-values and generalized confidence intervals. Journal of Statistical Planning and Inference. 2003; 115:103–121.
28. Zou KH, Hall WJ, Shapiro DE. Smooth Non-Parametric Receiver Operating Characteristic (ROC) Curves for Continuous Diagnostic Tests. Statistics in Medicine. 1997; 16:2143–2156. [PubMed: 9330425]
29. Faraggi D, Reiser B. Estimation of the Area Under the ROC Curve. Statistics in Medicine. 2002; 21:3093–3106. [PubMed: 12369084]
30. Zou KH, Hall WJ. Two transformation models for estimating an ROC curve derived from continuous data. Journal of Applied Statistics. 2000; 27(5):621–631.

31. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cut-off point. Biometrical Journal. 2005; 47:458–472. [PubMed: 16161804]

32. Schisterman E, Faraggi D, Reiser B. Adjusting the Generalized ROC Curve for Covariates. Statistics in Medicine. 2004; 23:3319–3331. [PubMed: 15490426]

33. Schisterma E, Reiser B, Faraggi D. ROC analysis for markers with mass at zero. Statistics in Medicine. 2006; 25:623–638. [PubMed: 16345033]

34. Molodianovitch K, Faraggi D, Reiser B. Comparing the Areas Under Two Correlated ROC Curves: Parametric and Non-Parametric Approaches. Biometrical Journal. 2007; 5:745–757.

35. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. Statistical Science. 2001; 16:101–117.

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

Table 1

Means (standard deviations) of neuropsychometric tests from the WU ADRC sample

Variable	CDR 0 (n=45)	CDR 0.5 (n=44)	CDR 1 (n=29)
Global factor	0.569(0.888)	−1.622(1.722)	−4.199(1.699)
Frontal factor	2.866(1.777)	0.373(2.212)	−2.682(2.067)
Parietal factor	1.803(1.295)	−0.241(2.051)	−2.377(2.549)
Temporal factor	4.085(2.249)	−0.986(3.315)	−5.855(3.223)
Associate Learning	0.741(0.890)	−0.579(0.888)	−1.501(0.871)
Logical Memory	0.730(0.848)	−0.858(0.895)	−1.766(0.402)
Digit Span Forward	0.579(0.806)	−0.212(0.892)	−1.210(1.127)
Digit Span Backward	0.546(0.923)	−0.400(0.853)	−1.824(1.410)
Visual Retention (10 s)	0.636(0.879)	−0.821(1.099)	−1.658(0.773)
Information	0.631(0.844)	−0.607(1.080)	−2.302(1.139)
Word Fluency	0.729(1.178)	−0.255(0.981)	−1.438(0.883)
Mental Control	0.463(0.612)	−0.374(1.197)	−1.715(1.130)
Boston Naming	0.588(0.531)	−0.497(1.635)	−3.072(2.148)
Visual Retention (copy)	0.202(0.667)	−0.551(1.864)	−1.769(2.398)

