Received: 1 June 2020 | Revised: 29 April 2021 | Accepted: 1 May 2021

DOI: 10.1002/pst.2131

WILEY

MAIN PAPER

Joint confidence region estimation on predictive values

Braydon J. Schaible | **Jingjing Yin**

Department of Biostatistics, Epidemiology and Environmental Health Sciences, Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, Georgia, USA

Correspondence
Jingjing Yin, Department of Biostatistics, Epidemiology and Environmental Health Sciences, Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, GA 30458, USA.
Email: jyin@georgiasouthern.edu

Abstract

For evaluating diagnostic accuracy of inherently continuous diagnostic tests/biomarkers, sensitivity and specificity are well-known measures both of which depend on a diagnostic cut-off, which is usually estimated. Sensitivity (specificity) is the conditional probability of testing positive (negative) given the true disease status. However, a more relevant question is "what is the probability of having (not having) a disease if a test is positive (negative)?". Such post-test probabilities are denoted as positive predictive value (PPV) and negative predictive value (NPV). The PPV and NPV at the same estimated cut-off are correlated, hence it is desirable to make the joint inference on PPV and NPV to account for such correlation. Existing inference methods for PPV and NPV focus on the individual confidence intervals and they were developed under binomial distribution assuming binary instead of continuous test results. Several approaches are proposed to estimate the joint confidence region as well as the individual confidence intervals of PPV and NPV. Simulation results indicate the proposed approaches perform well with satisfactory coverage probabilities for normal and non-normal data and, additionally, outperform existing methods with improved coverage as well as narrower confidence intervals for PPV and NPV. The Alzheimer's Disease Neuroimaging Initiative (ADNI) data set is used to illustrate the proposed approaches and compare them with the existing methods.

KEYWORDS

Box-Cox transformation, joint confidence region, predictive value, Youden index

1 | INTRODUCTION

In medical diagnostics, sensitivity and specificity are two frequently used measures for evaluating diagnostic performance by both researchers and practitioners. The sensitivity is the probability of a diseased subject being diagnosed as diseased while the specificity is the probability of a healthy subject being diagnosed as non-diseased. Therefore, sensitivity and specificity are the conditional probabilities of testing results given the true disease status. However, in practice, without performing the gold standard test (which often is more costly and/or involves more risky procedures), clinicians and patients do not know the true disease status. Therefore, a more relevant question for clinicians and patients is "what is the probability of having (not having) a disease under a positive (negative) test result?".[1] Such posterior probabilities are denoted as positive predictive value (PPV) and negative predictive value (NPV). Based on Bayes theorem, we can derive the equations of the posterior probabilities, PPV and NPV, as functions of sensitivity, specificity and prevalence rate of the disease. Unlike sensitivity and specificity, which is an invariant measure for diagnostic tests, PPV and

NPV depend on the prevalence of the disease and change with respect to different disease prevalence rates from different target populations.

Not much research has focused on the statistical methods for the inference based on PPV and/or NPV, and currently existing methods were developed under a binary-scale diagnostic test. For example, four confidence intervals (standard, standard adjusted, logit, logit-adjusted) were proposed based on binomial distribution properties for predictive values.[2] Additionally, Bayesian alternative methods[3] were proposed. However, for a diagnostic test with naturally continuous measurements, the value of the diagnostic cut-off c is needed to dichotomize the continuous measurements into binary (positive or negative) test results. If we apply the existing binomial confidence intervals, the test results must be binary or the cut-off for the continuous test is known and assumed to be fixed. For situations where the cut-off is unknown and needs to be estimated, the Youden index[4] is a widely used optimization criteria to find c in practice.[5–7] The corresponding optimal cut-off point gives the maximum of the sum of sensitivity and specificity. After the cut-off point based on Youden index is estimated, sensitivity and specificity can be calculated. Then, for any given disease prevalence, the values of PPV and NPV can be estimated.

Since the optimal cut-off point is estimated and PPV and NPV are both functions of the estimated sensitivity and specificity at the same cut-off point, PPV and NPV estimates are potentially correlated. If we want to make inferences about both PPV and NPV, in order to account for such potential correlation between them, we may consider a joint confidence region. The joint confidence region of the potentially correlated PPV and NPV can give more comprehensive information compared with the individual confidence intervals that are more commonly used in practice. Section 2 in the following text provides more details about its benefits in application. Past literature exists for about joint inference methods in *ROC* analysis. For example, there is a publication about the joint confidence region of the *AUC* and Youden index,[8] the joint confidence region of the sensitivity and specificity,[9] the joint confidence region based on empirical likelihood method for any pair of (sensitivity, specificity, cut-off point) given the third value is fixed,[10] and the joint confidence region of the sensitivity and specificity at the estimated optimal cut-off point considering the variability of Box-Cox transformation parameter.[11] Previous joint inference research were about the pre-test diagnostic accuracy measures, which values are obtained conditioning on the true disease status of each observation. In this research, the accuracy measures of interest are the predictive values, which are post-test measures that give user some idea about the diagnostic accuracy conditioning on the test results.

The rest of the paper is organized as follows. Section 2 gives the motivation of the research and discusses the advantage of the proposed methods over the two existing methods which were developed under binary-scale test settings.[2,3] Section 3 presents the proposed methods and has five subsections. Notations and preliminaries about the binormal model in an Receiver operating characteristic (ROC) setting are introduced in Section 3.1. In Sections 3.2 and 3.3, confidence regions of PPV and NPV based on generalized inference and parametric bootstrap under normality are discussed, respectively. In Section 3.4, the application of Box-Cox transformation for the estimation of joint confidence region under non-normal data is presented. Finally, monotonic transformation and inverse transformation of the proposed confidence region are presented in Section 3.5. Section 4 contains simulation results comparing the proposed methods as well as for the two existing methods of constructing the individual confidence intervals.[2,3] In Section 5, a real-world data example from the Alzheimer's Disease Neuroimaging Initiative (ADNI) is analyzed to illustrate the proposed confidence regions. Section 6 provides conclusions and extends the proposed setting for applying different weights or incorporating the variability of sensitivity and specificity due to other considerations. Furthermore, possible applications of the proposed methods beyond medical diagnostics are discussed.

2 | MOTIVATION

We introduced two existing methods[2,3] for estimating the confidence intervals of PPV and NPV previously, and both methods assume the test measurements are dichotomous (positive or negative) and use the properties of the binomial distribution to construct the confidence intervals. Before applying the two existing methods to calculate the binomial confidence intervals, for a diagnostic test with naturally continuous measurements, the value of the diagnostic cut-off c is needed. Given the two existing methods are constructed under binomial distribution, the test results must be binary and the cut-off point must be pre-known thus is considered as fixed. Therefore, the existing methods cannot account for the variability of the estimated cut-off point for a naturally continuous test/biomarker. This might cause some issues and in this research we will compare the two existing methods with our proposed methods by simulations under continuous biomarker settings.

Additionally, if we want to obtain the confidence intervals about both predictive values simultaneously, we typically need to apply a multiple testing/comparison adjustment in order to maintain the nominal confidence level for both PPV and NPV intervals together. The Bonferroni method is the most straightforward approach for multiple comparison and it is commonly applied. However, such multiple testing adjustment is conservative as it assumes independence of the two predictive value estimates, which is not a valid assumption as we have discussed earlier that the PPV and NPV are potentially correlated. In order to account for such correlation, we propose to estimate the joint confidence region of the predictive values. The proposed joint confidence region of PPV and NPV at the estimated cut-off point for a continuous biomarker defines an elliptical area around the point estimates of PPV and NPV. The elliptical region is expected to cover the true values of the PPV and NPV simultaneously with $100(1-\alpha)\%$ confidence. Hence, the corresponding simultaneous confidence intervals that are projected from the joint confidence region on either PPV and NPV domains, would give better coverage compared with the Bonferroni approach which assumes independence between the PPV and NPV. In order to compare the proposed with the Bonferroni-adjusted confidence intervals, additional illustrations are provided by the data example in Section 5.

In this research, we follow three steps to estimate the joint confidence region of the PPV and NPV for a candidate continuous biomarker with the need of estimating the diagnostic cut-off. Firstly, we estimate the optimal cut-off point, that is, $\hat{c}_o$, based on the Youden index criteria. Note that other cut-off selection methods such as Euclidean index, product of sensitivity and specificity and maximum absolute determinant[12,13] can be applied similarly in the proposed framework. We chose Youden index for illustration as it is the most well-known and straightforward method.[5–7] Secondly, we calculate the pre-test accuracy probabilities, that is, the sensitivity and specificity at the estimated optimal cut-off point associated with the Youden index and then the post-test accuracy probabilities PPV and NPV are derived using the Bayes equations. Note since sensitivity and specificity are obtained at the same estimated cut-off, and they are correlated, so are the predictive values. Finally, we derive the individual confidence intervals as well as the joint confidence region of the PPV and NPV. Note the joint confidence region can as well give the alternative confidence intervals that simultaneously maintain the type I error when making inference about both predictive values.

This paper has two highlights: (1) the proposed inference approach of PPV and NPV is developed under a continuous test setting. Generally, in the literature, the existing inference approaches of PPV and NPV are centered around binary-scale test results such as qualitative ratings by clinical evaluations. However, for most clinical practices, it is more common to have continuous test measurements for many screening and diagnostic tests; And our simulation results suggest that the proposed method outperforms the existing methods for continuous test settings; (2) the proposed estimation is based on the joint confidence region, which accounts for the potential correlation between PPV and NPV. In the existing diagnostic literature, prior research proposes methods for constructing individual confidence intervals, which consider PPV and NPV separately. Even after considering multiple testing adjustment, the results will be too conservative since the correlation between predictive values are not accounted for. The proposed joint confidence region estimation allows clinicians to observe the ranges of the post-test predictive accuracy measures simultaneously and comprehensively, with the correlation between the predictive values in mind, and hence better decisions about the diagnostic test can be made.

3 | METHODS

3.1 | Binormal model for ROC summary statistics

Let Y_1 and Y_2 denote the marker measurements for diseased and healthy subjects, respectively, and $F_{Y_1}(.)$ and $F_{Y_2}(.)$ denote the corresponding cumulative distribution functions (cdfs). Note that Y_1 and Y_2 are independent. Henceforth, let $\boldsymbol{\eta} = (Se, Sp)^T$ denote the vector of true values of sensitivity (Se) and specificity (Sp), and c denotes any given/known cut-off point. Assuming a higher marker measurement is associated with larger likelihood of having the disease, so if an individual has a marker measurement greater than or equal to c, he/she will be classified as disease positive, and if the marker measurement is less than c, disease negative. Sensitivity (Se) and specificity (Sp) at cut-off point (c) are

$$Se(c) = 1 - F_{Y_1}(c) \text{ and } Sp(c) = F_{Y_2}(c).$$

The optimal cut-off point c_o determined by Youden index is estimated as follows

$$c_o = \{c : max_c(Se(c) + Sp(c) - 1) = max_c(F_{Y_2}(c) - F_{Y_1}(c))\}.$$

By Bayes Theorem,

$$PPV(c) = \frac{Se(c)*P_d}{Se(c)*P_d + (1 - Sp(c))*(1 - P_d)} \tag{1}$$

and

$$NPV(c) = \frac{Sp(c)*(1 - P_d)}{(1 - Se(c))*P_d + Sp(c)*(1 - P_d)} \tag{2}$$

where P_d is the prevalence of disease.

From the PPV and NPV equations, we can conclude that the lower the prevalence, the higher the NPV, while the lower the PPV. Therefore, in order to estimate the PPV and NPV, we need to obtain a value of disease prevalence from the targeting population. Note only cohort studies can actually provide valid estimates of the prevalence of disease from the observed data. However, for diagnostic test evaluations, especially at early stages, the case–control studies are generally used, where the disease status of the patients are known during recruiting and samples are collected separately to form a case group from the diseased population and a control group from the non-diseased population. One major reason is that oversampling of disease subjects over healthy controls is necessary in order to have enough samples from the case group for the evaluation of the corresponding diagnostic test. This is especially true if rare disease is of interest. For case–control diagnostic test studies, the prevalence is generally pre-specified or estimated from external sources such as literature from past cohort studies, systematic reviews or meta analyses that pool results from multiple studies. Henceforth, in this research, the disease prevalence is assumed to be fixed and is pre-specified for the simulation study and data analysis.