Table 2

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

Three Independent Normal Distributions												
$(\mu_1, \sigma_1) = (0, 1)'$, $(\mu_2, \sigma_2) = (2.5, 1.1)'$, $(\mu_3, \sigma_3) = (3.69, 1.2)'$, $P_2 = 0.5$												
	Coverage Probability			Lower Tail			Upper Tail			Length of CI		
Sample Sizes	BTP	BTII	GI	BTP	BTII	GI	BTP	BTII	GI	BTP	BTII	GI
(10, 10, 10)	0.9766	0.9360	0.9606	0.0198	0.0502	0.0038	0.0036	0.0138	0.0356	0.8086	0.6364	0.6943
(30, 30, 30)	0.9792	0.9602	0.9572	0.0146	0.0276	0.0110	0.0062	0.0122	0.0318	0.5570	0.5117	0.4330
(20, 10, 10)	0.9774	0.9398	0.9632	0.0190	0.0494	0.0074	0.0036	0.0108	0.0294	0.8013	0.6280	0.6775
(50, 30, 30)	0.9740	0.9464	0.9490	0.0194	0.0382	0.0182	0.0066	0.0154	0.0328	0.5498	0.5032	0.4239
(50, 50, 50)	0.9752	0.9446	0.9564	0.0174	0.0338	0.0130	0.0074	0.0216	0.0306	0.4423	0.3121	0.3356
$(\mu_1, \sigma_1) = (0, 1)'$, (μ_2, σ_2) $(2.5, 1.1)'$, $(\mu_3, \sigma_3) = (4.31, 1.2)'$, $P_2 = 0.7$												
	Coverage Probability			Lower Tail			Upper Tail			Length of CI		
Sample Sizes	BTP	BTII	GI	BTP	BTII	GI	BTP	BTII	GI	BTP	BTII	GI
(10, 10, 10)	0.9834	0.9578	0.9476	0.0146	0.0242	0.0042	0.0020	0.0180	0.0482	0.7278	0.5694	0.6888
(30, 30, 30)	0.9772	0.9622	0.9536	0.0160	0.0236	0.0130	0.0068	0.0142	0.0334	0.4850	0.4445	0.3911
(20, 10, 10)	0.9826	0.9578	0.9506	0.0158	0.0238	0.0042	0.0016	0.0184	0.0452	0.7120	0.5531	0.6578
(50, 30, 30)	0.9786	0.9618	0.9546	0.0164	0.0256	0.0122	0.0050	0.0126	0.0332	0.4764	0.4352	0.3810
(50, 50, 50)	0.9752	0.9532	0.9498	0.0182	0.0254	0.0118	0.0066	0.0214	0.0384	0.3823	0.2681	0.3003
$(\mu_1, \sigma_1) = (0, 1)'$, $(\mu_2, \sigma_2) = (2.5, 1.1)'$, $(\mu_3, \sigma_3) = (4.73, 1.2)'$, $P_2 = 0.8$												
	Coverage Probability			Lower Tail			Upper Tail			Length of CI		
Sample Sizes	BTP	BTII	GI	BTP	BTII	GI	BTP	BTII	GI	BTP	BTII	GI
(10, 10, 10)	0.9720	0.9568	0.9358	0.0266	0.0206	0.0018	0.0014	0.0226	0.0624	0.6372	0.4999	0.6431
(30, 30, 30)	0.9848	0.9720	0.9464	0.0116	0.0142	0.0070	0.0036	0.0138	0.0466	0.4091	0.3704	0.3336
(20, 10, 10)	0.9684	0.9566	0.9360	0.0298	0.0212	0.0048	0.0018	0.0222	0.0592	0.6030	0.4771	0.6024
(50, 30, 30)	0.9812	0.9678	0.9468	0.0156	0.0146	0.0120	0.0032	0.0176	0.0412	0.3951	0.3608	0.3217
(50, 50, 50)	0.9800	0.9620	0.9478	0.0126	0.0214	0.0142	0.0074	0.0166	0.0380	0.3179	0.2188	0.2524
$(\mu_1, \sigma_1) = (0, 1)'$, $(\mu_2, \sigma_2) = (2.5, 1.1)'$, $(\mu_3, \sigma_3) = (5.51, 1.2)'$, $P_2 = 0.9$												
	Coverage Probability			Lower Tail			Upper Tail			Length of CI		

Three Independent Normal Distributions												
Sample Sizes	BTP	BTII	GI	BTP	BTII	GI	BTP	BTII	GI	BTP	BTII	GI
(10, 10, 10)	0.9468	0.8928	0.9310	0.0530	0.0000	0.0038	0.0002	0.1072	0.0652	0.4583	0.3639	0.5250
(30, 30, 30)	0.9820	0.9704	0.9430	0.0150	0.0114	0.0106	0.0030	0.0182	0.0464	0.2772	0.2591	0.2399
(20, 10, 10)	0.9244	0.8740	0.9216	0.0752	0.0000	0.0056	0.0004	0.1260	0.0728	0.4277	0.3454	0.4711
(50, 30, 30)	0.9828	0.9704	0.9420	0.0144	0.0076	0.0110	0.0028	0.0220	0.0470	0.2626	0.2474	0.2262
(50, 50, 50)	0.9856	0.9558	0.9512	0.0118	0.0192	0.0098	0.0026	0.0250	0.0390	0.2160	0.1464	0.1755

BTP: The confidence interval based on bootstrap percentiles.

BTII: The confidence interval presented in equ (15).

GI: The generalized confidence interval for Box-Cox transferred data.

Low tail (upper tail): One-sided coverage errors, i.e. the proportion of runs in which the lower (or upper) limit of the confidence interval excluded the true P_2 at nominal level 0.025.

Length of CI: the average length of two-sided confidence intervals for P_2.

Table 3

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

Three Independent Beta Distributions												
$(\alpha_1, \beta_1) = (1, 6)'$, $(\alpha_2, \beta_2) = (6, 6)'$, $(\alpha_3, \beta_3) = (9.6, 6)'$, $\boldsymbol{P_2} = 0.5$												
	Coverage Probability			Lower Tail			Upper Tail			Length of CI		
Sample Sizes	BTP	BTII	GI	BTP	BTII	GI	BTP	BTII	GI	BTP	BTII	GI
(10, 10, 10)	0.9722	0.9428	0.9664	0.0146	0.0392	0.0076	0.0132	0.0180	0.0260	0.7917	0.6168	0.6997
(30, 30, 30)	0.9744	0.9608	0.9484	0.0134	0.0260	0.0332	0.0122	0.0132	0.0184	0.5228	0.4725	0.3852
(20, 10, 10)	0.9728	0.9510	0.9680	0.0154	0.0350	0.0148	0.0118	0.0140	0.0172	0.7770	0.6055	0.6396
(50, 30, 30)	0.9728	0.9582	0.9436	0.0170	0.0264	0.0432	0.0102	0.0154	0.0132	0.5079	0.4611	0.3703
(50, 50, 50)	0.9672	0.9608	0.9324	0.0202	0.0240	0.0546	0.0126	0.0152	0.0130	0.4077	0.3851	0.2907
$(\alpha_1, \beta_1) = (1, 6)'$, $(\alpha_2, \beta_2) = (6, 6)'$, $(\alpha_3, \beta_3) = (12.6, 6)'$, $\boldsymbol{P_2} = 0.7$												
	Coverage Probability			Lower Tail			Upper Tail			Length of CI		
Sample Sizes	BTP	BTII	GI	BTP	BTII	GI	BTP	BTII	GI	BTP	BTII	GI
(10, 10, 10)	0.9706	0.9554	0.9740	0.0154	0.0212	0.0038	0.0140	0.0234	0.0222	0.7172	0.5571	0.7024
(30, 30, 30)	0.9710	0.9654	0.9692	0.0148	0.0164	0.0158	0.0142	0.0182	0.0150	0.4589	0.4178	0.3366
(20, 10, 10)	0.9810	0.9602	0.9744	0.0108	0.0214	0.0062	0.0082	0.0184	0.0194	0.6924	0.5361	0.5904
(50, 30, 30)	0.9720	0.9630	0.9622	0.0134	0.0188	0.0278	0.0146	0.0182	0.0100	0.4422	0.4035	0.3156
(50, 50, 50)	0.9696	0.9662	0.9580	0.0148	0.0188	0.0316	0.0156	0.0150	0.0104	0.3531	0.3360	0.2489
$(\alpha_1, \beta_1) = (1, 6)'$, $(\alpha_2, \beta_2) = (6, 6)'$, $(\alpha_3, \beta_3) = (15.3, 6)'$, $\boldsymbol{P_2} = 0.8$												
	Coverage Probability			Lower Tail			Upper Tail			Upper Tail		
Sample Sizes	BTP	BTII	GI	BTP	BTII	GI	BTP	BTII	GI	BTP	BTII	GI
(10, 10, 10)	0.9672	0.9520	0.9708	0.0256	0.0170	0.0032	0.0072	0.0310	0.0260	0.6234	0.4873	0.6605
(30, 30, 30)	0.9788	0.9708	0.9732	0.0128	0.0112	0.0092	0.0084	0.0180	0.0176	0.3841	0.3520	0.2907
(20, 10, 10)	0.9648	0.9470	0.9714	0.0280	0.0202	0.0046	0.0072	0.0328	0.0240	0.5808	0.4568	0.5249
(50, 30, 30)	0.9726	0.9656	0.9756	0.0176	0.0152	0.0106	0.0098	0.0192	0.0138	0.3661	0.3370	0.2683
(50, 50, 50)	0.9736	0.9656	0.9760	0.0140	0.0132	0.0108	0.0124	0.0212	0.0132	0.2976	0.2812	0.2126
$(\alpha_1, \beta_1) = (1, 6)'$, $(\alpha_2, \beta_2) = (6, 6)'$, $(\alpha_3, \beta_3) = (20.4, 6)'$, $\boldsymbol{P_2} = 0.9$												