Under binormal setting, that is, the diseased and healthy populations are both normally distributed as $Y_1 \sim Normal(\mu_1, \sigma_1^2)$ and $Y_2 \sim Normal(\mu_2, \sigma_2^2)$, sensitivity and specificity at the cut-off point c are expressed as

$$Se(c) = \Phi\left(\frac{\mu_1 - c}{\sigma_1}\right) \text{ and } Sp(c) = \Phi\left(\frac{c - \mu_2}{\sigma_2}\right), \tag{3}$$

where $\Phi(.)$ denotes the standard normal cumulative distribution function. The optimal cut-off point c_o based on the Youden index can be obtained analytically as follows.[14] When variances of diseased and healthy samples are not equal,

$$c_o = \frac{\mu_2(b^2 - 1) - a + b\sqrt{a^2 + (b^2 - 1)\sigma_2^2 \ln(b^2)}}{b^2 - 1}, \tag{4}$$

where $a = \mu_1 - \mu_2$ and $b = \frac{\sigma_1}{\sigma_2}$; when variances are equal, that is, $b = 1$,

$$c_o = \frac{\mu_1 + \mu_2}{2}.$$

The estimates of sensitivity and specificity at the optimal cut-off point can be obtained by substituting the estimates of μ_is and σ_i^2s ($i = 1, 2$) in (4) and then the estimates of the optimal cut-off in (3).

3.2 | The generalized inference approach (*GPQ*)

A pivot (Q), or a pivotal quantity is a random variable whose distribution does not depend on any of the distribution parameters. That is, for a random sample $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)'$ from a distribution with parameter set $\boldsymbol{\theta}$, that is, $\boldsymbol{Y} \sim F(\boldsymbol{y}|\boldsymbol{\theta})$ where $\boldsymbol{y}$ is the observed data of $\boldsymbol{Y}$, then the pivot $Q(\boldsymbol{Y}, \boldsymbol{\theta})$ has the same distribution for all values of $\boldsymbol{\theta}$. A common

example of pivot is the t-test statistics $T = (\bar{Y} - \mu)/(S/\sqrt{n})$, where $\bar{Y}$ and S are the sample mean and standard deviation for a random sample $\boldsymbol{Y}$ of size n generated from a normal distribution $N(\mu, \sigma^2)$. And T does not dependent on μ nor σ^2. More details about pivots please refer to the sect. 9.2 in the "Statistical Inference" book by Casella and Berger.[15]

Similarly, suppose that $\boldsymbol{Y} \sim F(\boldsymbol{y}|\, \psi, \boldsymbol{\nu})$ where ψ is the parameter of interest for estimation and $\boldsymbol{\nu}$ is a vector of nuisance parameters. A generalized pivotal quantity (***GPQ***) $R(\boldsymbol{Y}; \boldsymbol{Y}, \psi, \boldsymbol{\nu})$ has two properties[16]: (1) $R(\boldsymbol{Y}; \boldsymbol{Y}, \psi, \boldsymbol{\nu})$ has a distribution independent of parameters; (2) The value of $R(\boldsymbol{Y}; \boldsymbol{Y}, \psi, \boldsymbol{\nu})$ can estimate ψ. So we can use the second property of GPQs to estimate ψ and construct the confidence interval for ψ. The concepts of generalized confidence interval 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, especially for small samples.[17–21]

The ***GPQ*** for normal variances and means are well-known[19] as

$$R_{\sigma_i^2} = \frac{(n_i - 1)s_i^2}{V_i} \text{ and } R_{\mu_i} = \bar{y}_i - Z_i\sqrt{R_{\sigma_i^2}/n_i},$$

where $V_i = \left((n_i - 1)S_i^2/\sigma_i^2\right) \sim \chi^2_{n_i - 1}$ and $Z_i = \left(\sqrt{n_i}(\bar{Y}_i - \mu_i)/\sigma_i\right) \sim N(0,1)$ ($i = 1, 2$ for diseased and healthy groups, respectively). The generalized pivots for sensitivity (R_{Se}) and specificity (R_{Sp}) are

$$\begin{aligned} R_{Se} &= \Phi\left(\frac{R_{\mu_1} - R_{c_o}}{R_{\sigma_1}}\right) \text{ and} \\ R_{Sp} &= \Phi\left(\frac{R_{c_o} - R_{\mu_2}}{R_{\sigma_2}}\right), \end{aligned} \tag{5}$$

where R_{c_o} is the generalized pivot for the optimal cut-off point c_o. R_{c_o} can be calculated as

$$R_{c_o} = \frac{R_{\mu_2}\left(R_b^2 - 1\right) - R_a + R_b\sqrt{R_a^2 + \left(R_b^2 - 1\right)R_{\sigma_2^2}\ln\left(R_b^2\right)}}{R_b^2 - 1}, \tag{6}$$

where $R_a = R_{\mu_1} - R_{\mu_2}$, $R_b = (R_{\sigma_1}/R_{\sigma_2})$, and $R_{\sigma_i} = \sqrt{R_{\sigma_i^2}}$ for $i = 1, 2$ under heterogeneity, and

$$R_{c_o} = \frac{R_{\mu_1} + R_{\mu_2}}{2}$$

under homogeneity. Thus, R_{PPV} and R_{NPV} can be calculated from the estimates of sensitivity R_{Se} and specificity R_{Sp} as follows:

$$\begin{aligned} R_{PPV} &= \frac{R_{Se}p_d}{R_{Se}p_d + \left(1 - R_{Sp}\right)(1 - p_d)} \\ R_{NPV} &= \frac{R_{Sp}(1 - p_d)}{\left(1 - R_{Sp}\right)p_d + R_{Sp}(1 - p_d)} \end{aligned} \tag{7}$$

where p_d is the disease prevalence, which is pre-specified and assumed as a constant across all pivots.

3.2.1 | Computing Algorithm

1. Generate $V_i \sim \chi^2_{n_i - 1}$ for calculating $R_{\sigma_i^2}$. Generate $Z_i \sim N(0, 1)$ for calculating R_{μ_i}.
2. Calculate the GPQs of sensitivity and specificity $\boldsymbol{R_\eta} = (R_{Se}, R_{Sp})^T$ following (5). Calculate $\boldsymbol{R_\nu} = (R_{PPV}, R_{NPV})^T$ following (7).

3. Repeat above steps for $B = 2500$ times to obtain $\boldsymbol{R}_{\boldsymbol{\nu}}^b = \left(R_{PPV}^b, R_{NPV}^b\right)^T; b = 1, ..., B$. Note the variability of cut-off point estimation is accounted by resampling the GPQs. Iteration number for GPQ is set at 2500 as indicated by previous literature.[8,19,22]
4. Denote the 100($\alpha/2$)th and 100($1-\alpha/2$)th percentiles of R_{PPV}^b $(b = 1, ..., B)$ as $R_{PPV}(\alpha/2)$ and $R_{PPV}(1-\alpha/2)$. Denote the 100($\alpha/2$)th and 100($1-\alpha/2$)th percentiles of R_{NPV}^b $(b = 1, ..., B)$ as $R_{NPV}(\alpha/2)$ and $R_{NPV}(1-\alpha/2)$.
5. Calculate $\hat{\boldsymbol{\nu}}_{GPQ} = \frac{1}{B}\sum_{b=1}^{B} \boldsymbol{R}_{\boldsymbol{\nu}}^b$ and $\hat{\boldsymbol{\Sigma}}_{GPQ} = \frac{1}{B-1}\sum_{b=1}^{B}\left(\boldsymbol{R}_{\boldsymbol{\nu}}^b - \hat{\boldsymbol{\nu}}_{GPQ}\right)\left(\boldsymbol{R}_{\boldsymbol{\nu}}^b - \hat{\boldsymbol{\nu}}_{GPQ}\right)^T$. Compute the standardized version of $\boldsymbol{R}_{\boldsymbol{\nu}}^b$, that is, $\tilde{\boldsymbol{R}}_{\boldsymbol{\nu}}^b = \hat{\boldsymbol{\Sigma}}_{GPQ}^{-1/2}\left(\boldsymbol{R}_{\boldsymbol{\nu}}^b - \hat{\boldsymbol{\nu}}_{GPQ}\right)$, and its length/norm as $\| \tilde{\boldsymbol{R}}_{\boldsymbol{\nu}}^b \| = \sqrt{\left(\tilde{\boldsymbol{R}}_{\boldsymbol{\nu}}^b\right)^T \tilde{\boldsymbol{R}}_{\boldsymbol{\nu}}^b}$ for $b = 1, ..., B$. Denote the 100 $(1-\alpha)$th percentile of the set $\| \tilde{\boldsymbol{R}}_{\boldsymbol{\nu}} \|$ as $q_{\left\{\| \tilde{\boldsymbol{R}}_{\boldsymbol{\nu}} \|; 1-\alpha\right\}}$.

The 100$(1-\alpha)$% generalized (referred as ***GPQ*** hereafter) confidence region of $\boldsymbol{\nu} = (PPV, NPV)^T$ is

$$\left\{\boldsymbol{\nu} : (\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}_{GPQ})^T \hat{\boldsymbol{\Sigma}}_{GPQ}^{-1} (\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}_{GPQ}) \le q^2_{\left\{\| \tilde{\boldsymbol{R}}_{\boldsymbol{\nu}} \|; 1-\alpha\right\}}\right\},$$

where $\hat{\boldsymbol{\nu}}_{GPQ}$, $\hat{\boldsymbol{\Sigma}}_{GPQ}$ and $q_{\left\{\| \tilde{\boldsymbol{R}}_{\boldsymbol{\nu}} \|; 1-\alpha\right\}}$ are the values obtained in step 5 of the *Computing Algorithm*. The area of the generalized confidence region is estimated by $A_{GPQ} = \pi\left(q^2_{\left\{\| \tilde{\boldsymbol{R}}_{\boldsymbol{\nu}} \|; 1-\alpha\right\}}\right)\sqrt{|\hat{\boldsymbol{\Sigma}}_{GPQ}|}$ where $|\hat{\boldsymbol{\Sigma}}_{GPQ}|$ is the determinant of $\hat{\boldsymbol{\Sigma}}_{GPQ}$.

The corresponding simultaneous confidence intervals can be obtained as a by-product of the joint confidence region. Such confidence interval adjusts for multiple testing and maintains the type I error rate for estimating PPV and NPV together. With more general notations, with joint confidence region of $\boldsymbol{\nu} = (PPV, NPV)^T$ being $\left\{\boldsymbol{\nu} : (\boldsymbol{\nu} - \hat{\boldsymbol{\nu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}) \le q^2_{1-\alpha}\right\}$ where $\hat{\boldsymbol{\nu}} = \left(\hat{PPV}, \hat{NPV}\right)^T$ and $q^2_{1-\alpha}$ is the critical value of the confidence region (e.g., $q^2_{\left\{\| \tilde{\boldsymbol{R}}_{\boldsymbol{\nu}} \|; 1-\alpha\right\}}$ for ***GPQ*** method, etc.), the ***GPQ*** simultaneous confidence intervals for *PPV* and *NPV* can be estimated as follows:

$$\begin{aligned} &\hat{PPV} \pm q_{1-\alpha}\sqrt{\hat{\boldsymbol{\Sigma}}_{1,1}} \text{ and} \\ &\hat{NPV} \pm q_{1-\alpha}\sqrt{\hat{\boldsymbol{\Sigma}}_{2,2}}, \end{aligned} \tag{8}$$

where $q_{1-\alpha}$ is the square-root of $q^2_{1-\alpha}$. Note that the simultaneous confidence interval for either *PPV* or *NPV* is the projection of the confidence ellipse of $\boldsymbol{\eta}$ on the corresponding axis. Another more common way to adjust for multiple testing is the Bonferroni approach which was constructed similarly as the following individual confidence intervals, but at a type I error rate setting at the nominal level dividing by the number of tests (=k), that is, the critical value is $z_{1-\alpha/2k}$. However, the Bonferroni method is a well-known conservative approach, especially if the correlations between tests are large.

When only one measure, either PPV or NPV, is of interest, the individual confidence interval is more appropriate. The 100$(1-\alpha)$% individual confidence interval of *PPV* can be obtained based on the percentiles of the generalized pivots as the following

$$(R_{PPV}(\alpha/2),\ R_{PPV}(1-\alpha/2),)$$

where $R_{PPV}(\alpha/2)$ and $R_{PPV}(1-\alpha/2)$ are obtained in step 4 of the *Computing Algorithm*. Additionally, we can apply the standard normal quantile ($z_{1-\alpha/2}$) approach assuming R_{PPV} being asymptotically normal, as

$$\bar{R}_{PPV} \pm z_{1-\alpha/2}\sqrt{Var(R_{PPV})}$$

where $\bar{R}_{PPV}$ and $Var(R_{PPV})$ are the sample mean and sample variance of the simulated values R_{PPV} obtained in step 3 of the *Computing Algorithm*. Similar computations can be used to calculate the confidence interval for NPV using the ***GPQ*** method.