	Coverage Probability			Lower Tail			Upper Tail			Upper Tail		
Sample Sizes	BTP	BTII	GI	BTP	BTII	GI	BTP	BTII	GI	BTP	BTII	GI
(10, 10, 10)	0.9334	0.8802	0.9482	0.0574	0.0000	0.0028	0.0092	0.1198	0.0490	0.4763	0.3805	0.6004
(30, 30, 30)	0.9710	0.9734	0.9554	0.0164	0.0092	0.0046	0.0126	0.0174	0.0400	0.2853	0.2622	0.2336
(20, 10, 10)	0.9268	0.8714	0.9528	0.0658	0.0000	0.0048	0.0074	0.1286	0.0424	0.4287	0.3458	0.4404
(50, 30, 30)	0.9774	0.9648	0.9590	0.0140	0.0100	0.0058	0.0086	0.0252	0.0352	0.2611	0.2435	0.2104
(50, 50, 50)	0.9716	0.9740	0.9556	0.0154	0.0098	0.0054	0.0130	0.0162	0.0390	0.2163	0.2071	0.1662

BTP: The confidence interval based on bootstrap percentiles.

BTII: The confidence interval presented in equ (15).

GI: The generalized confidence interval for Box-Cox transferred data.

Low tail (upper tail): One-sided coverage errors, i.e. the proportion of runs in which the lower (or upper) limit of the confidence interval excluded the true P_2 at nominal level 0.025.

Length of CI: the average length of two-sided confidence intervals for P_2.