3.3 | Parametric bootstrap method (*PBoot*)

In Section 3.2, the GPQ method obtains resamples (pivots) from simulating the GPQs for normal means and variances, separately, and then obtains the pivots for other statistics that are functions of the normal means and variances. Since binormality is assumed, we can apply the parametric bootstrap method which resamples data directly from normal distributions with parameters set to be the sample means and sample variances of the observed data. The main difference between the GPQ method and the parametric bootstrap method is that the GPQ method resamples the pivotal estimates of normal means and variances directly from simulating the GPQs, while the parametric bootstrap method simulates biomarker values from normal distributions for the diseased and non-diseased groups and then calculates the maximum likelihood estimates (MLEs) of normal means and variances for the resampled data. Denote the sample means and variances as $\widehat{\mu}_i$ and $\widehat{\sigma}_i^2$ $(i = 1, 2)$, which are estimated from the observed biomarker measurements from the disease and healthy groups. We can obtain the confidence intervals and region using the parametric bootstrap method as follows:

3.3.1 | Computing Algorithm

1. Generate normally-distributed bootstrap samples $Y_1^b \sim N\left(\widehat{\mu}_1, \widehat{\sigma}_1^2\right)$ and $Y_2^b \sim N\left(\widehat{\mu}_2, \widehat{\sigma}_2^2,\right)$. Then obtain sample means and variances from the generated bootstrap samples Y_1^b and Y_2^b.
2. By replacing the population parameters μ and σ^2 with respective sample means and variances of the bootstrap samples. Calculate $\widehat{c}_0$ following (4). Calculate $\widehat{\boldsymbol{\eta}} = \left(\widehat{Se}, \widehat{Sp}\right)^T$ following (3). Calculate $\widehat{\boldsymbol{\nu}} = \left(\widehat{PPV}, \widehat{NPV}\right)^T$ following (1) and (2).
3. Repeat above steps for $B = 500$ times to obtain $\widehat{\boldsymbol{\nu}}^b = \left(\widehat{PPV}^b, \widehat{NPV}^b\right)^T; b = 1, ..., B$. Note the variability of cut-off point estimation is accounted by resampling the bootstrap samples. Bootstrap number is set at 500 as recommended by the previous literature that boostrap number exceeding 399 gives good performance if significance of the test is set at 0.05[23,24] and we set it slightly more than required to guarantee the desired performance from bootstrap.
4. Calculate $\widehat{\boldsymbol{\nu}}_{PBoot} = (1/B)\sum_{b=1}^{B} \widehat{\boldsymbol{\nu}}^b$ and $\widehat{\boldsymbol{\Sigma}}_{PBoot} = (1/B-1)\sum_{b=1}^{B} \left(\widehat{\boldsymbol{\nu}}^b - \widehat{\boldsymbol{\nu}}_{PBoot}\right)\left(\widehat{\boldsymbol{\nu}}^b - \widehat{\boldsymbol{\nu}}_{PBoot}\right)^T$. Compute the standardized version of $\widehat{\boldsymbol{\nu}}^b$, that is, $\tilde{\widehat{\boldsymbol{\nu}}}^b = \widehat{\boldsymbol{\Sigma}}_{PBoot}^{-1/2}\left(\widehat{\boldsymbol{\nu}}^b - \widehat{\boldsymbol{\nu}}_{PBoot}\right)$, and its length/norm as $\| \tilde{\widehat{\boldsymbol{\nu}}}^b \| = \sqrt{\left(\tilde{\widehat{\boldsymbol{\nu}}}^b\right)^T \tilde{\widehat{\boldsymbol{\nu}}}^b}$ for $b = 1, ..., B$. Denote the $100(1-\alpha)$th percentile of the set $\| \tilde{\widehat{\boldsymbol{\nu}}} \|$ as $q_{\left\{\| \tilde{\widehat{\boldsymbol{\nu}}} \|; 1-\alpha\right\}}$.

Similar to ***GPQ***, the $100(1-\alpha)\%$ parametric bootstrap (referred as ***PBoot*** hereafter) confidence region of $\boldsymbol{\nu} = (PPV, NPV)^T$ is

$$\left\{ \boldsymbol{\nu} : (\boldsymbol{\nu} - \widehat{\boldsymbol{\nu}}_{PBoot})^T \widehat{\boldsymbol{\Sigma}}_{PBoot}^{-1} (\boldsymbol{\nu} - \widehat{\boldsymbol{\nu}}_{PBoot}) \leq q^2_{\left\{\| \tilde{\widehat{\boldsymbol{\nu}}} \|; 1-\alpha\right\}} \right\},$$

and the corresponding area of the confidence region is $\pi\left(q^2_{\left\{\| \tilde{\widehat{\boldsymbol{\nu}}} \|; 1-\alpha\right\}}\right)\sqrt{|\widehat{\boldsymbol{\Sigma}}_{PBoot}|}$. The simultaneous and individual confidence intervals are obtained similarly as in Section 3.2.

3.4 | Without normality

When normality is not satisfied, it is a standard practice to use the Box-Cox transformation on the diagnostic test/biomarker measurements to achieve normality in both disease and non-disease groups due to the fact that the *ROC* curve is invariant under monotonic transformations.

For the jth ($j = 1, ..., n_i$) subject in the ith ($i = 1, 2$) group with each group having n_i observations, let

$$Y_{ij}^{(\lambda)} = \begin{cases} \dfrac{Y_{ij}^{\lambda} - 1}{\lambda} & \lambda \nleq 0, \\ log\left(Y_{ij}\right) & \lambda = 0 \end{cases}$$

where it is assumed that $Y_{ij}^{(\lambda)} \overset{\text{i.i.d}}{\sim} N(\mu_i, \sigma_i^2)$. Based on the observations from healthy and diseased groups, the log-likelihood function can be simplified as

$$\sum_{i}^{2}\sum_{j}^{n_i}\left[-\frac{1}{2}log(2\pi\sigma_i^2) - \frac{\left(Y_{ij}^{(\lambda)} - \mu_i\right)^2}{2\sigma_i^2} + (\lambda - 1)logY_{ij}\right]. \tag{9}$$

The MLE of λ can be obtained by maximizing the above function, denoted as $\widehat{\lambda}$.

For non-normal data, the Box-Cox transformation needs to be applied first to approximate normality and then the parametric inference methods based on normality can be applied to the transformed data. For example, to apply the ***PBoot*** method, firstly, we should Box-Cox transform the data from disease and healthy groups simultaneously to normal using the same λ values for both disease and healthy, and then obtain the sample mean and variances, $\widehat{\mu}_{1_{BoxCox}}, \widehat{\sigma^2}_{1_{BoxCox}}$, from the transformed data. Finally, by achieving normality approximately, we can then follow the *Computing Algorithm* in Sections 3.2 and 3.3 to estimate the joint confidence region as well as the individual confidence intervals. We denote the generalized pivotal method after Box-Cox transformation as ***GPQT*** and the parametric bootstrap method after Box-Cox transformation as ***PBootT***.

3.5 | Monotonic transformations of PPV and NPV

Note that since PPV and NPV are values in the range [0, 1], the performance of the proposed elliptical confidence regions, which assume the estimates are asymptotically normal, is not well maintained when the sample size is not large enough or the values of PPV and NPV are close to the boundary. Here we adopt the idea of monotonic transformation for quantities with restrictive range, such that the transformed quantity will approximate to normality faster, thus, the normal-based confidence intervals and regions are more precise. Common examples of monotonic transformations include the log transformation for odds ratios (OR) and the Fisher-Z transformation for correlation. For example, the confidence interval of $OR \in [0, 1]$, is usually calculated by exponentiating the confidence interval of the slope, that is, $\beta_1 = log(OR) \in [-\infty, \infty]$. Similarly, we can apply monotonic transformations, such as logit or arcsin-square-root transformations,[8] on probabilities (such as PPV and NPV), to improve the coverage of the confidence regions and intervals. Note the transformation discussed in this section refers to the transformation on the diagnostic accuracy measure/statistic, including the PPV and NPV, so that the transformed statistic will have an unbounded range and approximates to normal faster. The Box-Cox transformation in Section 3.4 refers to the transformation on the biomarker measurements so that the biomarker values are binormally distributed in both disease and non-disease groups.

Denote the transformation function as $h()$, hence transformed quantities as $\boldsymbol{\nu}^h = (h(PPV), h(NPV))$, and $\widehat{\boldsymbol{\nu}}^h$ and $\widehat{\boldsymbol{\Sigma}}^h$ denote the sample mean and the sample covariance matrix of $\boldsymbol{\nu}^h$. We apply the inverse transformation, $h^{-1}()$, to obtain the confidence region of the original values, $\boldsymbol{\nu} = (PPV, NPV)$, as

$$\left\{\boldsymbol{\nu} : \left[\boldsymbol{\nu} - h^{-1}\left(\widehat{\boldsymbol{\nu}}^h\right)\right]^T \left(\widehat{\boldsymbol{\Sigma}}^{inv}\right)^{-1} \left[\boldsymbol{\nu} - h^{-1}\left(\widehat{\boldsymbol{\nu}}^h\right)\right] \leq q_{1-\alpha}^2\right\} \tag{10}$$

where $\widehat{\boldsymbol{\Sigma}}^{inv} = \boldsymbol{J}_{inv}^T \widehat{\boldsymbol{\Sigma}}^h \boldsymbol{J}_{inv}$ is the inverse transformation of $\widehat{\boldsymbol{\Sigma}}^h$ and $\boldsymbol{J}_{inv}$. The Jacobian matrix $\boldsymbol{J}_{inv}$ is calculated by taking the first derivative of the inverse function $h^{-1}()$ evaluated at $\widehat{\boldsymbol{\nu}}^h$. Likewise, the individual confidence intervals of PPV (or NPV) can be obtained by inversely transforming the limits of confidence intervals (denote as lci and uci) of the transformed quantities accordingly as $[h^{-1}(lci), h^{-1}(uci)]$.

4 | SIMULATION STUDIES

Simulations were conducted under normal and gamma distributions (as an example of non-normal data) to assess the performance of the proposed confidence regions (i.e., the generalized inference (***GPQ*** and ***GPQT***) and the parametric bootstrap (***PBoot*** and ***PBootT***) approaches). Additionally, we also evaluated the coverage of individual confidence intervals for PPV and NPV and compared them with the existing binomial confidence intervals. In the process of

TABLE 1 Summary of approximate 95% joint confidence regions of ***GPQ*** method for (P_1, P_2) under normal distributions (based on 2000 simulations)

Sample Size	Estimate		Coverage			Width/area		
	PPV	NPV	PPV	NPV	CR	PPV	NPV	CR
$(\mu_1,\sigma_1^2)=(2.3205,3.9250)$, $(\mu_2,\sigma_2^2)=(0,1)$, $(\boldsymbol{P_1},\boldsymbol{P_2})=(0.4375,0.9643)$, $\mathbf{P_d}=0.1$								
(20, 30)	0.4499	0.9647	0.9520	0.9565	0.9505	0.3154	0.0381	0.0112
(30, 30)	0.4477	0.9647	0.9475	0.9560	0.9495	0.2852	0.0322	0.0085
(50, 40)	0.4454	0.9646	0.9445	0.9455	0.9395	0.2364	0.0255	0.0055
(50, 50)	0.4456	0.9647	0.9550	0.9460	0.9460	0.2215	0.0247	0.0050
(55, 65)	0.4419	0.9643	0.9580	0.9535	0.9515	0.2016	0.0234	0.0043
(75, 75)	0.4406	0.9645	0.9510	0.9485	0.9465	0.1806	0.0203	0.0034
(100,100)	0.4416	0.9645	0.9475	0.9500	0.9515	0.1568	0.0175	0.0025
$(\mu_1,\sigma_1^2)=(2.3205,3.9250)$, $(\mu_2,\sigma_2^2)=(0,1)$, $(\boldsymbol{P_1},\boldsymbol{P_2})=(0.8750,0.7500)$, $\mathbf{P_d}=0.5$								
(20, 30)	0.8764	0.7560	0.9575	0.9605	0.9570	0.1467	0.2019	0.0272
(30, 30)	0.8750	0.7553	0.9570	0.9515	0.9500	0.1332	0.1719	0.0209
(50, 40)	0.8749	0.7524	0.9440	0.9515	0.9465	0.1092	0.1373	0.0136
(50, 50)	0.8752	0.7529	0.9550	0.9575	0.9575	0.1013	0.1341	0.0123
(55, 65)	0.8751	0.7516	0.9535	0.9485	0.9460	0.0916	0.1256	0.0105
(75, 75)	0.8749	0.7519	0.9455	0.9455	0.9440	0.0820	0.1096	0.0082
(100,100)	0.8748	0.7514	0.9460	0.9495	0.9520	0.0708	0.0951	0.0061
$(\mu_1,\sigma_1^2)=(2.3205,3.9250)$, $(\mu_2,\sigma_2^2)=(0,1)$, $(\boldsymbol{P_1},\boldsymbol{P_2})=(0.9844,0.2500)$, $\mathbf{P_d}=0.9$								
(20, 30)	0.9842	0.2643	0.9475	0.9540	0.9390	0.0220	0.2273	0.0040
(30, 30)	0.9843	0.2616	0.9575	0.9545	0.9545	0.0197	0.1898	0.0031
(50, 40)	0.9845	0.2590	0.9420	0.9495	0.9480	0.0158	0.1476	0.0020
(50, 50)	0.9846	0.2585	0.9490	0.9430	0.9395	0.0145	0.1432	0.0018
(55, 65)	0.9843	0.2556	0.9430	0.9450	0.9400	0.0132	0.1324	0.0015
(75, 75)	0.9843	0.2540	0.9410	0.9465	0.9455	0.0118	0.1139	0.0012
(100,100)	0.9844	0.2535	0.9440	0.9505	0.9535	0.0100	0.0981	0.0009
$(\mu_1,\sigma_1^2)=(1.1712,0.2547)$, $(\mu_2,\sigma_2^2)=(0,1)$, $(\boldsymbol{P_1},\boldsymbol{P_2})=(0.2500,0.9844)$, $\mathbf{P_d}=0.1$								
(20, 30)	0.2609	0.9840	0.9570	0.9465	0.9395	0.2000	0.0235	0.0037
(30, 30)	0.2606	0.9844	0.9610	0.9525	0.9495	0.1883	0.0196	0.0030
(50, 40)	0.2563	0.9842	0.9520	0.9510	0.9510	0.1557	0.0156	0.0021
(50, 50)	0.2557	0.9842	0.9530	0.9570	0.9480	0.1414	0.0148	0.0018
(55, 65)	0.2552	0.9842	0.9505	0.9460	0.9460	0.1257	0.0135	0.0015
(75, 75)	0.2538	0.9844	0.9560	0.9455	0.9495	0.1135	0.0117	0.0012
(100,100)	0.2529	0.9843	0.9475	0.9515	0.9475	0.0979	0.0101	0.0009
$(\mu_1,\sigma_1^2)=(1.1712,0.2547)$, $(\mu_2,\sigma_2^2)=(0,1)$, $(\boldsymbol{P_1},\boldsymbol{P_2})=(0.7500,0.8750)$, $\mathbf{P_d}=0.5$								
(20, 30)	0.7538	0.8739	0.9545	0.9500	0.9485	0.1820	0.1551	0.0255
(30, 30)	0.7570	0.8754	0.9465	0.9545	0.9490	0.1722	0.1329	0.0208
(50, 40)	0.7530	0.8758	0.9455	0.9485	0.9400	0.1462	0.1063	0.0142
(50, 50)	0.7536	0.8754	0.9490	0.9410	0.9535	0.1342	0.1011	0.0123
(55, 65)	0.7513	0.8749	0.9555	0.9580	0.9545	0.1198	0.0933	0.0102
(75, 75)	0.7524	0.8759	0.9420	0.9465	0.9455	0.1097	0.0817	0.0081
(100,100)	0.7507	0.8749	0.9545	0.9475	0.9485	0.0952	0.0708	0.0061

(Continues)

TABLE 1 (Continued)

Sample Size	Estimate		Coverage			Width/area		
	PPV	NPV	PPV	NPV	CR	PPV	NPV	CR
$(\mu_1, \sigma_1^2) = (1.1712, 0.2547)$, $(\mu_2, \sigma_2^2) = (0,1)$, $(\boldsymbol{P_1}, \boldsymbol{P_2}) = (0.9643, 0.4375)$, $\mathbf{P_d} = 0.9$								
(20, 30)	0.9648	0.4509	0.9525	0.9435	0.9490	0.0340	0.3216	0.0102
(30, 30)	0.9646	0.4460	0.9520	0.9560	0.9470	0.0322	0.2851	0.0085
(50, 40)	0.9650	0.4471	0.9530	0.9520	0.9540	0.0271	0.2337	0.0058
(50, 50)	0.9646	0.4437	0.9590	0.9445	0.9520	0.0249	0.2214	0.0051
(55, 65)	0.9642	0.4409	0.9510	0.9460	0.9415	0.0223	0.2036	0.0042
(75, 75)	0.9647	0.4432	0.9505	0.9450	0.9390	0.0202	0.1811	0.0034
(100,100)	0.9643	0.4394	0.9525	0.9480	0.9465	0.0176	0.1567	0.0025

simulations, we found that the percentile-based ***GPQ*** confidence interval is generally more robust than its z-score-based ***GPQ*** counterpart. Thus, we report the results of the percentile-based confidence intervals. Additionally, we found that the logit transformation (as one common example of monotonic transformation in Section 3.5) greatly improves the coverage of the confidence regions as well as the corresponding confidence intervals. Therefore, in this section, we present the results of confidence regions and intervals utilizing the logit-transformation for all cases.

Sample sizes (n_1, n_2) were set as (20, 30), (30, 30), (50, 40), (50, 50), (55, 65), (75, 75), and (100,100). The PPV and NPV, (P_1, P_2), are calculated based on a set of values for sensitivity and specificity which were predetermined to be (0.70,0.90) and (0.90,0.70) to represent cases where sensitivity and specificity are far apart, and (0.80,0.90), (0.90,0.80), (0.85,0.95), and (0.95,0.85) to represent cases where the two quantities are closer. Prevalence rate is set at 0.1, 0.5, and 0.9 to cover a wide range of potential prevalence rates. A simulation consisting of 2000 runs was conducted under each setting of the combinations of sensitivity and specificity pairs, prevalence rates and the sample size pairs previously mentioned. Under 2000 simulation runs, we would expect the coverage probability of the confidence regions and intervals to fall between 0.94 and 0.96. Additionally, using the same simulation settings, we tested the two existing methods[2,3] for constructing the binomial confidence intervals using simulated continuous marker measurements. The simulation results for the two existing methods are presented in the Appendix in Tables A1-A4 as a supplementary document.

Tables 1 and 2 present the simulation results for the calculation of joint confidence regions for (P_1, P_2) at the nominal level of 95% under normal assumption for the ***GPQ*** and ***PBoot*** approaches, respectively. Both ***GPQ*** and ***PBoot*** provide satisfactory coverage close to the nominal level of 0.95, regardless of sample size, prevalence, or values of sensitivity and specificity, with ***PBoot*** being slightly worse with more liberal results (underestimated coverage). In addition, Figure 1 presents boxplots of the coverage probabilities under normal assumption for both ***GPQ*** and ***PBoot*** methods. With regards to the average area of the confidence region, the ***GPQ*** approach showed larger confidence region average areas than the ***PBoot*** confidence regions for all combinations of simulation settings.

Tables A1 and A2 present the simulation results for the two existing binomial confidence interval estimation methods, denoted as Mercaldo[2] and Stamey for the Bayesian version,[3] respectively, under binormal distributions. If the variability of the cut-off estimate is ignored, the binomial confidence intervals can be constructed directly from the binary test results assuming the cut-off is fixed. Such information loss from dichotomizing a continuous variable into binary are likely to increase the variability of the PPV and NPV estimates, thus, resulting in more conservative/wider individual confidence intervals. For both existing methods, coverage probabilities were overestimated and uniformly larger than the proposed confidence intervals, and the differences are more obvious at the boundaries, when the true PPV or NPV is high.

For the non-normal bi-gamma cases, ***GPQ*** with Box-Cox transformation (***GPQT***) approach performed poorly by the simulation results. Figure 2 presents the boxplots of the coverage probabilities of ***GPQT*** and ***PBootT*** under gamma distributions. For many settings, the coverage probabilities of the joint confidence region, that were obtained by the ***GPQT*** method, fall largely in the range of [0.86,0.89] and they are consistently lower than the nominal coverage probability of 0.95 across all settings. In addition, the ***GPQT***-based NPV, PPV confidence intervals are slightly liberal and the coverage probabilities centered around 0.92, which is lower than the nominal level of 0.95. In terms of average areas of the

TABLE 2 Summary of approximate 95% joint confidence regions of ***PBoot*** method for (P_1, P_2) under normal distributions (based on 2000 simulations)

Sample Size	Estimate		Coverage			Width/area		
	PPV	NPV	PPV	NPV	CR	PPV	NPV	CR
$(\mu_1,\sigma_1^2) = (2.3205, 1.9812)$, $(\mu_2,\sigma_2^2) = (0,1)$, $(\boldsymbol{P_1}, \boldsymbol{P_2}) = (0.4375, 0.9643)$, $\mathbf{P_d} = 0.1$								
(20, 30)	0.4699	0.9675	0.9525	0.9550	0.9520	0.3276	0.0407	0.0117
(30, 30)	0.4678	0.9664	0.9395	0.9440	0.9375	0.2963	0.0336	0.0089
(50, 40)	0.4592	0.9657	0.9370	0.9510	0.9435	0.2441	0.0262	0.0057
(50, 50)	0.4578	0.9658	0.9320	0.9445	0.9435	0.2275	0.0255	0.0052
(55, 65)	0.4517	0.9655	0.9375	0.9490	0.9450	0.2061	0.0240	0.0044
(75, 75)	0.4510	0.9653	0.9335	0.9515	0.9510	0.1849	0.0207	0.0034
(100,100)	0.4466	0.9650	0.9400	0.9585	0.9495	0.1593	0.0179	0.0026
$(\mu_1,\sigma_1^2) = (2.3205, 1.9812)$, $(\mu_2,\sigma_2^2) = (0,1)$, $(\boldsymbol{P_1}, \boldsymbol{P_2}) = (0.8750, 0.7500)$, $\mathbf{P_d} = 0.5$								
(20, 30)	0.8844	0.7706	0.9325	0.9525	0.9495	0.2171	0.1440	0.0278
(30, 30)	0.8857	0.7655	0.9380	0.9590	0.9445	0.1807	0.1282	0.0206
(50, 40)	0.8820	0.7592	0.9435	0.9560	0.9450	0.1419	0.1067	0.0135
(50, 50)	0.8810	0.7600	0.9340	0.9510	0.9440	0.1384	0.0996	0.0123
(55, 65)	0.8795	0.7575	0.9370	0.9515	0.9530	0.1298	0.0908	0.0106
(75, 75)	0.8789	0.7563	0.9380	0.9490	0.9485	0.1124	0.0813	0.0082
(100,100)	0.8775	0.7542	0.9435	0.9475	0.9405	0.0971	0.0705	0.0061
$(\mu_1,\sigma_1^2) = (2.3205, 1.9812)$, $(\mu_2,\sigma_2^2) = (0,1)$, $(\boldsymbol{P_1}, \boldsymbol{P_2}) = (0.9844, 0.2500)$, $\mathbf{P_d} = 0.9$								
(20, 30)	0.9856	0.2853	0.9330	0.9565	0.9505	0.0211	0.2579	0.0044
(30, 30)	0.9856	0.2712	0.9400	0.9595	0.9475	0.0188	0.2022	0.0031
(50, 40)	0.9852	0.2643	0.9370	0.9525	0.9455	0.0154	0.1538	0.0020
(50, 50)	0.9852	0.2633	0.9480	0.9505	0.9525	0.0142	0.1491	0.0018
(55, 65)	0.9848	0.2604	0.9455	0.9500	0.9470	0.0131	0.1381	0.0016
(75, 75)	0.9848	0.2583	0.9420	0.9500	0.9420	0.0116	0.1182	0.0012
(100,100)	0.9846	0.2563	0.9465	0.9490	0.9530	0.0101	0.1009	0.0009
$(\mu_1,\sigma_1^2) = (1.1712, 0.5047)$, $(\mu_2,\sigma_2^2) = (0,1)$, $(\boldsymbol{P_1}, \boldsymbol{P_2}) = (0.2500, 0.9844)$, $\mathbf{P_d} = 0.1$								
(20, 30)	0.2771	0.9861	0.9505	0.9235	0.9310	0.2172	0.0214	0.0037
(30, 30)	0.2737	0.9855	0.9520	0.9410	0.9450	0.2044	0.0188	0.0032
(50, 40)	0.2653	0.9850	0.9495	0.9445	0.9460	0.1660	0.0152	0.0021
(50, 50)	0.2649	0.9852	0.9540	0.9425	0.9515	0.1504	0.0142	0.0018
(55, 65)	0.2621	0.9851	0.9470	0.9435	0.9460	0.1320	0.0131	0.0015
(75, 75)	0.2589	0.9849	0.9510	0.9375	0.9460	0.1184	0.0115	0.0012
(100,100)	0.2559	0.9847	0.9550	0.9485	0.9515	0.1007	0.0100	0.0009
$(\mu_1,\sigma_1^2) = (1.1712, 0.5047)$, $(\mu_2,\sigma_2^2) = (0,1)$, $(\boldsymbol{P_1}, \boldsymbol{P_2}) = (0.7500, 0.8750)$, $\mathbf{P_d} = 0.5$								
(20, 30)	0.7683	0.8890	0.9555	0.9290	0.9410	0.1893	0.1459	0.0244
(30, 30)	0.7647	0.8833	0.9515	0.9380	0.9505	0.1809	0.1301	0.0210
(50, 40)	0.7601	0.8805	0.9535	0.9395	0.9510	0.1524	0.1055	0.0144
(50, 50)	0.7597	0.8808	0.9535	0.9405	0.9410	0.1385	0.0997	0.0123
(55, 65)	0.7581	0.8802	0.9515	0.9415	0.9465	0.1232	0.0919	0.0101
(75, 75)	0.7559	0.8788	0.9495	0.9455	0.9495	0.1124	0.0813	0.0082
(100,100)	0.7553	0.8781	0.9475	0.9425	0.9430	0.0972	0.0704	0.0061