Table 4

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

Three Independent Gamma Distributions												
$(\alpha_1, \beta_1) = (1, 6)'$, $(\alpha_2, \beta_2) = (4, 6)'$, $(\alpha_3, \beta_3) = (6.2, 6)'$, $\boldsymbol{P_2} = 0.5$												
	Coverage Probability			Lower Tail			Upper Tail			Length of CI		
Sample Sizes	BTP	BTII	GI	BTP	BTII	GI	BTP	BTII	GI	BTP	BTII	GI
(10, 10, 10)	0.9674	0.9498	0.9628	0.0306	0.0358	0.0038	0.0020	0.0144	0.0334	0.8131	0.6409	0.7133
(30, 30, 30)	0.9762	0.9632	0.9628	0.0176	0.0216	0.0130	0.0062	0.0152	0.0242	0.5550	0.5021	0.4271
(20, 10, 10)	0.9656	0.9518	0.9644	0.0318	0.0350	0.0074	0.0026	0.0132	0.0282	0.7986	0.6246	0.6627
(50, 30, 30)	0.9696	0.9620	0.9586	0.0234	0.0254	0.0172	0.0070	0.0126	0.0242	0.5347	0.4876	0.3988
(50, 50, 50)	0.9734	0.9654	0.9548	0.0176	0.0206	0.0174	0.0090	0.0140	0.0278	0.4370	0.4138	0.3231
$(\alpha_1, \beta_1) = (1, 6)'$, $(\alpha_2, \beta_2) = (4, 6)'$, $(\alpha_3, \beta_3) = (7.7, 6)'$, $\boldsymbol{P_2} = 0.7$												
	Coverage Probability			Lower Tail			Upper Tail			Length of CI		
Sample Sizes	BTP	BTII	GI	BTP	BTII	GI	BTP	BTII	GI	BTP	BTII	GI
(10, 10, 10)	0.9634	0.9600	0.9578	0.0342	0.0228	0.0026	0.0024	0.0172	0.0396	0.7463	0.5823	0.7318
(30, 30, 30)	0.9726	0.9696	0.9638	0.0204	0.0136	0.0088	0.0070	0.0168	0.0274	0.4905	0.4476	0.3874
(20, 10, 10)	0.9640	0.9652	0.9536	0.0340	0.0184	0.0048	0.0020	0.0164	0.0416	0.7179	0.5565	0.6386
(50, 30, 30)	0.9712	0.9660	0.9594	0.0224	0.0186	0.0108	0.0064	0.0154	0.0298	0.4667	0.4216	0.3571
(50, 50, 50)	0.9714	0.9736	0.9594	0.0208	0.0148	0.0148	0.0078	0.0116	0.0258	0.3819	0.3622	0.2905
$(\alpha_1, \beta_1) = (1, 6)'$, $(\alpha_2, \beta_2) = (4, 6)'$, $(\alpha_3, \beta_3) = (9.0, 6)'$, $\boldsymbol{P_2} = 0.8$												
	Coverage Probability			Lower Tail			Upper Tail			Length of CI		
Sample Sizes	BTP	BTII	GI	BTP	BTII	GI	BTP	BTII	GI	BTP	BTII	GI
(10, 10, 10)	0.9680	0.9572	0.9570	0.0228	0.0190	0.0022	0.0092	0.0238	0.0408	0.6473	0.5053	0.6967
(30, 30, 30)	0.9776	0.9760	0.9582	0.0136	0.0104	0.0084	0.0088	0.0136	0.0334	0.4166	0.3777	0.3378
(20, 10, 10)	0.9742	0.9626	0.9490	0.0162	0.0136	0.0036	0.0096	0.0238	0.0474	0.6096	0.4795	0.5771
(50, 30, 30)	0.9740	0.9738	0.9508	0.0136	0.0116	0.0102	0.0124	0.0146	0.0390	0.3864	0.3529	0.3053
(50, 50, 50)	0.9746	0.9744	0.9554	0.0104	0.0122	0.0118	0.0150	0.0134	0.0328	0.3213	0.3055	0.2518
$(\alpha_1, \beta_1) = (1, 6)'$, $(\alpha_2, \beta_2) = (4, 6)'$, $(\alpha_3, \beta_3) = (12.1, 6)'$, $\boldsymbol{P_2} = 0.9$												

Three Independent Gamma Distributions												
	Coverage Probability			Lower Tail			Upper Tail			Length of CI		
Sample Sizes	BTP	BTII	GI	BTP	BTII	GI	BTP	BTII	GI	BTP	BTII	GI
(10, 10, 10)	0.9292	0.8888	0.9526	0.0626	0.0000	0.0056	0.0082	0.1112	0.0418	0.4958	0.3872	0.6012
(30, 30, 30)	0.9754	0.9704	0.9568	0.0146	0.0100	0.0130	0.0100	0.0196	0.0302	0.3072	0.2840	0.2696
(20, 10, 10)	0.9264	0.8924	0.9502	0.0670	0.0000	0.0072	0.0066	0.1076	0.0426	0.4332	0.3544	0.4592
(50, 30, 30)	0.9726	0.9678	0.9598	0.0138	0.0122	0.0142	0.0136	0.0200	0.0260	0.2740	0.2561	0.2305
(50, 50, 50)	0.9678	0.9706	0.9520	0.0206	0.0120	0.0184	0.0116	0.0174	0.0296	0.2401	0.2295	0.1973

BTP: The confidence interval based on bootstrap percentiles.
BTII: The confidence interval presented in equ (15).
GI: The generalized confidence interval for Box-Cox transferred data.
Low tail (upper tail): One-sided coverage errors, i.e. the proportion of runs in which the lower (or upper) limit of the confidence interval excluded the true P_2 at nominal level 0.025.
Length of CI: the average length of two-sided confidence intervals for P_2.