(Continues)

TABLE 2 (Continued)

Sample Size	Estimate		Coverage			Width/area		
	PPV	NPV	PPV	NPV	CR	PPV	NPV	CR
$(\mu_1, \sigma_1^2) = (1.1712, 0.5047)$, $(\mu_2, \sigma_2^2) = (0,1)$, $(\boldsymbol{P_1}, \boldsymbol{P_2}) = (0.9643, 0.4375)$, $\mathbf{P_d} = 0.9$								
(20, 30)	0.9668	0.4831	0.9595	0.9270	0.9410	0.0352	0.3391	0.0107
(30, 30)	0.9667	0.4703	0.9525	0.9325	0.9435	0.0335	0.2964	0.0089
(50, 40)	0.9660	0.4580	0.9530	0.9520	0.9555	0.0282	0.2398	0.0060
(50, 50)	0.9657	0.4563	0.9525	0.9440	0.9530	0.0256	0.2273	0.0052
(55, 65)	0.9652	0.4519	0.9595	0.9480	0.9505	0.0228	0.2092	0.0043
(75, 75)	0.9652	0.4513	0.9580	0.9360	0.9495	0.0207	0.1848	0.0034
(100,100)	0.9651	0.4489	0.9540	0.9475	0.9515	0.0179	0.1598	0.0026

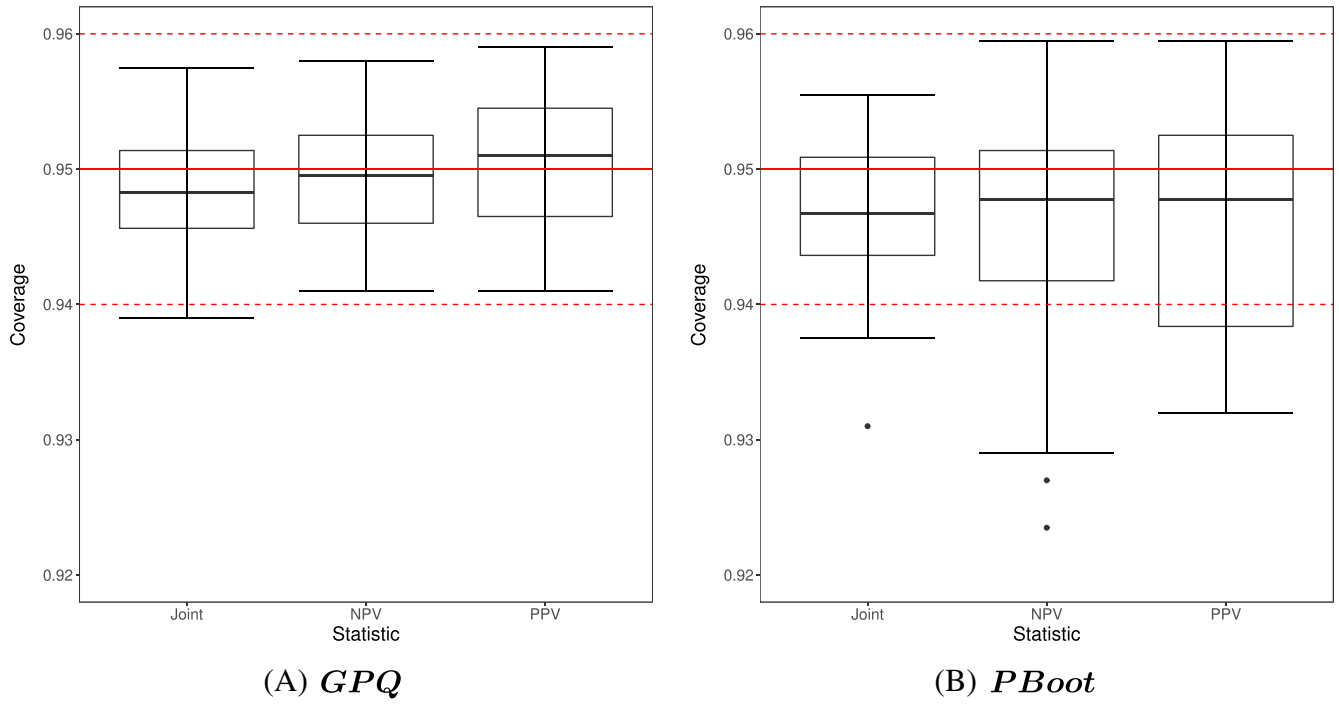
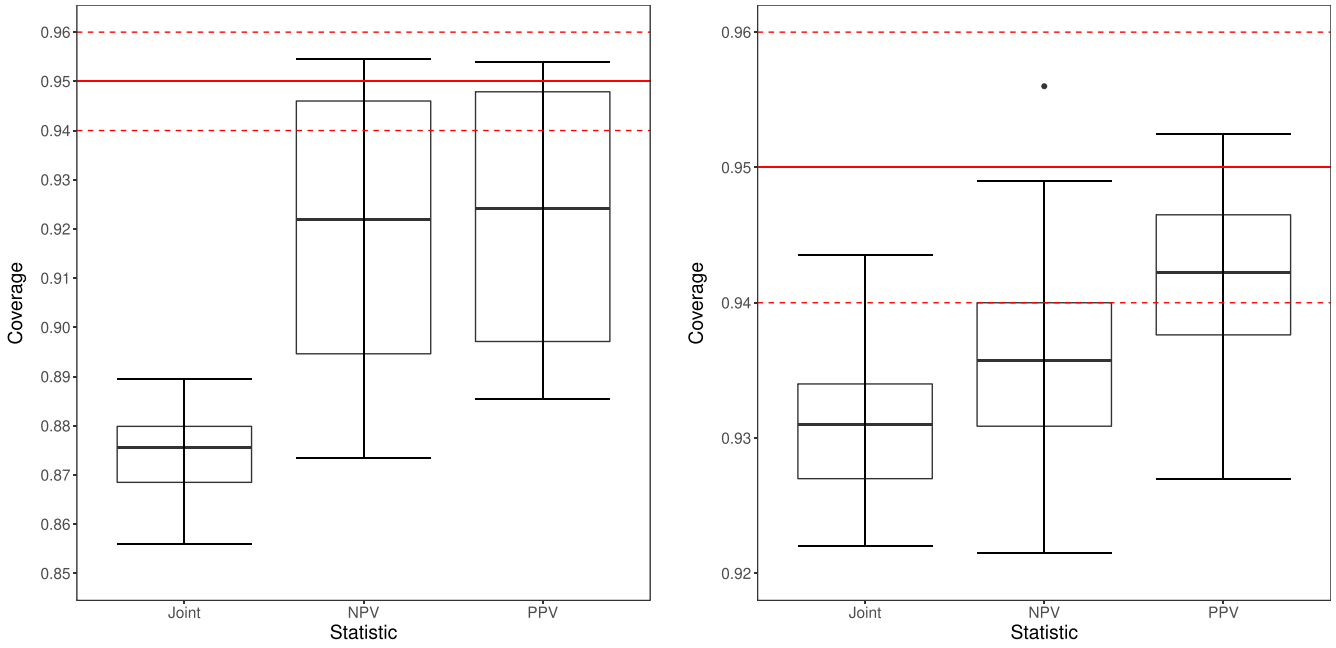


FIGURE 1 Boxplots of 95% confidence regions for PPV-NPV joint region, NPV, and PPV under bi-normal distributions for ***GPQ*** and ***PBoot*** methods

confidence regions, both ***GPQT*** and ***PBootT*** methods produce similar results, although the ***GPQT*** method tends to produce smaller areas. However, smaller area of ***GPQT*** is offset by its poorer coverage. Figure 2 demonstrates that ***GPQT*** uniformly underperforms compared to ***PBootT*** in terms of coverage probability, hence the simulation results of ***GPQT*** were not presented in a table and we do not recommend the ***GPQ*** method for non-normal data even after the Box-Cox transformation.

Table 3 presents simulation results of joint confidence regions for $\boldsymbol{\nu} = (P_1, P_2)^T$ at the nominal level of 95% under gamma distributions using the ***PBootT*** approach. The ***PBootT*** method provides at least 92% coverage of the joint confidence region, which underestimated the nominal coverage. We should apply ***PBootT*** method if Box-Cox transformation is needed to approximate normality for the observed data. Tables A3 and A4, present the simulation results under bi-gamma distributions for the Mercaldo[2] and Stamey[3] methods, respectively. Similar to the simulation results under normal distributions, both existing methods tend to be conservative and overestimate the coverage for the individual confidence intervals.



(A) ***GPQT*** (B) ***PBootT***

FIGURE 2 Boxplots of 95% confidence regions for PPV-NPV joint region, NPV, and PPV under bi-gamma distributions for ***GPQT*** and ***PBootT*** methods. *Note:* the scale for the y-axis (coverage) in (A) is different from (B) as ***GPQT*** has a lot lower coverage for the joint CR

5 | DATA EXAMPLE

Alzheimer's disease (AD) is a specific type of dementia characterized by loss of memory and other cognitive abilities that are important for performing daily activities. Mild cognitive impairment (MCI) due to AD, is a quickly progressing form of dementia caused by, typically undiagnosed, AD. While there is no cure for cognitively impaired (AD or MCI) patients, ascertaining an early diagnosis is crucial for developing treatment plans to slow down disease progression and for establishing legal and financial arrangements prior to further cognitive decline. The ADNI[25] is a push to advance AD research through the discovery and development of AD biomarkers. The data used in this example contains 114 cognitively healthy controls and 301 cognitively impaired (AD or MCI due to AD) cases.

Figure 3 presents the Q-Q plots of two common biomarkers used for MCI and AD diagnosis, the intracranial cerebrospinal fluid volume (ICV) and total Tau (TAU). ICV is normally distributed for both cognitively healthy and cognitively impaired groups, while TAU is not. Thus, we use ICV to illustrate the ***GPQ*** confidence region under normality and TAU to illustrate ***PBootT*** without normality. For both situations, logit transformations of PPV and NPV were applied and the resulting confidence region and intervals were inversely transformed to obtain those of the original PPV and NPV probabilities.

The prevalence of cognitively impaired disease (AD or MCI) cannot be estimated from the ADNI data. In addition, prevalence varies widely depending on demographics, especially age and education.[26] Prior research[27] concluded the prevalence estimates of MCI ranges from 5% to 36.7%. A more recent summary of Alzheimer's disease suggested that approximately 15% to 20% of people age 65 or older have MCI and among individuals with MCI who were tracked for 5 years or longer, an average of 38 percent developed AD.[28] In our analysis, since we combined MCI and AD as one disease group, we need to estimate the prevalence of the combined group for participants of age 54.4 and older. Suppose MCI prevalence is 5%–20% for subjects aged 54.4 and older, the prevalence of AD is then calculated by multiplying MCI prevalence by 0.38, which is around 2%–8%, therefore, the total prevalence of (MCI + AD) is 7%–28%.

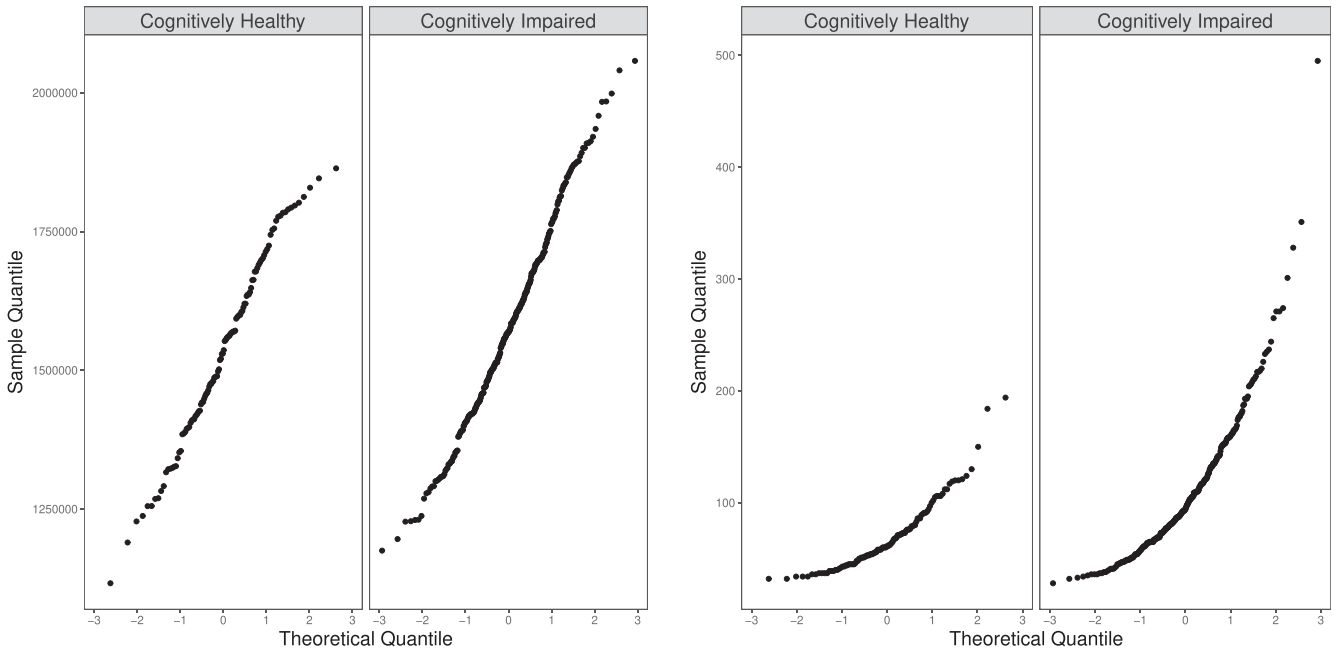
Therefore, in order to obtain more valid confidence interval and region estimations, three analyses were conducted for varying prevalence estimates of 10%, 20%, and 30% for this ADNI population of participants aged 54.4 years and older. In addition, for comparison purposes, the individual confidence intervals for PPV and NPV were adjusted for multiple testing using the Bonferroni method (i.e., the confidence level is set at $100(1-\alpha/2)\%$). Figures 4–6 present the elliptical confidence regions of PPV and NPV using the ***GPQ*** method for biomarker ICV and the ***PBootT*** methods for