Table 5

Estimated confidence intervals for the probability of detecting early stage Alzheimer's disease using psychometric tests from WU ADRC (sensitivity to full disease and specificity are assumed to equal to 0.8).

Confidence Intervals for the test covariates											
	Non-parametric					Parametric					
		BTP		BTII		Normal GI			Box-cox GI		
Variables	$\hat{P}_2^{NP}$	**lb**	**ub**	**lb**	**ub**	$\hat{P}_2^{P}$	**lb**	**ub**	$\hat{P}_2^{BoxP}$	**lb**	**ub**
Global factor	**0.7727**	0.5718	0.8875	0.5788	0.9371	0.6107	0.3594	0.7591	0.6267	0.3533	0.7738
Frontal factor*	0.5718	0.1986	0.7440	0.1977	0.7842	**0.4340**	0.1997	0.6058	0.4375	0.2189	0.5987
Parietal factor**	0.3708	0.1412	0.6866	0.1132	0.6877	0.2290	0.0000	0.4834	**0.2499**	0.0000	0.4650
Temporal factor*	0.7440	0.5431	0.8875	0.5527	0.9343	**0.6330**	0.3926	0.7862	0.6337	0.4128	0.7952
Associate Learning**	0.6005	0.2123	0.7440	0.2401	0.7961	0.3383	0.0695	0.5338	**0.3436**	0.0854	0.5501
Logical Memory	**0.7153**	0.4856	0.8875	0.5109	0.9225	0.6749	0.4883	0.7985	0.6739	0.4984	0.8021
Digit Span Forward	**0.3708**	0.0551	0.7153	0.1206	0.7159	0.1939	0.0000	0.4161	0.2001	0.0000	0.4057
Digit Span Backward	**0.3421**	0.0838	0.8588	0.1139	0.8959	0.2050	0.0000	0.4697	0.2145	0.0000	0.4502
Visual Retention (10 s)	**0.7153**	0.2273	0.8301	0.3120	0.9332	0.3336	0.1198	0.4961	0.3466	0.1231	0.5112
Information	**0.4282**	0.1699	0.8014	0.2112	0.8062	0.4387	0.1821	0.6181	0.4378	0.1943	0.6105
Word Fluency**	0.2273	0.0551	0.5718	0.0000	0.5490	0.1643	0.0000	0.3785	**0.1799**	0.0000	0.3817
Boston Naming	**0.3421**	0.1412	0.6866	0.0723	0.7036	0.3270	0.0562	0.5002	0.3406	0.0810	0.5206
Mental Control	**0.4569**	0.0551	0.6579	0.0395	0.7649	0.2369	0.0153	0.4039	0.2375	0.0361	0.4135
Visual Retention (copy)	**0.4856**	0.0551	0.7877	0.0000	0.8868	0	0.0000	0.0905	0	0.0000	0.1039

BTP: The confidence interval based on bootstrap percentiles.

BTII: Confidence interval is computed by the BTII approach when normality for the data or the Box-Cox transformed data is not satisfied.

GI: The normality is satisfied. Confidence interval is computed by the generalized inference approach.

Box-Cox: The normality of Box-Cox transformed data is satisfied. Confidence interval is computed by the Box-Cox generalized inference approach.

$\hat{P}_2^{NP}$: The nonparametric estimation of P_2 in equ (9).

$\hat{P}_2^{P}$: The parametric estimation based on normality of P_2 in equ (3).

$\hat{P}_2^{BoxP}$: The parametric estimation based on normality after Box-Cox transformation of P_2.

*: Normality is satisfied. The generalized confidence interval is preferred.

** : Normality is satisfied for Box-Cox transformed data. The generalized confidence interval for transformed data is preferred.

Note: The most appropriate point estimates are highlighted.