TABLE 3 Summary of approximate 95% joint confidence regions of ***PBootT*** method for (P_1, P_2) under gamma distributions (based on 2000 simulations)

Sample Size	Estimate		Coverage			Width/area		
	PPV	NPV	PPV	NPV	CR	PPV	NPV	CR
$(\alpha_1, \beta) = (2.4741, 0.1850)$, $(\alpha_2, \beta) = (5, 1)$, $(\boldsymbol{P_1}, \boldsymbol{P_2}) = (0.4375, 0.9643)$, $\mathbf{P_d} = 0.1$								
(20, 30)	0.4768	0.9701	0.9515	0.9330	0.9350	0.3926	0.0425	0.0158
(30, 30)	0.4712	0.9688	0.9395	0.9315	0.9220	0.3505	0.0347	0.0116
(50, 40)	0.4599	0.9673	0.9400	0.9425	0.9330	0.2862	0.0267	0.0073
(50, 50)	0.4521	0.9670	0.9270	0.9360	0.9240	0.2686	0.0261	0.0066
(55, 65)	0.4506	0.9669	0.9335	0.9300	0.9220	0.2468	0.0243	0.0056
(75, 75)	0.4453	0.9664	0.9500	0.9420	0.9330	0.2187	0.0210	0.0043
(100,100)	044359	0.9660	0.9445	0.9365	0.9370	0.1899	0.0180	0.0032
$(\alpha_1, \beta) = (2.4741, 0.1850)$, $(\alpha_2, \beta) = (5, 1)$, $(\boldsymbol{P_1}, \boldsymbol{P_2}) = (0.8750, 0.7500)$, $\mathbf{P_d} = 0.5$								
(20, 30)	0.8826	0.7852	0.9465	0.9275	0.9265	0.1782	0.2288	0.0385
(30, 30)	0.8824	0.7768	0.9380	0.9355	0.9355	0.1565	0.1905	0.0283
(50, 40)	0.8786	0.7669	0.9380	0.9435	0.9350	0.1286	0.1469	0.0177
(50, 50)	0.8772	0.7648	0.9375	0.9415	0.9310	0.1216	0.1433	0.0165
(55, 65)	0.8769	0.7640	0.9465	0.9560	0.9435	0.1115	0.1335	0.0140
(75, 75)	0.8761	0.7624	0.9400	0.9445	0.9335	0.0985	0.1151	0.0107
(100,100)	0.8752	0.7597	0.9475	0.9400	0.9300	0.0857	0.0989	0.0079
$(\alpha_1, \beta) = (2.4741, 0.1850)$, $(\alpha_2, \beta) = (5, 1)$, $(\boldsymbol{P_1}, \boldsymbol{P_2}) = (0.9844, 0.2500)$, $\mathbf{P_d} = 0.9$								
(20, 30)	0.9855	0.3051	0.9470	0.9290	0.9310	0.0267	0.2944	0.0068
(30, 30)	0.9851	0.2857	0.9465	0.9450	0.9385	0.0236	0.2276	0.0047
(50, 40)	0.9848	0.2745	0.9305	0.9325	0.9235	0.0188	0.1673	0.0028
(50, 50)	0.9847	0.2727	0.9480	0.9305	0.9285	0.7663	0.1609	0.0026
(55, 65)	0.9847	0.2708	0.9455	0.9390	0.9295	0.0161	0.1484	0.0022
(75, 75)	0.9845	0.2659	0.9365	0.9395	0.9340	0.0142	0.1254	0.0016
(100,100)	0.9845	0.2628	0.9395	0.9345	0.9355	0.0123	0.1061	0.0012
$(\alpha_1, \beta) = (26.0183, 3.3500)$, $(\alpha_2, \beta) = (5, 1)$, $(\boldsymbol{P_1}, \boldsymbol{P_2}) = (0.2500, 0.9844)$, $\mathbf{P_d} = 0.1$								
(20, 30)	0.2872	0.9867	0.9350	0.9125	0.9230	0.2335	0.0245	0.0048
(30, 30)	0.2871	0.9862	0.9425	0.9190	0.9130	0.2217	0.0224	0.0043
(50, 40)	0.2757	0.9854	0.9445	0.9350	0.9315	0.1779	0.0184	0.0029
(50, 50)	0.2717	0.9855	0.9395	0.9310	0.9255	0.1572	0.0170	0.0024
(55, 65)	0.2676	0.9853	0.9360	0.9225	0.9160	0.1371	0.0154	0.0019
(75, 75)	0.2628	0.9851	0.9495	0.9400	0.9315	0.1216	0.0137	0.0015
(100,100)	0.2606	0.9849	0.9450	0.9350	0.9315	0.1038	0.0119	0.0011
$(\alpha_1, \beta) = (26.0183, 3.3500)$, $(\alpha_2, \beta) = (5, 1)$, $(\boldsymbol{P_1}, \boldsymbol{P_2}) = (0.7500, 0.8750)$, $\mathbf{P_d} = 0.5$								
(20, 30)	0.7790	0.8947	0.9445	0.9250	0.9225	0.1951	0.1640	0.0297
(30, 30)	0.7756	0.8895	0.9370	0.9290	0.9275	0.1857	0.1512	0.0262
(50, 40)	0.7673	0.8827	0.9365	0.9490	0.9260	0.1553	0.1272	0.0187
(50, 50)	0.7669	0.8849	0.9360	0.9355	0.9275	0.1400	0.1176	0.0154
(55, 65)	0.7625	0.8829	0.9445	0.9275	0.9295	0.1245	0.1069	0.0125
(75, 75)	0.7619	0.8816	0.9460	0.9380	0.9315	0.1137	0.0958	0.0103
(100,100)	0.7585	0.8800	0.9415	0.9355	0.9360	0.0977	0.0831	0.0077
$(\alpha_1, \beta) = (26.0183, 3.3500)$, $(\alpha_2, \beta) = (5, 1)$, $(\boldsymbol{P_1}, \boldsymbol{P_2}) = (0.9643, 0.4375)$, $\mathbf{P_d} = 0.9$								

TABLE 3 (Continued)

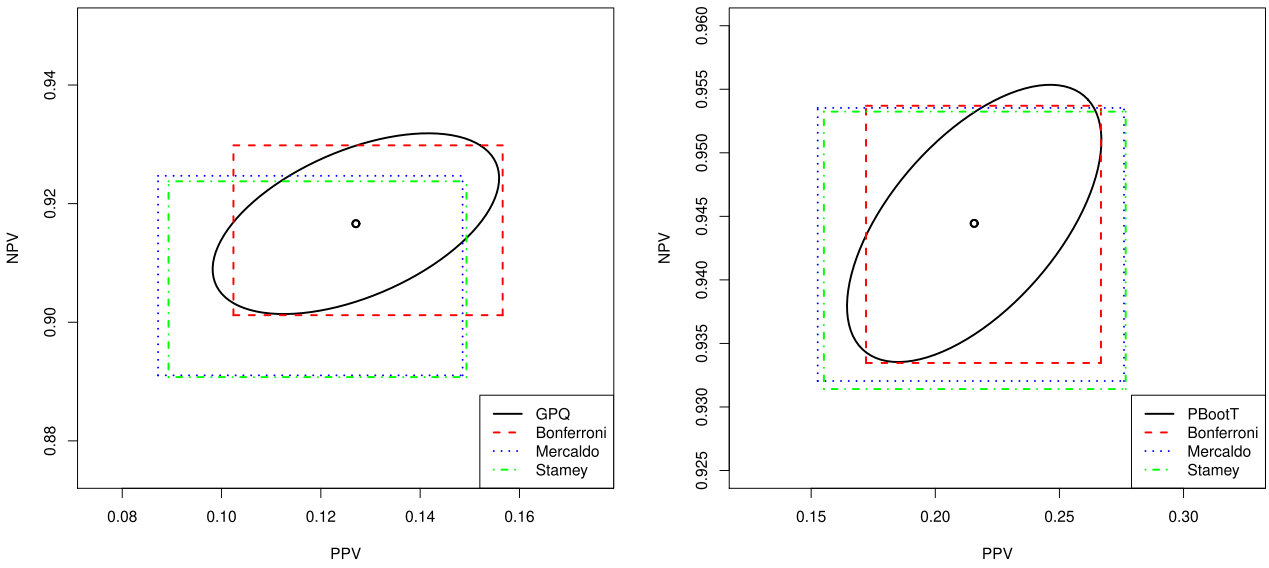
Sample Size	Estimate		Coverage			Width/area		
	PPV	NPV	PPV	NPV	CR	PPV	NPV	CR
(20, 30)	0.9688	0.5018	0.9525	0.9215	0.9195	0.0357	0.3890	0.0133
(30, 30)	0.9688	0.4883	0.9450	0.9280	0.9280	0.0336	0.3556	0.0113
(50, 40)	0.9673	0.4700	0.9390	0.9370	0.9295	0.0283	0.2905	0.0078
(50, 50)	0.9671	0.4668	0.9495	0.9390	0.9340	0.0255	0.2720	0.0066
(55, 65)	0.9664	0.4623	0.9420	0.9400	0.9240	0.0227	0.2472	0.0053
(75, 75)	0.9663	0.4587	0.9355	0.9360	0.9330	0.0206	0.2207	0.0043
(100,100)	0.9657	0.4524	0.9510	0.9435	0.9360	0.0179	0.1912	0.0032



(A) ICV (B) TAU

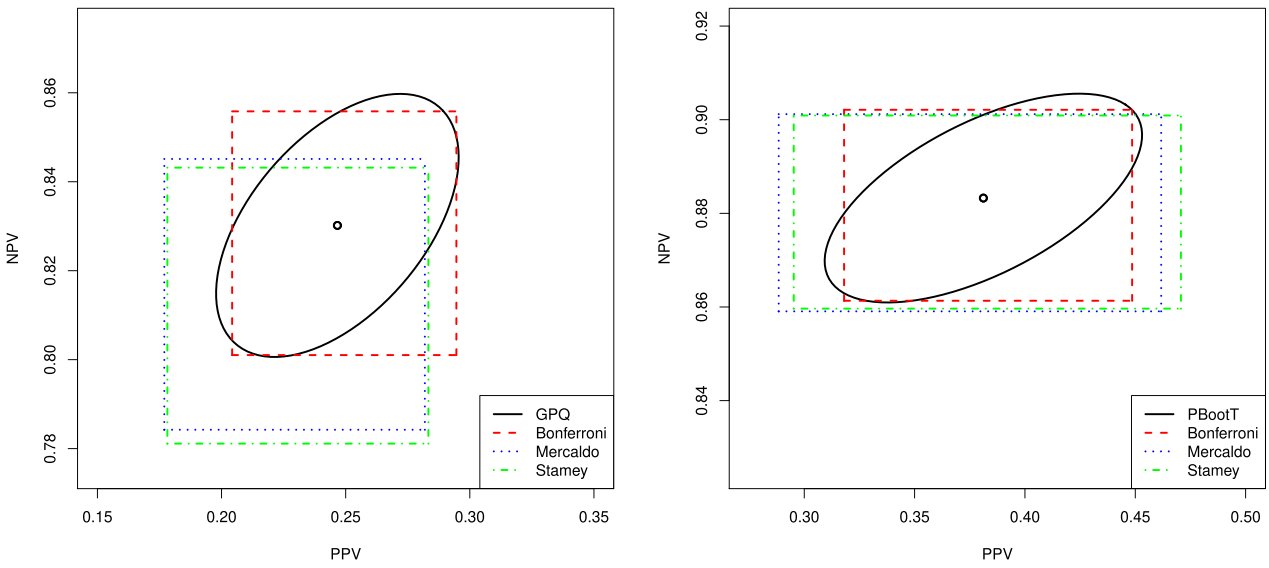
FIGURE 3 Q-Q plots of (A) marker ICV and (B) marker TAU. Values of marker ICV for both cognitively healthy and cognitively impaired groups are normally distributed, while values of marker TAU are not

biomarker TAU, along with the corresponding Bonferroni corrected rectangular regions, under pre-specified prevalence of 10%, 20%, and 30%, respectively. Additionally, results for the individual confidence intervals using the two existing methods[2,3] are provided and the results are adjusted for multiple testing using the Bonferroni method similarly. From Figures 4-6, we can easily see that the coverage area of the rectangular confidence region estimated using the Bonferroni method (rectangular) is much larger than the elliptical area obtained by the proposed confidence region, regardless of the method used for constructing the individual confidence intervals. Therefore, the Bonferroni method would produce confidence regions to be too conservative as compared to the proposed confidence region which considers the correlation between the two predictive values as well as simultaneously maintain the type I error for both. In addition, among all rectangular regions, we observe the two existing confidence intervals based on the binomial distribution give larger areas than the proposed confidence intervals derived for a continuous biomarker. Furthermore, we observed that the elliptical regions are tilted and the differences in the lengths of two axis are relatively large indicating there is a strong correlation between PPV and NPV estimates, and thus the proposed elliptical joint confidence region is preferred.



(A) ICV (B) TAU

FIGURE 4 95% Joint confidence regions of PPV (P_1) and NPV (P_2) based on the proposed ***GPQ*** and ***PBootT*** methods for (A) biomarker ICV and (B) biomarker TAU. Three rectangular regions formed by joining the respective confidence intervals with Bonferroni correction for PPV and NPV were plotted. The confidence intervals are calculated based on (1) the proposed individual ***GPQ*** (for ICV) and ***PBootT*** (for TAU) confidence intervals (C.I.); (2) Mercaldo C.I.s; (3) Stamey C.I.s. The limits of the C.I.s as well as the area of the confidence regions were given in Table 4. The estimates of P_1 and P_2 were determined using an estimate of 10% for disease prevalence. The ***GPQ*** confidence region in (A) is given by the elliptical equation $((x-0.1271)^2/0.2887^2)+((y-0.9166)^2/0.1551^2)=1$ with major axis being in the direction of vector $\pm(-1,-0.6001)^T$ and with point (0.1271,0.9166) as the origin. The ***PBootT*** confidence region in (B) is inversely transformed and given by the elliptical equation $((x-0.3329)^2/0.3341^2)+((y-0.9442)^2/0.1112^2)$ with major axis being in the direction of vector $\pm(-1,-0.3936)^T$ and with point (0.3329,0.9442) as the origin



(A) ICV (B) TAU

FIGURE 5 95% Joint confidence regions of PPV (P_1) and NPV (P_2) based on the proposed ***GPQ*** and ***PBootT*** methods for (A) biomarker ICV and (B) biomarker TAU. Three rectangular regions formed by joining the respective confidence intervals with Bonferroni correction for PPV and NPV were plotted. The confidence intervals are calculated based on (1) the proposed individual ***GPQ*** (for ICV) and ***PBootT*** (for TAU) confidence intervals (C.I.); (2) Mercaldo C.I.s; (3) Stamey C.I.s. The limits of the C.I.s as well as the area of the confidence regions were given in Table 5. The estimates of P_1 and P_2 were determined using an estimate of 20% for disease prevalence. The ***GPQ*** confidence region in (A) is given by the elliptical equation $((x-0.2467)^2/0.2963^2)+((y-0.8302)^2/0.1592^2)=1$ with major axis being in the direction of vector $\pm(-1,-0.6529)^T$ and with point (0.2467,0.8302) as the origin. The ***PBootT*** confidence region in (B) is inversely transformed and given by the elliptical equation $((x-0.5276)^2/0.3378^2)+((y-0.8827)^2/0.1160^2)$ with major axis being in the direction of vector $\pm(-1,-0.4059)^T$ and with point (0.5276,0.8827) as the origin

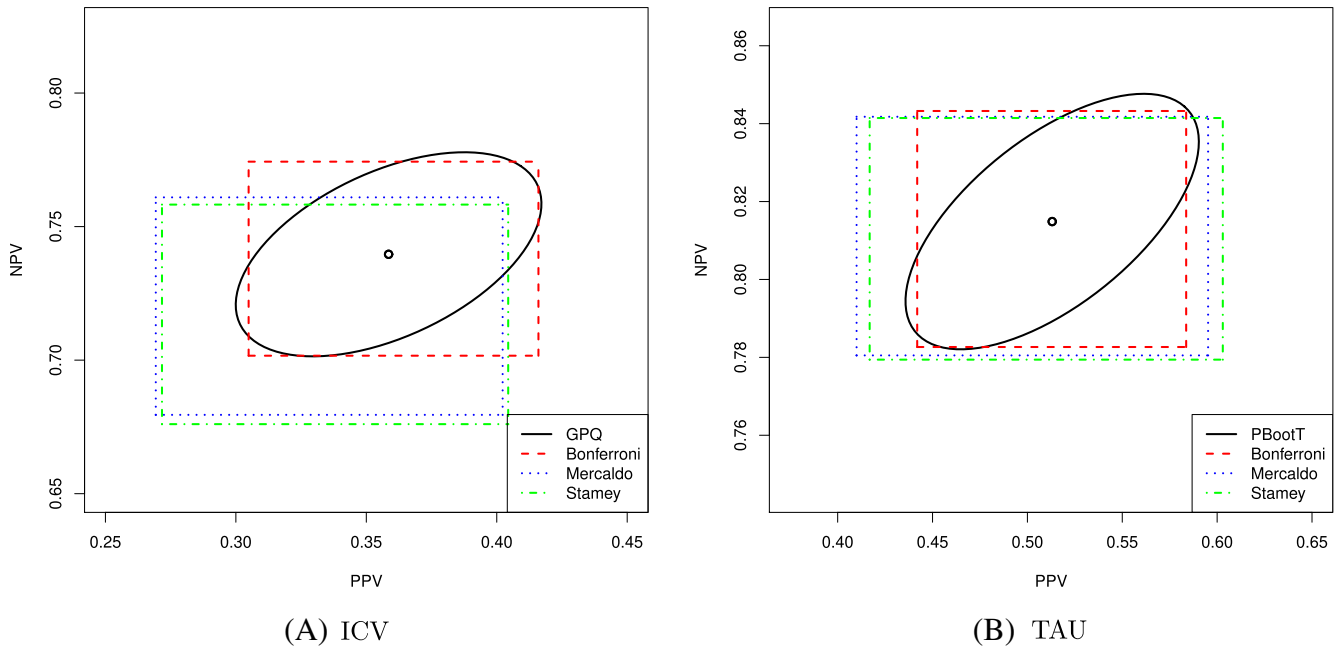


FIGURE 6 95% joint confidence regions of PPV (P_1) and NPV (P_2) based on the proposed ***GPQ*** and ***PBootT*** methods for (A) biomarker ICV and (B) biomarker TAU. Three rectangular regions formed by joining the respective confidence intervals with Bonferroni correction for PPV and NPV were plotted. The confidence intervals are calculated based on (1) the proposed individual ***GPQ*** (for ICV) and ***PBootT*** (for TAU) confidence intervals (C.I.); (2) Mercaldo C.I.s; (3) Stamey C.I.s. The limits of the C.I.s as well as the area of the confidence regions were given in Table 6. The estimates of P_1 and P_2 were determined using an estimate of 30% for disease prevalence. The ***GPQ*** confidence region in (A) is given by the elliptical equation $((x-0.3585)^2/0.2833^2)+((y-0.7396)^2/0.1550^2)=1$ with major axis being in the direction of vector $\pm(-1,-0.6127)^T$ and with point (0.3585,0.7396) as the origin. The ***PBootT*** confidence region in (B) is inversely transformed and given by the elliptical equation $((x-0.6564)^2/0.3437^2)+((y-0.8140)^2/0.1159^2)$ with major axis being in the direction of vector $\pm(-1,-0.4150)^T$ and with point (0.6564,0.8140) as the origin

TABLE 4 Summary of simultaneous confidence region and interval estimations about PPV (P_1) and NPV (P_2) using a prevalence estimate of 10%

CR	Method	ICV	TAU		
Area	Elliptical	0.0012	0.0014		
	Rectangular	0.0016	0.0020		
	$\text{Rectangular}_{Mercaldo}$	0.0021	0.0027		
	$\text{Rectangular}_{Stamey}$	0.0020	0.0027		
CI		Point.Est	CI	Point.Est	CI
PPV	95% Simul. CI	0.1271	(0.1009,0.1588)	0.3329	(0.2672,0.4058)
	95% Bonfer. CI	0.1271	(0.1024,0.1566)	0.3329	(0.2717,0.4003)
	95% Bonfer. $\text{CI}_{Mercaldo}$	0.1144	(0.0872,0.1486)	0.2077	(0.1526,0.2760)
	95% Bonfer. CI_{Stamey}	0.1139	(0.0893,0.1493)	0.2048	(0.1551.0.2767)
NPV	95% Simul. CI	0.9166	(0.9001,0.9306)	0.9442	(0.9352,0.9521)
	95% Bonfer. CI	0.9166	(0.9012,0.9298)	0.9442	(0.9359,0.9515)
	95% Bonfer. $\text{CI}_{Mercaldo}$	0.9093	(0.8910,0.9247)	0.9437	(0.9320,0.9535)
	95% Bonfer. CI_{Stamey}	0.9090	(0.8908,0.9238)	0.9434	(0.9314,0.9532)

Note: Area: Area of the confidence regions (CR). Elliptical: The elliptical joint confidence regions estimated by ***GPQ*** method for marker ICV and ***PBootT*** method for marker TAU, respectively. Rectangular: The rectangular confidence regions formed by two Bonferroni-corrected individual confidence intervals of PPV and NPV estimated by ***GPQ*** method for marker ICV and ***PBootT*** method for marker TAU, respectively. Rectangular-Mercaldo: The rectangular confidence regions formed by two Bonferroni-corrected individual confidence intervals of PPV and NPV estimated by Mercaldo's method for both markers ICV and TAU (with Box-Cox transformation), respectively. Rectangular-Stamey: The rectangular confidence regions formed by two Bonferroni-corrected individual confidence intervals of PPV and NPV estimated by Stamey's method for both markers ICV and TAU (with Box-Cox transformation), respectively. Simul. CI: Simultaneous confidence intervals (CI) derived from the elliptical joint confidence regions. Bonfer. CI: Bonferroni-corrected simultaneous confidence intervals.

TABLE 5 Summary of simultaneous confidence region and interval estimations about PPV (P_1) and NPV (P_2) using a prevalence estimate of 20%

CR	Method	ICV	TAU		
Area	Elliptical	0.0039	0.0032		
	Rectangular	0.0049	0.0045		
	$\text{Rectangular}_{Mercaldo}$	0.0064	0.0073		
	$\text{Rectangular}_{Stamey}$	0.0065	0.0072		
CI		Point.Est	CI	Point.Est	CI
PPV	95% Simul. CI	0.2467	(0.2011,0.2987)	0.5276	(0.4488,0.6050)
	95% Bonfer. CI	0.2467	(0.2043,0.2946)	0.5276	(0.4558,0.5983)
	95% Bonfer. $\text{CI}_{Mercaldo}$	0.2251	(0.1769,0.2819)	0.3709	(0.2884,0.4617)
	95% Bonfer. CI_{Stamey}	0.2244	(0.1782,0.2833)	0.3668	(0.2952,0.4707)
NPV	95% Simul. CI	0.8302	(0.7985,0.8577)	0.8827	(0.8644,0.8989)
	95% Bonfer. CI	0.8302	(0.8010,0.8558)	0.8827	(0.8661,0.8975)
	95% Bonfer. $\text{CI}_{Mercaldo}$	0.8166	(0.7842,0.8451)	0.8817	(0.8591,0.9012)
	95% Bonfer. CI_{Stamey}	0.8162	(0.7811,0.8432)	0.8810	(0.8597,0.9009)

Note: Area: Area of the confidence regions. Elliptical: The elliptical joint confidence regions estimated by ***GPQ*** method for marker ICV and ***PBootT*** method for marker TAU, respectively. Rectangular: The rectangular confidence regions formed by two Bonferroni-corrected simultaneous confidence intervals of PPV and NPV. Simul. CI: Simultaneous confidence intervals derived from the elliptical joint confidence regions. Bonfer. CI: Bonferroni-corrected individual confidence intervals.

TABLE 6 Summary of simultaneous confidence region and interval estimations about PPV (P_1) and NPV (P_2) using a prevalence estimate of 30%

CR	Method	ICV	TAU		
Area	Elliptical	0.0061	0.0043		
	Rectangular	0.0081	0.0062		
	$\text{Rectangular}_{Mercaldo}$	0.0108	0.0114		
	$\text{Rectangular}_{Stamey}$	0.0109	0.0115		
CI		Point.Est	CI	Point.Est	CI
PPV	95% Simul. CI	0.3585	(0.3023,0.4190)	0.6564	(0.5809,0.7247)
	95% Bonfer. CI	0.3585	(0.3049,0.4160)	0.6564	(0.5877,0.7191)
	95% Bonfer. $\text{CI}_{Mercaldo}$	0.3325	(0.2693,0.4023)	0.5027	(0.4100,0.5952)
	95% Bonfer. CI_{Stamey}	0.3315	(0.2717,0.4044)	0.4983	(0.4169,0.6030)
NPV	95% Simul. CI	0.7396	(0.6996,0.7760)	0.8140	(0.7870,0.8384)
	95% Bonfer. CI	0.7396	(0.7017,0.7743)	0.8140	(0.7894,0.8364)
	95% Bonfer. $\text{CI}_{Mercaldo}$	0.7221	(0.6795,0.7609)	0.8131	(0.7805,0.8418)
	95% Bonfer. CI_{Stamey}	0.7215	(0.6760,0.7582)	0.8119	(0.7794,0.8414)

Note: Area: Area of the confidence regions. Elliptical: The elliptical joint confidence regions estimated by ***GPQ*** method for marker ICV and ***PBootT*** method for marker TAU, respectively. Rectangular: The rectangular confidence regions formed by two Bonferroni-corrected simultaneous confidence intervals of PPV and NPV. Simul. CI: Simultaneous confidence intervals derived from the elliptical joint confidence regions. Bonfer. CI: Bonferroni-corrected simultaneous confidence intervals.

Full results of the analysis of ICV and TAU can be found in Tables 4–6 for prevalence estimates of 10%, 20%, and 30%, respectively. These include point estimates and 95% confidence intervals (both the simultaneous interval derived from the elliptical region and the Bonferroni corrected interval) for PPV and NPV, as well as the areas of respective confidence regions. Additionally, results for the Bonferroni-adjusted rectangular regions using Mercaldo's[2] and Stamey's[3] methods are also presented. The areas in the rectangular confidence regions, regardless of the method used, are larger than those produced using the elliptical method.

6 | CONCLUSIONS AND DISCUSSIONS

PPVs and NPVs at the optimal cut-off point associated with the Youden index are important measures for evaluating a biomarker's diagnostic accuracy and, additionally, they are post-test accuracy measures (i.e., conditioning on the test results), which is more intuitive and practical. The joint inference of PPV and NPV offers a comprehensive view of the diagnostic potential of a biomarker. The proposed elliptical joint confidence region maintains the type I error rate when evaluating a biomarker through both PPV and NPV under test positive and negative, respectively. In the meantime, it takes into account the correlation between PPV and NPV, something that has not been previously addressed. Furthermore, past research of PPV and NPV has focused on binary tests. To apply those methods for a continuous test, dichotomizing at a pre-specified cut-off is needed, however, the variability of the cut-off estimate is not accounted for. The proposed methods are framed for continuous biomarker/test measurements so the estimation variability of the cut-off is considered, which leads to superior performance compared to the existing methods as seen in the simulations.

As the simulation results indicate, all methods using the logit-transformed probabilities perform better than the same method without the logit transformation. The ***GPQ*** method performs quite well under normality, and slightly outperforms the ***PBoot*** method, with closer coverage probabilities to the nominal level and smaller areas (thus more precise). However, under non-normal conditions, the parametric bootstrap method with Box-Cox transformation (***PbootT***) markedly outperforms the generalized inference approach under Box-Cox transformation (***GPQT***). Thus, ***GPQT*** approach is not recommended for non-normal data, while the ***PBoot*** and ***PBootT*** methods can be utilized under normal and non-normal conditions respectively. To further improve the performance of the proposed confidence region under non-normal situations, future research can use non-parametric methods such as the empirical estimate[8] or the kernel smoothed version[9] of the sensitivity and specificity to plug into the Bayes equations for estimating the PPV and NPV quantities while using the non-parametric bootstrap methods for the variance estimation. Moreover, the proposed confidence intervals outperform the two existing confidence interval estimation methods uniformly for the continuous test if the estimation of the diagnostic cut-off is needed. Finally, the data example provides evidence that the proposed elliptical joint confidence region produces smaller areas compared to the rectangular, Bonferroni-adjusted region, and is thus, more preferred.

For different types of disease diagnosis, sensitivity and specificity may possess different degrees of importance. For example, a diagnostic or screening test of high sensitivity may be more preferred if the disease has high mortality and a cure is available, such as tests for the early detection of cervical cancer in-situ, for which a simple laser surgery would treat effectively. In other cases, the consequence of having a false positive is very serious (e.g., a false diagnosis of leukemia resulting in strong doses of chemotherapy with severe side-effects, large amount of medical expenditures and a heavy psychological burden), then a diagnostic test of high specificity is required. The Youden index criterion for the cut-off selection weighs the sensitivity and specificity equally important and, thus, may not be practical for some settings. In such settings, it is recommended to use the weighted Youden index,[29] which is a sum of weighted sensitivity and specificity that multiplies different weight coefficients to address varying degrees of importance.

Furthermore, in this research, we assume prevalence is fixed and the pre-test diagnostic accuracy probabilities (sensitivity and specificity) depend merely on the choice of the cut-off and the proposed method accounted for the variability of estimating the unknown cut-off. In practice, various factors can influence the performance of a diagnostic test, resulting in different values of sensitivity and specificity, such as the testing location, the clinicians, the patient characteristics. Additionally, different patient characteristics may lead to different disease prevalence. To deal with various sources of confounding, the regression analysis can be applied to account for the variability of pre-test accuracy measures as well as the prevalence due to confounding for estimating the confidence region and intervals of the post-test accuracy probabilities—the PPV and NPV.

In addition to evaluating diagnostic accuracy, the joint inference methods for the PPV and NPV could be applied to novel trial design topics, most notably studies utilizing enrichment and/or adaptive designs. Adaptive trials use information obtained throughout a clinical trial to adapt aspects of the study design such as patient enrollment, treatment allocation, and sample size planning. For the purpose of patient selection in enrichment studies, first, a predictive logistic regression model using the screening data as predictors is developed to predict patient response to treatment. The condition for enrichment analysis would be responders versus non-responders, which is comparable to test positive versus negative for a diagnostic test evaluation. Then the question of interest would be, "What's the probability that a subject responds at the end of the study given their predicted probability of responding to treatment based on the screening data?" Enrichment studies are especially popular in the Alzheimer's Disease (AD) therapeutic area where well-known AD biomarkers, such as cerebrospinal fluid, can be used to predict an individual's probability of responding to

treatment, prior to enrollment.[30,31] This allows sponsors to enrich their studies with subjects who are more likely to respond positively to study treatments. The post-test probabilities of response to treatment could be summarized using the proposed joint confidence region method in order to make more informed decisions regarding enrollment criteria, which may significantly reduce sample size, increase power, and generally increase the chances of detecting a statistically significant treatment effect.[32] Note that although large prevalence rates are less common for disease diagnosis, in broader applications, such as enrichment designs, the "prevalence" is the treatment response rate in the general population, which can be of large probability values. In the simulation, we have settings at $P_d = 0.9$ to demonstrate the performance under broader applications.

An R program containing functions for the proposed methods is included in the supplementary document.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

DATA AVAILABILITY STATEMENT

The data used in Section 5 are openly available on the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu).

ORCID

Jingjing Yin https://orcid.org/0000-0003-4843-613X

REFERENCES

1. Zhou X-H, McClish DK, Obuchowski NA. *Statistical Methods in Diagnostic Medicine*. Vol 569. Hoboken, New Jersey, USA: Wiley-Interscience; 2009.
2. Mercaldo ND, Lau KF, Zhou XH. Confidence intervals for predictive values with an emphasis to case–control studies. *Stat Med*. 2007;26(10):2170-2183.
3. Stamey JD, Holt MM. Bayesian interval estimation for predictive values from case-control studies. *Commun Stat Simul Comput*. 2009;39(1):101-110.
4. Youden W. Index for rating diagnostic tests. *Cancer*. 1950;3(1):32-35.
5. Verniquet A, Kakel R. How accurate is pulse pressure variation as a predictor of fluid responsiveness? *Anesthesiology*. 2012;116(3):740.
6. Gnjidic D, Hilmer SN, Blyth FM, et al. Polypharmacy cutoff and outcomes: five or more medicines were used to identify community-dwelling older men at risk of different adverse outcomes. *J Clin Epidemiol*. 2012;65(9):989-995.
7. Pan H-C, Jenq C-C, Tsai M-H, et al. Risk models and scoring systems for predicting the prognosis in critically ill cirrhotic patients with acute kidney injury: a prospective validation study. *PLoS One*. 2012;7(12):e51094.
8. Yin J, Tian L. Joint confidence region estimation for area under roc curve and youden index. *Stat Med*. 2014;33(6):985-1000.
9. Yin J, Tian L. Joint inference about sensitivity and specificity at the optimal cut-off point associated with Youden index. *Comput Stat Data Anal*. 2014;77:1-13.
10. Adimari G, Chiogna M. Simple nonparametric confidence regions for the evaluation of continuous-scale diagnostic tests. *Int J Biostat*. 2010;6(1):1557-4679.
11. Bantis LE, Nakas CT, Reiser B. Construction of confidence regions in the roc space after the estimation of the optimal youden index-based cut-off point. *Biometrics*. 2014;70(1):212-223.
12. Hajian-Tilaki K. The choice of methods in determining the optimal cut-off value for quantitative diagnostic test evaluation. *Stat Methods Med Res*. 2018;27(8):2374-2383.
13. Hua J, Tian L. A comprehensive and comparative review of optimal cut-points selection methods for diseases with multiple ordinal stages. *J Biopharm Stat*. 2019;30(1):46-68.
14. Schisterman EF, Perkins N. Confidence intervals for the youden index and corresponding optimal cut-point. *Commun Stat Simul Comput*. 2007;36(3):549-563.
15. Casella G, Berger RL. *Statistical Inference*. Vol 2. Pacific Grove, CA: Cengage Learning; 2002.
16. Weerahandi S. Generalized confidence intervals. *J Am Stat Assoc*. 1993;88(423):899-905.
17. Weerahandi S. Anova under unequal error variances. *Biometrics*. 1995;51(2):589-599.
18. Weerahandi S, Berger VW. Exact inference for growth curves with intraclass correlation structure. *Biometrics*. 1999;55(3):921-924.
19. Krishnamoorthy K, Lu Y. Inferences on the common mean of several normal populations based on the generalized variable method. *Biometrics*. 2003;59(2):237-247.
20. Tian L, Cappelleri JC. A new approach for interval estimation and hypothesis testing of a certain intraclass correlation coefficient: the generalized variable method. *Stat Med*. 2004;23(13):2125-2135.
21. Lin S, Lee JC, Wang R. Generalized inferences on the common mean vector of several multivariate normal populations. *J Stat Plan Infer*. 2007;137(7):2240-2249.

22. Tian L. Confidence intervals for $p(y_1 > y_2)$ with normal outcomes in linear models. *Stat Med*. 2008;27(21):4221-4237.
23. Davidson R, MacKinnon JG. Bootstrap tests: how many bootstraps? *Econ Rev*. 2000;19(1):55-68.
24. Wilcox RR. *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy*. New York, NY, USA: Springer; 2010.
25. Mueller SG, Weiner MW, Thal LJ, et al. The alzheimer's disease neuroimaging initiative. *Neuroimaging Clin*. 2005;15(4):869-877.
26. Roberts R, Knopman DS. Classification and epidemiology of mci. *Clin Geriatr Med*. 2013;29(4):753-772.
27. Sachdev PS, Lipnicki DM, Kochan NA, et al. The prevalence of mild cognitive impairment in diverse geographical and ethnocultural regions: the cosmic collaboration. *PLoS One*. 2015;10(11):e0142388.
28. A. Association et al. 2018 alzheimer's disease facts and figures. *Alzheimers Dement*. 2018;14(3):367-429.
29. Rücker G, Schumacher M. Summary roc curve based on a weighted youden index for selecting an optimal cutpoint in meta-analysis of diagnostic accuracy. *Stat Med*. 2010;29(30):3069-3078.
30. Ballard C et al. Enrichment factors for clinical trials in mild-to-moderate alzheimer's disease. *Alzheimer's Dement Transl Res Clin Interventions*. 2019;5:164-174.
31. Wolz R et al. Enrichment of clinical trials in mci due to ad using markers of amyloid and neurodegeneration. *Neurology*. 2016;87(12):1235-1241.
32. Holland D, McEvoy LK, Desikan RS, Dale AM, Initiative ADN, et al. Enrichment and stratification for predementia alzheimer disease clinical trials. *PLoS One*. 2012;7(10):e47739.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Schaible BJ, Yin J. Joint confidence region estimation on predictive values. *Pharmaceutical Statistics*. 2021;1–21. https://doi.org/10.1002/pst.2131